ENTROPIC INFERENCE AND THE FOUNDATIONS OF PHYSICS

ARIEL CATICHA
Department of Physics, University at Albany–SUNY

Foreword

This book was specially prepared by Ariel Caticha, from SUNY-Albany, for a tutorial on the subject of Entropic Inference and the Foundations of Physics, to be presented at EBEB-2012, the 11th Brazilian Meeting on Bayesian Statistics, held on March 18-22 at Amparo, São Paulo.

The organizing committee of EBEB 2012 had some goals for this conference, including:

- To publish high quality proceedings, in the hope of transforming the character of the EBEB meetings, from the current – national meeting with some important international guests, to a future – international meeting with an active participation of the local scientific community.
- To promote the interaction of the statistics community with researchers from other areas. This includes statistical model development for foreign application areas, but also comprises learning, incorporation or adaptation by statistical science of new forms of probabilistic modeling or alternative uncertainty representations originated in other fields.
- To have some in-depth tutorials, including the production of textbooks and other didactic materials. These tutorials should benefit all of us, but are specially intended for upper level under-graduate and graduate students trying to acquire familiarity with new areas of research.

Ariel Caticha, in collaboration with the SUNY-Albany Information Physics / Bayesian Statistics group, helped us to accomplish all of the aforementioned goals. For their efforts, we are very grateful.

Julio Michael Stern,
for the EBEB 2012 Organizing Committee,
São Paulo, February 22, 2012.

Contents

Preface

Science consists in using information about the world for the purpose of predicting, explaining, understanding, and/or controlling phenomena of interest. The basic difficulty is that the available information is usually insufficient to attain any of those goals with certainty. A central concern in these lectures will be the problem of inductive inference, that is, the problem of reasoning under conditions of incomplete information.

Our goal is twofold. First, to develop the main tools for inference — probability and entropy — and to demonstrate their use. And second, to demonstrate their importance for physics. More specifically our goal is to clarify the conceptual foundations of physics by deriving the fundamental laws of statistical mechanics and of quantum mechanics as examples of inductive inference. Perhaps all physics can be derived in this way.

The level of these lectures is somewhat uneven. Some topics are fairly advanced — the subject of recent research — while some other topics are very elementary. I can give two related reasons for including both in the same book. The first is pedagogical: these are lectures — the easy stuff has to be taught too. More importantly, the standard education of physicists includes a very inadequate study of probability and even of entropy. The result is a widespread misconception that these "elementary" subjects are trivial and unproblematic — that the real problems of theoretical and experimental physics lie elsewhere.

As for the second reason, it is inconceivable that the interpretations of probability and of entropy would turn out to bear no relation to our understanding of physics. Indeed, if the only notion of probability at our disposal is that of a frequency in a large number of trials one might be led to think that the ensembles of statistical mechanics must be real, and to regard their absence as an urgent problem demanding an immediate solution — perhaps an ergodic solution. One might also be led to think that analogous ensembles are needed in quantum theory perhaps in the form of parallel worlds. Similarly, if the only available notion of entropy is derived from thermodynamics, one might end up thinking that entropy is some physical quantity that can be measured in the lab, and that it has little or no relevance beyond statistical mechanics.

It is very worthwhile to revisit the "elementary" basics because usually the basics are not elementary at all, and even more importantly, because they are so fundamental.

Acknowledgements: Most specially I am indebted to C. R. Rodríguez and to

N. Caticha, whose views on these matters have profoundly influenced my own, but I have also learned much from discussions with many colleagues and friends: D. Bartolomeo, C. Cafaro, V. Dose, K. Earle, R. Fischer, A. Garrett, A. Giffin, P. Goggans, A. Golan, M. I. Gomez, P. Goyal, M. Grendar, D. T. Johnson, K. Knuth, S. Nawaz, R. Preuss, T. Seidenfeld, J. Skilling, R. Spekkens, and C.-Y. Tseng. I would also like to thank all the students who over the years have taken my course on *Information Physics*; their questions and doubts have very often helped clear my own questions and doubts. I would also like to express my special gratitude to Julio Stern for his continued encouragement to get my lectures published and to J. Stern, C. A. de Bragança Pereira, A. Polpo, M. Lauretto and M. A. Diniz, organizers of EBEB 2012 for undertaking their publication.

Albany, February 2012.

Chapter 1

Inductive Inference and Physics

The process of drawing conclusions from available information is called inference. When the available information is sufficient to make unequivocal, unique assessments of truth we speak of making deductions: on the basis of a certain piece of information we deduce that a certain proposition is true. The method of reasoning leading to deductive inferences is called logic. Situations where the available information is insufficient to reach such certainty lie outside the realm of logic. In these cases we speak of doing inductive inference, and the methods deployed are those of probability theory and entropic inference.

1.1 Probability

The question of the meaning and interpretation of the concept of probability has long been controversial. Needless to say the interpretations offered by various schools are at least partially successful or else they would already have been discarded. But the different interpretations are not equivalent. They lead people to ask different questions and to pursue their research in different directions. Some questions may become essential and urgent under one interpretation while totally irrelevant under another. And perhaps even more important: under different interpretations equations can be used differently and this can lead to different predictions.

The frequency interpretation

Historically the *frequentist* interpretation has been the most popular: the probability of a random event is given by the relative number of occurrences of the event in a sufficiently large number of identical and independent trials. The appeal of this interpretation is that it seems to provide an empirical method to estimate probabilities by counting over the ensemble of trials. The magnitude

of a probability is obtained solely from the observation of many repeated trials and does not depend on any feature or characteristic of the observers. Probabilities interpreted in this way have been called *objective*. This view dominated the fields of statistics and physics for most of the 19th and 20th centuries (see, *e.g.*, [von Mises 1957]).

One disadvantage of the frequentist approach has to do with matters of rigor: what precisely does one mean by 'random'? If the trials are sufficiently identical, shouldn't one always obtain the same outcome? Also, if the interpretation is to be validated on the basis of its operational, empirical value, how large should the number of trials be? Unfortunately, the answers to these questions are neither easy nor free from controversy. By the time the tentative answers have reached a moderately acceptable level of sophistication the intuitive appeal of this interpretation has long been lost. In the end, it seems the frequentist interpretation is most useful when left a bit vague.

A more serious objection is the following. In the frequentist approach the notion of an ensemble of trials is central. In cases where there is a natural ensemble (tossing a coin, or a die, spins in a lattice, etc.) the frequency interpretation seems natural enough. But for many other problems the construction of an ensemble is at best highly artificial. For example, consider the probability of there being life in Mars. Are we to imagine an ensemble of Mars planets and solar systems? In these cases the ensemble would be purely hypothetical. It offers no possibility of an empirical determination of a relative frequency and this defeats the original goal of providing an objective operational interpretation of probabilities as frequencies. In yet other problems there is no ensemble at all: consider the probability that the nth digit of the number π be 7. Are we to imagine alternative universes with different values for the number π? It is clear that there a number of interesting problems where one suspects the notion of probability could be quite useful but which nevertheless lie outside the domain of the frequentist approach.

The Bayesian interpretations

According to the Bayesian interpretations, which can be traced back to Bernoulli and Laplace, but have only achieved popularity in the last few decades, a probability reflects the confidence, the degree of belief of an individual in the truth of a proposition. These probabilities are said to be *Bayesian* because of the central role played by Bayes' theorem – a theorem which is actually due to Laplace. This approach enjoys several advantages. One is that the difficulties associated with attempting to pinpoint the precise meaning of the word 'random' can be avoided. Bayesian probabilities are not restricted to repeatable events; they allow us to reason in a consistent and rational manner about unique, singular events. Thus, in going from the frequentist to the Bayesian interpretations the domain of applicability and therefore the usefulness of the concept of probability is considerably enlarged.

The crucial aspect of Bayesian probabilities is that different individuals may have different degrees of belief in the truth of the very same proposition, a

fact that is described by referring to Bayesian probabilities as being *subjective*. This term is somewhat misleading because there are (at least) two views on this matter, one is the so-called subjective Bayesian or *personalistic* view (see, *e.g.*, [Savage 1972; Howson Urbach 1993; Jeffrey 2004]), and the other is the *objective* Bayesian view (see *e.g.* [Jeffreys 1939; Cox, 1946; Jaynes 1985, 2003; Lucas 1970]). For an excellent elementary introduction with a philosophical perspective see [Hacking 2001]. According to the subjective view, two reasonable individuals faced with the same evidence, the same information, can legitimately differ in their confidence in the truth of a proposition and may therefore assign different probabilities. Subjective Bayesians accept that an individual can change his or her beliefs, merely on the basis of introspection, reasoning, or even revelation.

At the other end of the Bayesian spectrum, the objective Bayesian view considers the theory of probability as an extension of logic. It is said then that a probability measures a degree of *rational* belief. It is assumed that the objective Bayesian has thought so long and hard about how probabilities are assigned that no further reasoning will induce a revision of beliefs except when confronted with new information. In an ideal situation two different individuals will, on the basis of the same information, assign the same probabilities.

Subjective or objective?

Whether Bayesian probabilities are subjective or objective is still a matter of dispute. Our position is that they lie somewhere in between. Probabilities will always retain a "subjective" element because translating information into probabilities involves judgments and different people will inevitably judge differently.

On the other hand, it is a presupposition of thought itself that some beliefs are better than others — otherwise why go through the trouble of thinking? And they are "objectively" better in that they provide better guidance about how to cope with the world. The adoption of better beliefs has real consequences. Similarly, not all probability assignments are equally useful and it is plausible that what makes some assignments better than others is that they represent or reflect some objective feature of the world. One might even say that what makes them better is that they provide a better guide to the "truth". It is the conviction that posterior probabilities are somehow objectively better than prior probabilities that provides the justification for going through the troubles of gathering information and using it to update our beliefs.

We shall find that while the subjective element in probabilities can never be completely eliminated, the rules for processing information, that is, the rules for updating probabilities, are themselves quite objective. This means that the new information can be objectively processed and incorporated into our posterior probabilities. Thus, it is quite possible to continuously suppress the subjective elements while enhancing the objective elements as we process more and more information.

Thus, probabilities can be characterized by both subjective and objective elements and, ultimately, it is their objectivity that makes probabilities use-

ful. There is much to be gained by rejecting the sharp subjective/objective dichotomy and replacing it with a continuous spectrum of intermediate possibilities.[1]

1.2 Designing a framework for inductive inference

A common hope in both science and philosophy has been to find a secure foundation for knowledge on which to build science, mathematics, and philosophy. So far the search has not been successful and everything indicates that such indubitable foundation is nowhere to be found. Accordingly, we adopt a pragmatic attitude: there are ideas about which we can have greater or lesser confidence, and from these we can infer the plausibility of others; but there is nothing about which we can have full certainty and complete knowledge.

Inductive inference in its Bayesian/entropic form is a framework designed for the purpose of coping with the world in a rational way in situations where the information available is incomplete. The framework must solve two related problems. First, it must allow for convenient representations of states of partial knowledge — this is handled through the introduction of probabilities. Second, it must allow us to update from one state of knowledge to another when new information becomes available — this is handled through the introduction of relative entropy as the tool for updating. *The theory of probability cannot be separate from a theory for updating probabilities.*

The framework for inference will be constructed by a process of *eliminative induction.* The objective is to design the appropriate tools, which in our case, means designing the theory of probability and entropy. The different ways in which probabilities and entropies are defined and handled will lead to different inference schemes and one can imagine a vast variety of possibilities. To select one we must first have a clear idea of the function that those tools are supposed to perform, that is, we must specify *design criteria* or *design specifications* that the desired inference framework must obey. Finally, in the *eliminative* part of the process one proceeds to systematically rule out all those inference schemes that fail to comply with the design criteria — that is, that fail to perform as desired.

There is no implication that an inference framework designed in this way is in any way "true", or that it succeeds because it achieves some special intimate agreement with reality. Instead, the claim is pragmatic: the method succeeds to the extent that *the inference framework works as designed* and its performance will be deemed satisfactory as long as it leads to scientific models that are empirically adequate. Whatever design criteria are chosen, they are meant to be only provisional — just like everything else in science, there is no reason to consider them immune from further change and improvement.

[1] This position bears a resemblance to the rejection of the fact/value dichotomy advocated in [Putnam 1991, 2003].

The pros and cons of eliminative induction have been the subject of considerable philosophical research (e.g. [Earman 1992; Hawthorne 1993; Godfrey-Smith 2003]). On the negative side, eliminative induction, like any other form of induction, is not guaranteed to work. On the positive side, eliminative induction adds an interesting twist to Popper's scientific methodology. According to Popper scientific theories can never be proved right, they can only be proved false; a theory is corroborated only to the extent that all attempts at falsifying it have failed. Eliminative induction is fully compatible with Popper's notions but the point of view is just the opposite. Instead of focusing on *failure* to falsify one focuses on *success*: it is the successful falsification of all rival theories that corroborates the surviving one. The advantage is that one acquires a more explicit understanding of why competing theories are eliminated.

In chapter 2 we address the problem of the design and construction of probability theory as a tool for inference. In other words, we show that degrees of rational belief, those measures of plausibility that we require to do inference, should be manipulated and calculated according to the ordinary rules of the calculus of probabilities.

The problem of designing a theory for updating probabilities is addressed mostly in chapter 6 and then completed in chapter 8. We discuss the central question "What is information?" and show that there is a unique method to update from an old set of beliefs codified in a prior probability distribution into a new set of beliefs described by a new, posterior distribution when the information available is in the form of a constraint on the family of acceptable posteriors. In this approach the tool for inference is entropy. A central achievement is the complete unification of Bayesian and entropic methods.

1.3 Entropic Physics

Once the framework of entropic inference has been constructed we deploy it to clarify the conceptual foundations of physics.

Prior to the work of Jaynes it was suspected that there was a connection between thermodynamics and information theory. But the connection took the form of an analogy between the two fields: Shannon's information theory was designed to be useful in engineering[2] while thermodynamics was meant to be "true" by virtue of reflecting "laws of nature". The gap was enormous; to this day many still think that the analogy is purely accidental. With the work of Jaynes, however, it became clear that the connection is not an accident: the crucial link is that both situations involve reasoning with incomplete information. This development was significant for many subjects — engineering, statistics, computation — but for physics the impact of such a change in perspective is absolutely enormous: thermodynamics and statistical mechanics provided the first example of a fundamental theory that, instead of being a direct image of nature, should be interpreted as a scheme for inference about nature. Beyond

[2]Even as late as 1961 Shannon expressed doubts that information theory would ever find application in fields other than communication theory. [Tribus 1978]

the impact on statistical mechanics itself, the obvious question is: Are there other examples? The answer is yes.

Our goal in chapter 5 is to provide an explicit discussion of statistical mechanics as an example of entropic inference; the chapter is devoted to discussing and clarifying the foundations of thermodynamics and statistical mechanics. The development is carried largely within the context of Jaynes' MaxEnt formalism and we show how several central topics such as the equal probability postulate, the second law of thermodynamics, irreversibility, reproducibility, and the Gibbs paradox can be considerably clarified when viewed from the information/inference perspective.

In chapters 9 and 10 we explore new territory. These chapters are devoted to deriving quantum theory as an example of entropic inference. The challenge is that the theory involves dynamics and time in a fundamental way. It is significant that the full framework of entropic inference derived in chapters 6 an 8 is needed here — the old entropic methods developed by Shannon and Jaynes are no longer sufficient.

The payoff is considerable. A vast fraction of the quantum formalism is derived and the entropic approach offers new insights into many topics that are central to quantum theory: the interpretation of the wave function, the wave-particle duality, the quantum measurement problem, the introduction and interpretation of observables other than position, including momentum, the corresponding uncertainty relations, and most important, it leads to a theory of entropic time. The overall conclusion is that *the laws of quantum mechanics are not laws of nature; they are rules for processing information about nature.*

Chapter 2

Probability

Our goal is to establish the theory of probability as the general theory for reasoning on the basis of incomplete information. This requires us to tackle two different problems. The first problem is to figure out how to achieve a quantitative description of a state of partial knowledge. Once this is settled we address the second problem of how to update from one state of knowledge to another when new information becomes available.

Throughout we will assume that the subject matter – the set of propositions the truth of which we want to assess – has been clearly specified. This question of what it is that we are actually talking about is much less trivial than it might appear at first sight.[1] Nevertheless, it will not be discussed further.

The first problem, that of describing or characterizing a state of partial knowledge, requires that we quantify the degree to which we believe each proposition in the set is true. The most basic feature of these beliefs is that they form an interconnected web that must be internally consistent. The idea is that in general the strengths of one's beliefs in some propositions are constrained by one's beliefs in other propositions; beliefs are not independent of each other. For example, the belief in the truth of a certain statement a is strongly constrained by the belief in the truth of its negation, not-a: the more I believe in one, the less I believe in the other.

The second problem, that of updating from one consistent web of beliefs to another when new information becomes available, will be addressed for the special case that the information is in the form of data. The basic updating strategy reflects the conviction that what we learned in the past is valuable, that the web of beliefs should only be revised to the extent required by the data. We will see that this principle of *minimal updating* leads to the uniquely natural rule that is widely known as Bayes' rule. (More general kinds of information can also be processed using the minimal updating principle but they require a more sophisticated tool, namely, relative entropy. This topic will be extensively

[1] Consider the example of quantum mechanics: Are we talking about particles, or about experimental setups, or both? Are we talking about position variables, or about momenta, or both? Or neither? Is it the position of the particles or the position of the detectors?

explored later.) As an illustration of the enormous power of Bayes' rule we will briefly explore its application to data analysis.

2.1 The design of probability theory

Science requires a framework for inference on the basis of incomplete information. We will show that the quantitative measures of *plausibility* or *degrees of belief* that are the tools for reasoning should be manipulated and calculated using the ordinary rules of the calculus of probabilities — and *therefore* probabilities *can* be interpreted as degrees of belief.

The procedure we follow differs in one remarkable way from the traditional way of setting up physical theories. Normally one starts with the mathematical formalism, and then one proceeds to try to figure out what the formalism might possibly mean; one tries to append an interpretation to it. This is a very difficult problem; historically it has affected not only statistical physics — what is the meaning of probabilities and of entropy — but also quantum theory — what is the meaning of wave functions and amplitudes. Here we proceed in the opposite order, we first decide what we are talking about, degrees of belief or degrees of plausibility (we use the two expressions interchangeably) and then we *design* rules to manipulate them; we design the formalism, we construct it to suit our purposes. The advantage of this approach is that the issue of meaning, of interpretation, is settled from the start.

2.1.1 Rational beliefs?

Before we proceed further it may be important to emphasize that the degrees of belief discussed here are those held by an idealized rational agent that would not be subject to the practical limitations under which we humans operate. Different individuals may hold different beliefs and it is certainly important to figure out what those beliefs might be — perhaps by observing their gambling behavior — but this is not our present concern. Our objective is neither to assess nor to describe the subjective beliefs of any particular individual. Instead we deal with the altogether different but very common problem that arises when we are confused and we want some guidance about what we are *supposed* to believe. Our concern here is not so much with beliefs as they actually are, but rather, with beliefs as they *ought* to be — at least as they ought to be to deserve to be called *rational*. We are concerned with the ideal standard of rationality that we humans ought to attain at least when discussing scientific matters.

The concept of rationality is notoriously difficult to pin down. One thing we can say is that rational beliefs are constrained beliefs. The essence of rationality lies precisely in the existence of some constraints — not everything goes. We need to identify some *normative criteria of rationality* and the difficulty is to find criteria that are sufficiently general to include all instances of rationally justified belief. Here is our first criterion of rationality:

> *The inference framework must be based on assumptions that have wide appeal and universal applicability.*

Whatever guidelines we pick they must be of general applicability — otherwise they fail when most needed, namely, when not much is known about a problem. Different rational agents can reason about different topics, or about the same subject but on the basis of different information, and therefore they could hold different beliefs, but they must agree to follow the same rules. What we seek here are not the specific rules of inference that will apply to this or that specific instance; what we seek is to identify some few features that all instances of rational inference might have in common.

The second criterion is that

> *The inference framework must not be self-refuting.*

It may not be easy to identify criteria of rationality that are sufficiently general and precise. Perhaps we can settle for the more manageable goal of avoiding irrationality in those glaring cases where it is easily recognizable. And this is the approach we take: rather than providing a precise criterion of rationality to be carefully followed, we design a framework with the more modest goal of avoiding some forms of irrationality that are perhaps sufficiently obvious to command general agreement. The basic desire is that the web of rational beliefs must avoid inconsistencies. If a quantity can be inferred in two different ways the two ways must agree. As we shall see this requirement turns out to be extremely restrictive.

Finally,

> *The inference framework must be useful in practice — it must allow quantitative analysis.*

Otherwise, why bother?

Whatever specific design criteria are chosen, one thing must be clear: they are justified on purely pragmatic grounds and therefore they are meant to be only provisional. Rationality itself is not immune to change and improvement. Given some criteria of rationality we proceed to construct models of the world, or better, models that will help us deal with the world — predict, control, and explain the facts. The process of improving these models — better models are those that lead to more accurate predictions, more accurate control, and more lucid and encompassing explanations of more facts, not just the old facts but also of new and hopefully even unexpected facts — may eventually suggest improvements to the rationality criteria themselves. Better rationality leads to better models which leads to better rationality and so on. The method of science is not independent from the contents of science.

2.1.2 Quantifying rational belief

In order to be useful we require an inference framework that allows quantitative reasoning. The first obvious question concerns the type of quantity that will

represent the intensity of beliefs. Discrete categorical variables are not adequate for a theory of general applicability; we need a much more refined scheme.

Do we believe proposition a more or less than proposition b? Are we even justified in comparing propositions a and b? The problem with propositions is not that they cannot be compared but rather that the comparison can be carried out in too many different ways. We can classify propositions according to the degree we believe they are true, their plausibility; or according to the degree that we desire them to be true, their utility; or according to the degree that they happen to bear on a particular issue at hand, their relevance. We can even compare propositions with respect to the minimal number of bits that are required to state them, their description length. The detailed nature of our relations to propositions is too complex to be captured by a single real number. What we claim is that a single real number is sufficient to measure one specific feature, the sheer intensity of rational belief. This should not be too controversial because it amounts to a tautology: an "intensity" is precisely the type of quantity that admits no more qualifications than that of being more intense or less intense; it is captured by a single real number.

However, some preconception about our subject is unavoidable; we need some rough notion that a belief is not the same thing as a desire. But how can we know that we have captured pure belief and not belief contaminated with some hidden desire or something else? Strictly we can't. We hope that our mathematical description captures a sufficiently purified notion of rational belief, and we can claim success only to the extent that the formalism proves to be useful.

The inference framework will capture two intuitions about rational beliefs. First, we take it to be a defining feature of the intensity of *rational* beliefs that if a is more believable than b, and b more than c, then a is more believable than c. Such transitive rankings can be implemented using real numbers we are again led to claim that

> *Degrees of rational belief (or, as we shall later call them, probabilities) are represented by real numbers.*

Before we proceed further we need to establish some notation. The following choice is standard.

Notation

For every proposition a there exists its negation not-a, which will be denoted $\tilde{a}$. If a is true, then $\tilde{a}$ is false and vice versa.

Given any two propositions a and b the conjunction "a AND b" is denoted ab or $a \wedge b$. The conjunction is true if and only if both a and b are true.

Given a and b the disjunction "a OR b" is denoted by $a \vee b$ or (less often) by $a + b$. The disjunction is true when either a or b or both are true; it is false when both a and b are false.

Typically we want to quantify the degree of belief in $a \vee b$ and in ab in the context of some background information expressed in terms of some proposition

c in the same universe of discourse as a and b. Such propositions we will write as $a \vee b|c$ and $ab|c$.

The real number that represents the degree of belief in $a|b$ will initially be denoted $[a|b]$ and eventually in its more standard form $p(a|b)$ and all its variations.

Degrees of rational belief will range from the extreme of total certainty, $[a|a] = v_T$, to total disbelief, $[\tilde{a}|a] = v_F$. The transitivity of the ranking scheme implies that there is a single value v_F and a single v_T.

The representation of OR and AND

The inference framework is designed to include a second intuition concerning rational beliefs:

> *In order to be rational our beliefs in $a \vee b$ and ab must be somehow related to our separate beliefs in a and b.*

Since the goal is to design a quantitative theory, we require that these relations be represented by some functions F and G,

$$[a \vee b|c] = F([a|c], [b|c], [a|bc], [b|ac]) \tag{2.1}$$

and

$$[ab|c] = G([a|c], [b|c], [a|bc], [b|ac]) \ . \tag{2.2}$$

Note the *qualitative* nature of this assumption: what is being asserted is the existence of some unspecified functions F and G and not their specific functional forms. The same F and G are meant to apply to all propositions; what is being *designed* is a single inductive scheme of universal applicability. Note further that the arguments of F and G include all four possible degrees of belief in a and b in the context of c and not any potentially questionable subset.

The functions F and G provide a representation of the Boolean operations AND and OR. The requirement that F and G reflect the appropriate associative and distributive properties of the Boolean AND and OR turns out to be extremely constraining. Indeed, we will show that there is essentially a single representation that is equivalent to probability theory. (All allowed representations are equivalent to each other.)

In section 3 the associativity of OR is shown to lead to a constraint that requires the function F to be equivalent to the sum rule for probabilities. In section 4 we focus on the distributive property of AND over OR and the corresponding constraint leads to the product rule for probabilities.[2]

[2]Our subject is degrees of rational belief but the algebraic approach followed here [Caticha 2009] can be pursued in its own right irrespective of any interpretation. It was used in [Caticha 1998] to derive the manipulation rules for complex numbers interpreted as quantum mechanical amplitudes; in [Knuth 2003] in the mathematical problem of assigning real numbers (valuations) on general distributive lattices; and in [Goyal et al 2010] to justify the use of complex numbers for quantum amplitudes.

Our method will be *design by eliminative induction*: now that we have identified a sufficiently broad class of theories — quantitative theories of universal applicability, with degrees of belief represented by real numbers and the operations of conjunction and disjunction represented by functions — we can start weeding the unacceptable ones out.

An aside on the Cox axioms

The development of probability theory in the following sections follows a path clearly inspired by [Cox 1946]. A brief comment may be appropriate.

Cox derived the sum and product rules by focusing on the properties of conjunction and negation. He assumed as one of his axioms that the degree of belief in a proposition a conditioned on b being true, which we write as $[a|b]$, is related to the degree of belief corresponding to its negation, $[\tilde{a}|b]$, through some definite but initially unspecified function f,

$$[\tilde{a}|b] = f\left([a|b]\right) \ . \tag{2.3}$$

This statement expresses the intuition that the more one believes in $a|b$, the less one believes in $\tilde{a}|b$.

A second Cox axiom is that the degree of belief of "a AND b given c," written as $[ab|c]$, must depend on $[a|c]$ and $[b|ac]$,

$$[ab|c] = g\left([a|c], [b|ac]\right) \ . \tag{2.4}$$

This is also very reasonable. When asked to check whether "a AND b" is true, we first look at a; if a turns out to be false the conjunction is false and we need not bother with b; therefore $[ab|c]$ must depend on $[a|c]$. If a turns out to be true we need to take a further look at b; therefore $[ab|c]$ must also depend on $[b|ac]$. Strictly $[ab|c]$ could in principle depend on all four quantities $[a|c]$, $[b|c]$, $[a|bc]$ and $[b|ac]$, an objection that has a long history. It was partially addressed in [Tribus 1969; Smith Erickson 1990; Garrett 1996].

Cox's important contribution was to realize that consistency constraints derived from the associativity property of AND and from the compatibility of AND with negation were sufficient to demonstrate that degrees of belief should be manipulated according to the laws of probability theory. We shall not pursue this line of development here. See [Cox 1946; Jaynes 1957a, 2003].

2.2 The sum rule

Our first goal is to determine the function F that represents OR. The space of functions of four arguments is very large. To narrow down the field we initially restrict ourselves to propositions a and b that are mutually exclusive in the context of d. Thus,

$$[a \vee b|d] = F([a|d], [b|d], v_F, v_F) \ , \tag{2.5}$$

which effectively restricts F to a function of only two arguments,

$$[a \vee b|d] = F([a|d], [b|d]) \ . \tag{2.6}$$

2.2.1 The associativity constraint

As a minimum requirement of rationality we demand that the assignment of degrees of belief be consistent: if a degree of belief can be computed in two different ways the two ways must agree. How else could we claim to be rational? All functions F that fail to satisfy this constraint must be discarded.

Consider any three mutually exclusive statements a, b, and c in the context of a fourth d. The consistency constraint that follows from the associativity of the Boolean OR,

$$(a \vee b) \vee c = a \vee (b \vee c) \ , \tag{2.7}$$

is remarkably constraining. It essentially determines the function F. Start from

$$[a \vee b \vee c|d] = F\left([a \vee b|d], [c|d]\right) = F\left([a|d], [b \vee c|d]\right) \ . \tag{2.8}$$

Use F again for $[a \vee b|d]$ and also for $[b \vee c|d]$, we get

$$F\{F\left([a|d], [b|d]\right), [c|d]\} = F\{[a|d], F\left([b|d], [c|d]\right)\} \ . \tag{2.9}$$

If we call $[a|d] = x$, $[b|d] = y$, and $[c|d] = z$, then

$$F\{F(x, y), z\} = F\{x, F(y, z)\} \ . \tag{2.10}$$

Since this must hold for arbitrary choices of the propositions a, b, c, and d, we conclude that *in order to be of universal applicability* the function F must satisfy (2.10) for arbitrary values of the real numbers (x, y, z). Therefore the function F must be associative.

Remark: The requirement of universality is crucial. Indeed, in a universe of discourse with a discrete and finite set of propositions it is conceivable that the triples (x, y, z) in (2.10) do not form a dense set and therefore one cannot conclude that the function F must be associative for arbitrary values of x, y, and z. For each specific finite universe of discourse one could design a tailor-made, single-purpose model of inference that could be consistent, i.e. it would satisfy (2.10), without being equivalent to probability theory. However, we are concerned with designing a theory of inference of universal applicability, a single scheme applicable to *all universes of discourse* whether discrete and finite or otherwise. And the scheme is meant to be used by *all rational agents* irrespective of their state of belief — which need not be discrete. Thus, a framework designed for broad applicability requires that the values of x form a dense set.[3]

[3]The possibility of alternative probability models was raised in [Halpern 1999]. That these models are ruled out by universality was argued in [Van Horn 2003] and [Caticha 2009].

2.2.2 The general solution and its regraduation

Equation (2.10) is a functional equation for F. It is easy to see that there exist an infinite number of solutions. Indeed, by direct substitution one can check that eq.(2.10) is satisfied by any function of the form

$$F\left(x,y\right)=\phi^{-1}\left(\phi\left(x\right)+\phi\left(y\right)\right)\ , \tag{2.11}$$

where ϕ is an arbitrary invertible function. What is not so easy to to show is this is also the *general* solution, that is, given ϕ one can calculate F and, conversely, given any associative F one can calculate the corresponding ϕ. Cox's proof of this result is given in section 2.2.4 [Cox 1946; Jaynes 1957a; Aczel 1966].

The significance of eq.(2.11) becomes apparent once it is rewritten as

$$\phi\left(F\left(x,y\right)\right)=\phi\left(x\right)+\phi\left(y\right)\quad\text{or}\quad\phi\left(\left[a\vee b|d\right]\right)=\phi\left(\left[a|d\right]\right)+\phi\left(\left[b|d\right]\right)\ . \tag{2.12}$$

This last form is central to Cox's approach to probability theory. Note that there was nothing particularly special about the original representation of degrees of plausibility by the real numbers $[a|d],[b|d],\ldots$ Their only purpose was to provide us with a ranking, an ordering of propositions according to how plausible they are. Since the function $\phi(x)$ is monotonic, the same ordering can be achieved using a new set of positive numbers,

$$\xi(a|d)\overset{\text{def}}{=}\phi([a|d]),\quad\xi(b|d)\overset{\text{def}}{=}\phi([b|d]),...\tag{2.13}$$

instead of the old. The original and the regraduated scales are equivalent because by virtue of being invertible the function ϕ is monotonic and therefore preserves the ranking of propositions. However, the regraduated scale is much more convenient because, instead of the complicated rule (2.11), the OR operation is now represented by a much simpler rule,

$$\xi\left(a\vee b|d\right)=\xi\left(a|d\right)+\xi\left(b|d\right)\ , \tag{2.14}$$

just a sum rule. Thus, the new numbers are neither more nor less correct than the old, they are just considerably more convenient.

Perhaps one can make the logic of regraduation a little bit clearer by considering the somewhat analogous situation of introducing the quantity temperature as a measure of degree of "hotness". Clearly any acceptable measure of "hotness" must reflect its transitivity — if a is hotter than b and b is hotter than c then a is hotter than c — which explains why temperatures are represented by real numbers. But the temperature scales can be quite arbitrary. While many temperature scales may serve equally well the purpose of ordering systems according to their hotness, there is one choice — the absolute or Kelvin scale — that turns out to be considerably more convenient because it simplifies the mathematical formalism. Switching from an arbitrary temperature scale to the Kelvin scale is one instance of a convenient regraduation. (The details of temperature regraduation are given in chapter 3.)

In the old scale, before regraduation, we had set the range of degrees of belief from one extreme of total disbelief, $[\tilde{a}|a] = v_F$, to the other extreme of total certainty, $[a|a] = v_T$. At this point there is not much that we can say about the regraduated $\xi_T = \phi(\nu_T)$ but $\xi_F = \phi(\nu_F)$ is easy to evaluate. Setting $d = \tilde{a}$ in eq.(2.14) gives

$$\xi\,(a \vee b|\tilde{a}) = \xi\,(a|\tilde{a}) + \xi\,(b|\tilde{a}) \ . \tag{2.15}$$

Since $a \vee b|\tilde{a}$ is true if and only if $b|\tilde{a}$ is true, the corresponding degrees of belief must coincide,

$$\xi\,(a \vee b|\tilde{a}) = \xi\,(b|\tilde{a}) \ , \tag{2.16}$$

and therefore

$$\xi\,(a|\tilde{a}) = \xi_F = 0 \ . \tag{2.17}$$

2.2.3 The general sum rule

The restriction to mutually exclusive propositions in the sum rule eq.(2.14) can be easily lifted. Any proposition a can be written as the disjunction of two mutually exclusive ones, $a = (ab) \vee (a\tilde{b})$ and similarly $b = (ab) \vee (\tilde{a}b)$. Therefore for any two *arbitrary* propositions a and b we have

$$a \vee b = (ab) \vee (a\tilde{b}) \vee (\tilde{a}b) \tag{2.18}$$

Since each of the terms on the right are mutually exclusive the sum rule (2.14) applies,

$$\begin{aligned} \xi(a \vee b|d) &= \xi(ab|d) + \xi(a\tilde{b}|d) + \xi(\tilde{a}b|d) + [\xi(ab|d) - \xi(ab|d)] \\ &= \xi(ab \vee a\tilde{b}|d) + \xi(ab \vee \tilde{a}b|d) - \xi(ab|d) \ , \end{aligned} \tag{2.19}$$

which leads to the general sum rule,

$$\xi(a \vee b|d) = \xi(a|d) + \xi(b|d) - \xi(ab|d) \ . \tag{2.20}$$

2.2.4 Cox's proof

Understanding the proof that eq.(2.11) is the general solution of the associativity constraint, eq.(2.10), is not necessary for understanding other topics in this book. This section may be skipped on a first reading. The proof given below, due to Cox, [Cox 1946] takes advantage of the fact that our interest is not just to find the most general mathematical solution but rather that we want the most general solution where the function F is to be used for the purpose of inference. This allows us to impose additional constraints on F.

The general strategy in solving equations such as (2.10) is to take partial derivatives to transform the functional equation into a differential equation and then to proceed to solve the latter. Fortunately we can assume that the allowed functions F are continuous and twice differentiable. Indeed, since inference is just quantified common sense, had the function F turned out to be non-differentiable serious doubt would be cast on the legitimacy of the whole scheme.

Furthermore, common sense also requires that $F(x, y)$ be monotonic increasing in both its arguments. Consider a change in the first argument $x = [a|d]$ while holding the second $y = [b|d]$ fixed. A strengthening of one's belief in $a|d$ must be reflected in a corresponding strengthening in ones's belief in $a \vee b|d$. Therefore $F(x, y)$ must be monotonic increasing in its first argument. An analogous line of reasoning shows that $F(x.y)$ must be monotonic increasing in the second argument as well. Therefore,

$$\frac{\partial F(x,y)}{\partial x} \geq 0 \quad \text{and} \quad \frac{\partial F(x,y)}{\partial y} \geq 0 \ . \tag{2.21}$$

Let

$$r \stackrel{\text{def}}{=} F(x,y) \quad \text{and} \quad s \stackrel{\text{def}}{=} F\left(y,z\right) \ , \tag{2.22}$$

and let partial derivatives be denoted by subscripts,

$$F_1(x,y) \stackrel{\text{def}}{=} \frac{\partial F\left(x,y\right)}{\partial x} \geq 0 \quad \text{and} \quad F_2(x,y) \stackrel{\text{def}}{=} \frac{\partial F(x,y)}{\partial y} \geq 0 \ . \tag{2.23}$$

Then eq.(2.10) and its derivatives with respect to x and y are

$$F\left(r,z\right) = F\left(x,s\right) \ , \tag{2.24}$$

$$F_1(r,z)F_1(x,y) = F_1(x,s) \ , \tag{2.25}$$

and

$$F_1(r,z)F_2(x,y) = F_2(x,s)F_1(y,z) \ . \tag{2.26}$$

Eliminating $F_1(r, z)$ from these last two equations we get

$$K(x,y) = K(x,s)F_1(y,z) \ . \tag{2.27}$$

where

$$K(x,y) = \frac{F_2(x,y)}{F_1(x,y)} \ . \tag{2.28}$$

Multiplying eq.(2.27) by $K(y, z)$ and using (2.28) we get

$$K(x,y)K(y,z) = K(x,s)F_2(y,z) \ . \tag{2.29}$$

Differentiating the right hand side of eq.(2.29) with respect to y and comparing with the derivative of eq.(2.27) with respect to z, we have

$$\frac{\partial}{\partial y}\left(K\left(x,s\right)F_2\left(y,z\right)\right) = \frac{\partial}{\partial z}\left(K\left(x,s\right)F_1\left(y,z\right)\right) = \frac{\partial}{\partial z}\left(K\left(x,y\right)\right) = 0. \tag{2.30}$$

Therefore,

$$\frac{\partial}{\partial y}\left(K\left(x,y\right)K\left(y,z\right)\right) = 0, \tag{2.31}$$

or,

$$\frac{1}{K\left(x,y\right)}\frac{\partial K\left(x,y\right)}{\partial y}=-\frac{1}{K\left(y,z\right)}\frac{\partial K\left(y,z\right)}{\partial y} \tag{2.32}$$

since the left hand side is independent of z while the right hand side is independent of of x it must be that they depend only on y,

$$\frac{1}{K\left(x,y\right)}\frac{\partial K\left(x,y\right)}{\partial y}\stackrel{\text{def}}{=}h\left(y\right) \tag{2.33}$$

Integrate using the fact that $K\geq 0$ because both F_1 and F_2 are positive, to get

$$K(x,y)=K(x,0)\,\exp\int_0^y h(y')dy'. \tag{2.34}$$

Similarly,

$$K\left(y,z\right)=K\left(0,z\right)\,\exp-\int_0^y h(y')dy', \tag{2.35}$$

so that

$$K\left(x,y\right)=\alpha\,\frac{H\left(x\right)}{H\left(y\right)}, \tag{2.36}$$

where $\alpha=K(0,0)$ is a constant and $H(x)$ is the *positive* function

$$H(x)\stackrel{\text{def}}{=}\exp\left[-\int_0^x h(x')dx'\right]\geq 0\ . \tag{2.37}$$

On substituting back into eqs.(2.27) and (2.29) we get

$$F_1(y,z)=\frac{H(s)}{H(y)}\qquad\text{and}\qquad F_2(y,z)=\alpha\,\frac{H(s)}{H(z)}. \tag{2.38}$$

Next, use $s=F(y,z)$, so that

$$ds=F_1(y,z)dy+F_2(y,z)dz\ . \tag{2.39}$$

Substituting (2.38) we get

$$\frac{ds}{H(s)}=\frac{dy}{H(y)}+\alpha\frac{dz}{H(z)}\ . \tag{2.40}$$

This is easily integrated. Let

$$\phi\left(x\right)=\phi\left(0\right)\,\exp\left(\int_0^x\frac{dx'}{H(x')}\right)\ , \tag{2.41}$$

be the integrating factor, so that $dx/H(x)=d\phi(x)/\phi(x)$. Then

$$\phi\left(F\left(y,z\right)\right)=\phi\left(y\right)\phi^{\alpha}\left(z\right)\ , \tag{2.42}$$

where a multiplicative constant of integration has been absorbed into the constant $\phi\,(0)$. Applying this function ϕ twice in eq.(2.10) we obtain

$$\phi(x)\phi^{\alpha}(y)\phi^{\alpha}(z) = \phi(x)\phi^{\alpha}(y)\phi^{\alpha^2}(z) \ , \tag{2.43}$$

so that $\alpha = 1$,

$$\phi\,(F\,(y,z)) = \phi\,(y)\,\phi\,(z) \ , \tag{2.44}$$

(The second possibility $\alpha = 0$ is discarded because it leads to $F(x,y) = x$ which is not useful for inference.)

This completes the proof that eq.(2.11) is the general solution of eq.(2.10): Given any $F(x,y)$ that satisfies eq.(2.10) one can construct the corresponding $\phi(x)$ using eqs.(2.28), (2.32), (2.37), and (2.41). Furthermore, since $\phi(x)$ is an exponential its sign is dictated by the constant $\phi\,(0)$ which is positive because the right hand side of eq.(2.44) is positive. Finally, since $H(x) \geq 0$, eq. (2.37), the regraduating function $\phi(x)$ is a monotonic function of its variable x.

2.3 The product rule

Next we consider the function G in eq.(2.2) that represents AND. Once the original plausibilities are regraduated by ϕ according to eq.(2.13), the new function G for the plausibility of a conjunction reads

$$\xi(ab|c) = G[\xi(a|c), \xi(b|c), \xi(a|bc), \xi(b|ac)] \ . \tag{2.45}$$

The space of functions of four arguments is very large so we first narrow it down to just two. Then, we require that the representation of AND be compatible with the representation of OR that we have just obtained. This amounts to imposing a consistency constraint that follows from the distributive properties of the Boolean AND and OR. A final trivial regraduation yields the product rule of probability theory.

2.3.1 From four arguments down to two

We will separately consider special cases where the function G depends on only two arguments, then three, and finally all four arguments. Using commutivity, $ab = ba$, the number of possibilities can be reduced to seven:

$$\xi(ab|c) = G^{(1)}[\xi(a|c), \xi(b|c)] \tag{2.46}$$

$$\xi(ab|c) = G^{(2)}[\xi(a|c), \xi(a|bc)] \tag{2.47}$$

$$\xi(ab|c) = G^{(3)}[\xi(a|c), \xi(b|ac)] \tag{2.48}$$

$$\xi(ab|c) = G^{(4)}[\xi(a|bc), \xi(b|ac)] \tag{2.49}$$

$$\xi(ab|c) = G^{(5)}[\xi(a|c), \xi(b|c), \xi(a|bc)] \tag{2.50}$$

$$\xi(ab|c) = G^{(6)}[\xi(a|c), \xi(a|bc), \xi(b|ac)] \tag{2.51}$$

$$\xi(ab|c) = G^{(7)}[\xi(a|c), \xi(b|c), \xi(a|bc), \xi(b|ac)] \tag{2.52}$$

We want a function G that is of general applicability. This means that the arguments of $G^{(1)} \dots G^{(7)}$ can be varied independently. Our goal is to go down the list and eliminate those possibilities that are clearly unsatisfactory.

First some notation: complete certainty is denoted ξ_T, while complete disbelief is $\xi_F = 0$, eq.(2.17). Derivatives are denoted with a subscript: the derivative of $G^{(3)}(x, y)$ with respect to its second argument y is $G^{(3)}_2(x, y)$.

Type 1: $\xi(ab|c) = G^{(1)}[\xi(a|c), \xi(b|c)]$

The function $G^{(1)}$ is unsatisfactory because it does not take possible correlations between a and b into account. For example, when a and b are mutually exclusive — say, $b = \tilde{a}d$, for some arbitrary d — we have $\xi(ab|c) = \xi_F$ but there are no constraints on either $\xi(a|c) = x$ or $\xi(b|c) = y$. Thus, in order that $G^{(1)}(x, y) = \xi_F$ for arbitrary choices of x and y, $G^{(1)}$ must be a constant which is unacceptable.

Type 2: $\xi(ab|c) = G^{(2)}[\xi(a|c), \xi(a|bc)]$

This function is unsatisfactory because it overlooks the plausibility of $b|c$ [Smith Erickson 1990]. For example: let $a =$ "X is big" and $b =$ "X is big and green" so that $ab = b$. Then

$$\xi(b|c) = G^{(2)}[\xi(a|c), \xi(a|abc)] \quad \text{or} \quad \xi(b|c) = G^{(2)}[\xi(a|c), \xi_T] \ , \tag{2.53}$$

which is clearly unsatisfactory since "green" does not figure anywhere on the right hand side.

Type 3: $\xi(ab|c) = G^{(3)}[\xi(a|c), \xi(b|ac)]$

As we shall see this function turns out to be satisfactory.

Type 4: $\xi(ab|c) = G^{(4)}[\xi(a|bc), \xi(b|ac)]$

This function strongly violates common sense: when $a = b$ we have $\xi(a|c) = G^{(4)}(\xi_T, \xi_T)$, so that $\xi(a|c)$ takes the same constant value irrespective of what a might be [Smith Erickson 1990].

Type 5: $\xi(ab|c) = G^{(5)}[\xi(a|c), \xi(b|c), \xi(a|bc)]$

This function turns out to be equivalent either to $G^{(1)}$ or to $G^{(3)}$ and can therefore be ignored. The proof follows from associativity, $(ab)c|d = a(bc)|d$, which leads to the constraint

$$\begin{aligned} &G^{(5)}\Big[G^{(5)}[\xi(a|d), \xi(b|d), \xi(a|bd)], \xi(c|d), G^{(5)}[\xi(a|cd), \xi(b|cd), \xi(a|bcd)]\Big] \\ &= G^{(5)}[\xi(a|d), G^{(5)}[\xi(b|d), \xi(c|d), \xi(b|cd)], \xi(a|bcd)] \end{aligned}$$

and, with the appropriate identifications,

$$G^{(5)}[G^{(5)}(x,y,z),u,G^{(5)}(v,w,s)] = G^{(5)}[x,G^{(5)}(y,u,w),s] \ . \tag{2.54}$$

Since the variables $x, y \ldots s$ can be varied independently of each other we can take a partial derivative with respect to z,

$$G_1^{(5)}[G^{(5)}(x,y,z),u,G^{(5)}(v,w,s)]G_3^{(5)}(x,y,z) = 0 \ . \tag{2.55}$$

Therefore, either

$$G_3^{(5)}(x,y,z) = 0 \quad \text{or} \quad G_1^{(5)}[G^{(5)}(x,y,z),u,G^{(5)}(v,w,s)] = 0 \ . \tag{2.56}$$

The first possibility says that $G^{(5)}$ is independent of its third argument which means that it is of the type $G^{(1)}$ that has already been ruled out. The second possibility says that $G^{(5)}$ is independent of its first argument which means that it is already included among the type $G^{(3)}$.

Type 6: $\xi(ab|c) = G^{(6)}[\xi(a|c),\xi(a|bc),\xi(b|ac)]$

This function turns out to be equivalent either to $G^{(3)}$ or to $G^{(4)}$ and can therefore be ignored. The proof — which we omit because it is analogous to the proof above for type 5 — also follows from associativity, $(ab)c|d = a(bc)|d$.

Type 7: $\xi(ab|c) = G^{(7)}[\xi(a|c),\xi(b|c),\xi(a|bc),\xi(b|ac)]$

This function turns out to be equivalent either to $G^{(5)}$ or $G^{(6)}$ and can therefore be ignored. Again the proof which uses associativity, $(ab)c|d = a(bc)|d$, is omitted because it is analogous to type 5.

Conclusion:

The possible functions G that are viable candidates for a general theory of inductive inference are equivalent to type $G^{(3)}$,

$$\xi(ab|c) = G[\xi(a|c),\xi(b|ac)] \ . \tag{2.57}$$

2.3.2 The distributivity constraint

The OR function G will be determined by requiring that it be compatible with the regraduated AND function F, which is just a sum. Consider three statements a, b, and c, where the last two are mutually exclusive, in the context of a fourth, d. Distributivity of AND over OR,

$$a\,(b \vee c) = ab \vee ac \ , \tag{2.58}$$

implies that $\xi\,(a\,(b \vee c)\,|d)$ can be computed in two ways,

$$\xi\,(a\,(b \vee c)\,|d) = \xi\,((ab|d) \vee (ac|d)) \ . \tag{2.59}$$

Using eq.(2.14) and (2.57) leads to

$$G[\xi\,(a|d)\,,\xi\,(b|ad)+\xi\,(c|ad)]=\;G[\xi\,(a|d)\,,\xi\,(b|ad)]+G[\xi\,(a|d)\,,\xi\,(c|ad)]\;,$$

which we rewrite as

$$G\,(u,v+w)=G\,(u,v)+G\,(u,w)\;\;, \tag{2.60}$$

where $\xi\,(a|d)=u$, $\xi\,(b|ad)=v$, and $\xi\,(c|ad)=w$.

To solve the functional equation (2.60) we first transform it into a differential equation. Differentiate with respect to v and w,

$$\frac{\partial^2 G\,(u,v+w)}{\partial v\partial w}=0\;, \tag{2.61}$$

and let $v+w=z$, to get

$$\frac{\partial^2 G\,(u,z)}{\partial z^2}=0\;, \tag{2.62}$$

which shows that G is linear in its second argument,

$$G(u,v)=A(u)v+B(u)\;. \tag{2.63}$$

Substituting back into eq.(2.60) gives $B(u)=0$. To determine the function $A(u)$ we note that the degree to which we believe in $ad|d$ is exactly the degree to which we believe in $a|d$ by itself.[4] Therefore,

$$\xi(a|d)=\xi(ad|d)=G[\xi(a|d),\xi(d|ad)]=G[\xi(a|d),\xi_T]\;\;, \tag{2.64}$$

or,

$$u=A(u)\xi_T\Rightarrow A(u)=\frac{u}{\xi_T}\;. \tag{2.65}$$

Therefore

$$G\,(u,v)=\frac{uv}{\xi_T}\quad\text{or}\quad\frac{\xi\,(ab|d)}{\xi_T}=\frac{\xi\,(a|d)}{\xi_T}\frac{\xi\,(b|ad)}{\xi_T}\;\;. \tag{2.66}$$

The constant ξ_T is easily regraduated away: just normalize ξ to $p=\xi/\xi_T$. The corresponding regraduation of the sum rule, eq.(2.20) is equally trivial. The degrees of belief ξ range from total disbelief $\xi_F=0$ to total certainty ξ_T. The corresponding regraduated values are $p_F=0$ and $p_T=1$.

The main result:

In the regraduated scale the AND operation is represented by a simple product rule,

$$p\,(ab|d)=p\,(a|d)\;p\,(b|ad)\;\;, \tag{2.67}$$

[4]This argument is due to N. Caticha, private communication (2009).

and the OR operation is represented by the sum rule,

$$p\,(a \vee b|d) = p\,(a|d) + p\,(b|d) - p(ab|d)\ . \tag{2.68}$$

Degrees of belief p measured in this particularly convenient regraduated scale will be called "probabilities". The degrees of belief p range from total disbelief $p_F = 0$ to total certainty $p_T = 1$.

Conclusion:

> *A state of partial knowledge —a web of interconnected rational beliefs— is mathematically represented by quantities that are to be manipulated according to the rules of probability theory. Degrees of rational belief are probabilities.*

Other equivalent representations are possible but less convenient; the choice is made on purely pragmatic grounds.

2.4 Some remarks on the sum and product rules

2.4.1 On meaning, ignorance and randomness

The product and sum rules can be used as the starting point for a theory of probability: Quite independently of what probabilities could possibly mean, we can develop a formalism of real numbers (measures) that are manipulated according to eqs.(2.67) and (2.68). This is the approach taken by Kolmogorov. The advantage is mathematical clarity and rigor. The disadvantage, of course, is that in actual applications the issue of meaning, of interpretation, turns out to be important because it affects how and why probabilities are used. It affects how one sets up the equations and it even affects our perception of what counts as a solution.

The advantage of the approach due to Cox is that the issue of meaning is clarified from the start: the theory was designed to apply to degrees of belief. Consistency requires that these numbers be manipulated according to the rules of probability theory. This is all we need. There is no reference to measures of sets or large ensembles of trials or even to random variables. This is remarkable: it means that we can apply the powerful methods of probability theory to thinking and reasoning about problems where nothing random is going on, and to single events for which the notion of an ensemble is either absurd or at best highly contrived and artificial. Thus, probability theory is *the* method for consistent reasoning in situations where the information available might be insufficient to reach certainty: probability is *the* tool for dealing with uncertainty and ignorance.

This interpretation is not in conflict with the common view that probabilities are associated with randomness. It may, of course, happen that there is

an unknown influence that affects the system in unpredictable ways and that there is a good reason why this influence remains unknown, namely, it is so complicated that the information necessary to characterize it cannot be supplied. Such an influence we call 'random'. Thus, being random is just one among many possible reasons why a quantity might be uncertain or unknown.

2.4.2 Independent and mutually exclusive events

In special cases the sum and product rules can be rewritten in various useful ways. Two statements or events a and b are said to be *independent* if the probability of one is not altered by information about the truth of the other. More specifically, event a is independent of b (given c) if

$$p\,(a|bc) = p\,(a|c) \ . \tag{2.69}$$

For independent events the product rule simplifies to

$$p(ab|c) = p(a|c)p(b|c) \quad \text{or} \quad p(ab) = p(a)p(b) \ . \tag{2.70}$$

The symmetry of these expressions implies that $p\,(b|ac) = p\,(b|c)$ as well: if a is independent of b, then b is independent of a.

Two statements or events a_1 and a_2 are *mutually exclusive* given b if they cannot be true simultaneously, i.e., $p(a_1a_2|b) = 0$. Notice that neither $p(a_1|b)$ nor $p(a_2|b)$ need vanish. For mutually exclusive events the sum rule simplifies to

$$p(a_1 + a_2|b) = p(a_1|b) + p(a_2|b). \tag{2.71}$$

The generalization to many mutually exclusive statements $a_1, a_2, \ldots, a_n$ (mutually exclusive given b) is immediate,

$$p(a_1 + a_2 + \cdots + a_n|b) = \sum_{i=1}^{n} p(a_i|b) \ . \tag{2.72}$$

If one of the statements $a_1, a_2, \ldots, a_n$ is necessarily true, i.e., they cover all possibilities, they are said to be *exhaustive*. Then their conjunction is necessarily true, $a_1 + a_2 + \cdots + a_n = \top$, so that for any b,

$$p(\top|b) = p(a_1 + a_2 + \cdots + a_n|b) = 1. \tag{2.73}$$

If, in addition to being exhaustive, the statements $a_1, a_2, \ldots, a_n$ are also mutually exclusive then

$$p(\top) = \sum_{i=1}^{n} p(a_i) = 1 \ . \tag{2.74}$$

A useful generalization involving the probabilities $p(a_i|b)$ conditional on any arbitrary proposition b is

$$\sum_{i=1}^{n} p(a_i|b) = 1 \ . \tag{2.75}$$

The proof is straightforward:

$$p(b) = p(b\top) = \sum_{i=1}^{n} p(ba_i) = p(b)\sum_{i=1}^{n} p(a_i|b) \ . \tag{2.76}$$

2.4.3 Marginalization

Once we decide that it is legitimate to quantify degrees of belief by real numbers p the problem becomes how do we assign these numbers. The sum and product rules show how we should assign probabilities to some statements once probabilities have been assigned to others. Here is an important example of how this works.

We want to assign a probability to a particular statement b. Let $a_1, a_2, \ldots, a_n$ be mutually exclusive and exhaustive statements and suppose that the probabilities of the conjunctions ba_j are known. We want to calculate $p(b)$ given the joint probabilities $p(ba_j)$. The solution is straightforward: sum $p(ba_j)$ over all a_js, use the product rule, and eq.(2.75) to get

$$\sum_j p(ba_j) = p(b)\sum_j p(a_j|b) = p(b) \ . \tag{2.77}$$

This procedure, called marginalization, is quite useful when we want to eliminate uninteresting variables a so we can concentrate on those variables b that really matter to us. The distribution $p(b)$ is referred to as the marginal of the joint distribution $p(ab)$.

For a second use of formulas such as these suppose that we happen to know the conditional probabilities $p(b|a)$. When a is known we can make good inferences about b, but what can we tell about b when we are uncertain about the actual value of a? Then we proceed as follows. Use of the sum and product rules gives

$$p(b) = \sum_j p(ba_j) = \sum_j p(b|a_j)p(a_j) \ . \tag{2.78}$$

This is quite reasonable: the probability of b is the probability we would assign if the value of a were precisely known, averaged over all as. The assignment $p(b)$ clearly depends on how uncertain we are about the value of a. In the extreme case when we are totally certain that a takes the particular value a_k we have $p(a_j) = \delta_{jk}$ and we recover $p(b) = p(b|a_k)$ as expected.

2.5 The expected value

Suppose we know that a quantity x can take values x_i with probabilities p_i. Sometimes we need an estimate for the quantity x. What should we choose? It seems reasonable that those values x_i that have larger p_i should have a dominant contribution to x. We therefore make the following reasonable choice: The expected value of the quantity x is denoted by $\langle x \rangle$ and is given by

$$\langle x \rangle \stackrel{\text{def}}{=} \sum_i p_i \, x_i \ . \tag{2.79}$$

The term 'expected' value is not always an appropriate one because $\langle x \rangle$ may not be one of the actually allowed values x_i and, therefore, it is not a value we

would expect. The expected value of a die toss is $(1 + \cdots + 6)/6 = 3.5$ which is not an allowed result.

Using the average $\langle x \rangle$ as an estimate for the expected value of x is reasonable, but it is also somewhat arbitrary. Alternative estimates are possible; for example, one could have chosen the value for which the probability is maximum — this is called the 'mode'. This raises two questions.

The first question is whether $\langle x \rangle$ is a good estimate. If the probability distribution is sharply peaked all the values of x that have appreciable probabilities are close to each other and to $\langle x \rangle$. Then $\langle x \rangle$ is a good estimate. But if the distribution is broad the actual value of x may deviate from $\langle x \rangle$ considerably. To describe quantitatively how large this deviation might be we need to describe how broad the probability distribution is.

A convenient measure of the width of the distribution is the root mean square (*rms*) deviation defined by

$$\Delta x \stackrel{\text{def}}{=} \left\langle (x - \langle x \rangle)^2 \right\rangle^{1/2}. \tag{2.80}$$

The quantity Δx is also called the standard deviation, its square $(\Delta x)^2$ is called the variance. The term 'variance' may suggest variability or spread but there is no implication that x is necessarily fluctuating or that its values are spread; Δx merely refers to our incomplete knowledge about x.

If $\Delta x \ll \langle x \rangle$ then x will not deviate much from $\langle x \rangle$ and we expect $\langle x \rangle$ to be a good estimate.

The definition of Δx is somewhat arbitrary. It is dictated both by common sense and by convenience. Alternatively we could have chosen to define the width of the distribution as $\langle |x - \langle x \rangle| \rangle$ or $\langle (x - \langle x \rangle)^4 \rangle^{1/4}$ but these definitions are less convenient for calculations.

Now that we have a way of deciding whether $\langle x \rangle$ is a good estimate for x we may raise a second question: Is there such a thing as the "best" estimate for x? Consider another estimate x'. We expect x' to be precise provided the deviations from it are small, i.e., $\langle (x - x')^2 \rangle$ is small. The best x' is that for which its variance is a minimum

$$\frac{d}{dx'} \langle (x - x')^2 \rangle \bigg|_{x'_{\text{best}}} = 0, \tag{2.81}$$

which implies $x'_{\text{best}} = \langle x \rangle$. Conclusion: $\langle x \rangle$ is the best estimate for x when by "best" we mean the estimate with the smallest variance. But other choices are possible, for example, had we actually decided to minimize the width $\langle |x - x'| \rangle$ the best estimate would have been the median, $x'_{\text{best}} = x_m$, a value such that $\text{Prob}(x < x_m) = \text{Prob}(x > x_m) = 1/2$.

We conclude this section by mentioning two important identities that will be repeatedly used in what follows. The first is that the average deviation from the mean vanishes,

$$\langle x - \langle x \rangle \rangle = 0, \tag{2.82}$$

because deviations from the mean are just as likely to be positive and negative. The second useful identity is

$$\left\langle (x - \langle x \rangle)^2 \right\rangle = \langle x^2 \rangle - \langle x \rangle^2. \tag{2.83}$$

The proofs are trivial — just use the definition (2.79).

2.6 The binomial distribution

Suppose the probability of a certain event α is θ. The probability of α not happening is $1 - \theta$. Using the theorems discussed earlier we can obtain the probability that α happens m times in N independent trials. The probability that α happens in the first m trials and not-α or $\tilde{\alpha}$ happens in the subsequent $N - m$ trials is, using the product rule for independent events, $\theta^m(1-\theta)^{N-m}$. But this is only one particular ordering of the m αs and the $(N-m)$ $\tilde{\alpha}$s. There are

$$\frac{N!}{m!(N-m)!} = \binom{N}{m} \tag{2.84}$$

such orderings. Therefore, using the sum rule for mutually exclusive events, the probability of m αs in N independent trials irrespective of the particular order of αs and $\tilde{\alpha}$s is

$$P(m|N,\theta) = \binom{N}{m} \theta^m (1-\theta)^{N-m}. \tag{2.85}$$

This is called the binomial distribution. θ is a parameter that labels the distributions $P(m|N,\theta)$; its interpretation is given by $P(1|1,\theta) = \theta$.

Using the binomial theorem (hence the name of the distribution) one can show these probabilities are correctly normalized:

$$\sum_{m=0}^{N} P(m|N,\theta) = \sum_{m=0}^{N} \binom{N}{m} \theta^m (1-\theta)^{N-m} = (\theta + (1-\theta))^N = 1. \tag{2.86}$$

The range of applicability of this distribution is enormous. Whenever trials are independent of each other (i.e., the outcome of one trial has no influence on the outcome of another, or alternatively, knowing the outcome of one trial provides us with no information about the possible outcomes of another) the distribution is binomial. Independence is the crucial feature.

The expected number of αs is

$$\langle m \rangle = \sum_{m=0}^{N} m\, P(m|N,\theta) = \sum_{m=0}^{N} m \binom{N}{m} \theta^m (1-\theta)^{N-m}.$$

This sum over m is complicated. The following elegant trick is useful. Consider

the sum

$$S(\theta,\phi) = \sum_{m=0}^{N} m \binom{N}{m} \theta^m \phi^{N-m},$$

where θ and ϕ are independent variables. After we calculate S we will replace ϕ by $1-\theta$ to obtain the desired result, $\langle m\rangle = S(\theta, 1-\theta)$. The calculation of S is easy once we realize that $m\,\theta^m = \theta\,\frac{\partial}{\partial\theta}\theta^m$. Then, using the binomial theorem

$$S(\theta,\phi) = \theta\,\frac{\partial}{\partial\theta}\sum_{m=0}^{N}\binom{N}{m}\theta^m\phi^{N-m} = \theta\,\frac{\partial}{\partial\theta}\,(\theta+\phi)^N = N\theta\,(\theta+\phi)^{N-1}\,.$$

Replacing ϕ by $1-\theta$ we obtain our best estimate for the expected number of αs

$$\langle m\rangle = N\theta\,. \tag{2.87}$$

This is the best estimate, but how good is it? To answer we need to calculate Δm. The variance is

$$(\Delta m)^2 = \Big\langle (m-\langle m\rangle)^2 \Big\rangle = \langle m^2\rangle - \langle m\rangle^2,$$

which requires we calculate $\langle m^2\rangle$,

$$\langle m^2\rangle = \sum_{m=0}^{N} m^2 P(m|N,\theta) = \sum_{m=0}^{N} m^2 \binom{N}{m}\theta^m(1-\theta)^{N-m}.$$

We can use the same trick we used before to get $\langle m\rangle$:

$$S'(\theta,\phi) = \sum_{m=0}^{N} m^2\binom{N}{m}\theta^m\phi^{N-m} = \theta\,\frac{\partial}{\partial\theta}\,\left(\theta\,\frac{\partial}{\partial\theta}\,(\theta+\phi)^N\right).$$

Therefore,

$$\langle m^2\rangle = (N\theta)^2 + N\theta(1-\theta), \tag{2.88}$$

and the final result for the *rms* deviation Δm is

$$\Delta m = \sqrt{N\theta\,(1-\theta)}. \tag{2.89}$$

Now we can address the question of how good an estimate $\langle m\rangle$ is. Notice that Δm grows with N. This might seem to suggest that our estimate of m gets worse for large N but this is not quite true because $\langle m\rangle$ also grows with N. The ratio

$$\frac{\Delta m}{\langle m\rangle} = \sqrt{\frac{(1-\theta)}{N\theta}} \propto \frac{1}{N^{1/2}}, \tag{2.90}$$

shows that while both the estimate $\langle m\rangle$ and its uncertainty Δm grow with N, the relative uncertainty decreases.

2.7 Probability vs. frequency: the law of large numbers

It is important to note that the "frequency" $f = m/N$ of αs obtained in one N-trial sequence is not equal to θ. For one given fixed value of θ, the frequency f can take any one of the values $0/N, 1/N, 2/N, \ldots N/N$. What is equal to θ is not the frequency itself but its expected value. Using eq.(2.87),

$$\langle f \rangle = \langle \frac{m}{N} \rangle = \theta \ . \tag{2.91}$$

For large N the distribution is quite narrow and the probability that the observed frequency of αs differs from θ tends to zero as $N \to \infty$. Using eq.(2.89),

$$\Delta f = \Delta \left(\frac{m}{N} \right) = \frac{\Delta m}{N} = \sqrt{\frac{\theta \left(1-\theta\right)}{N}} \propto \frac{1}{N^{1/2}}. \tag{2.92}$$

The same ideas are more precisely conveyed by a theorem due to Bernoulli known as the *law of large numbers*. A simple proof of the theorem involves an inequality due to Tchebyshev. Let $\rho\left(x\right) dx$ be the probability that a variable X lies in the range between x and $x + dx$,

$$P\left(x < X < x + dx\right) = \rho\left(x\right) dx.$$

The variance of X satisfies

$$\left(\Delta x\right)^2 = \int \left(x - \langle x \rangle\right)^2 \rho\left(x\right)\ dx \geq \int_{|x-\langle x\rangle| \geq \varepsilon} \left(x - \langle x \rangle\right)^2 \rho\left(x\right)\ dx\,,$$

where ε is an arbitrary constant. Replacing $\left(x - \langle x \rangle\right)^2$ by its least value ε^2 gives

$$\left(\Delta x\right)^2 \geq \varepsilon^2 \int_{|x-\langle x\rangle| \geq \varepsilon} \rho\left(x\right)\ dx = \varepsilon^2\ P\left(|x - \langle x \rangle| \geq \varepsilon\right),$$

which is Tchebyshev's inequality,

$$P\left(|x - \langle x \rangle| \geq \varepsilon\right) \leq \left(\frac{\Delta x}{\varepsilon} \right)^2 \ . \tag{2.93}$$

Next we prove Bernoulli's theorem. Consider first a special case. Let θ be the probability of outcome α in an experiment E, $P\left(\alpha|E\right) = \theta$. In a sequence of N independent repetitions of E the probability of m outcomes α is binomial. Substituting

$$\langle f \rangle = \theta \quad \text{and} \quad \left(\Delta f\right)^2 = \frac{\theta\left(1-\theta\right)}{N}$$

into Tchebyshev's inequality we get Bernoulli's theorem,

$$P\left(|f - \theta| \geq \varepsilon\,|E^N\right) \leq \frac{\theta\left(1-\theta\right)}{N\varepsilon^2} \ . \tag{2.94}$$

Therefore, the probability that the observed frequency f is appreciably different from θ tends to zero as $N \to \infty$. Or equivalently: for any small ε, the probability that the observed frequency $f = m/N$ lies in the interval between $\theta - \varepsilon/2$ and $\theta + \varepsilon/2$ tends to unity as $N \to \infty$.,

$$\lim_{N\to\infty} P\left(|f-\theta| \leq \varepsilon \,|E^N\right) = 1 \ . \tag{2.95}$$

In the mathematical/statistical literature this result is commonly stated in the form

$$f \longrightarrow \theta \quad \textit{in probability.} \tag{2.96}$$

The qualifying words 'in probability' are crucial: we are not saying that the observed f tends to θ for large N. What vanishes for large N is not the difference $f-\theta$ itself, but rather the *probability* that $|f-\theta|$ is larger than a certain (small) amount.

Thus, probabilities and frequencies are related to each other but they are not the same thing. Since $\langle f\rangle = \theta$, one might perhaps be tempted to define the probability θ in terms of the expected frequency $\langle f\rangle$, but this does not work. The problem is that the notion of expected value presupposes that the concept of probability has already been defined. Defining probability in terms of expected values would be circular.[5] We can express this important point in yet a different way: We cannot define probability as a limiting frequency $\lim_{N\to\infty} f$ because there exists no frequency function $f = f(N)$; the limit makes no sense.

The law of large numbers is easily generalized beyond the binomial distribution. Consider the average

$$x = \frac{1}{N}\sum_{r=1}^{N} x_r \ , \tag{2.97}$$

where $x_1, \ldots, x_N$ are N independent variables with the same mean $\langle x_r\rangle = \mu$ and variance $\mathrm{var}(x_r) = (\Delta x_r)^2 = \sigma^2$. (In the previous discussion leading to eq.(2.94) each variable x_r is either 1 or 0 according to whether outcome α happens or not in the rth repetition of the experiment E.)

To apply Tchebyshev's inequality, eq.(2.93), we need the mean and the variance of x. Clearly,

$$\langle x\rangle = \frac{1}{N}\sum_{r=1}^{N}\langle x_r\rangle = \frac{1}{N}N\mu = \mu \ . \tag{2.98}$$

Furthermore, since the x_r are independent, their variances are additive. For example,

$$\mathrm{var}(x_1 + x_2) = \mathrm{var}(x_1) + \mathrm{var}(x_2) \ . \tag{2.99}$$

(Prove it.) Therefore,

$$\mathrm{var}(x) = \sum_{r=1}^{N}\mathrm{var}(\frac{x_r}{N}) = N\left(\frac{\sigma}{N}\right)^2 = \frac{\sigma^2}{N} \ . \tag{2.100}$$

[5] Expected values can be introduced independently of probability –see [Jeffrey 2004]– but this does not help make probabilities equal to frequencies either.

Tchebyshev's inequality now gives,

$$P\left(|x-\mu| \geq \varepsilon | E^N\right) \leq \frac{\sigma^2}{N\varepsilon^2} \tag{2.101}$$

so that for any $\varepsilon > 0$

$$\lim_{N\to\infty} P\left(|x-\mu| \geq \varepsilon | E^N\right) = 0 \quad \text{or} \quad \lim_{N\to\infty} P\left(|x-\mu| \leq \varepsilon | E^N\right) = 1\,, \tag{2.102}$$

or

$$x \longrightarrow \mu \quad \textit{in probability.} \tag{2.103}$$

Again, what vanishes for large N is not the difference $x-\mu$ itself, but rather the *probability* that $|x-\mu|$ is larger than any given small amount.

2.8 The Gaussian distribution

The Gaussian distribution is quite remarkable, it applies to a wide variety of problems such as the distribution of errors affecting experimental data, the distribution of velocities of molecules in gases and liquids, the distribution of fluctuations of thermodynamical quantities, and so on and on. One suspects that a deeply fundamental reason must exist for its wide applicability. The Central Limit Theorem discussed below provides an explanation.

2.8.1 The de Moivre-Laplace theorem

The Gaussian distribution turns out to be a special case of the binomial distribution. It applies to situations when the number N of trials and the expected number of αs, $\langle m\rangle = N\theta$, are both very large (i.e., N large, θ not too small).

To find an analytical expression for the Gaussian distribution we note that when N is large the binomial distribution,

$$P(m|N,\theta) = \frac{N!}{m!(N-m)!}\,\theta^m(1-\theta)^{N-m},$$

is very sharply peaked: $P(m|N,\theta)$ is essentially zero unless m is very close to $\langle m\rangle = N\theta$. This suggests that to find a good approximation for P we need to pay special attention to a very small range of m. One might be tempted to follow the usual approach and directly expand in a Taylor series but a problem becomes immediately apparent: if a small change in m produces a small change in P then we only need to keep the first few terms, but in our case P is a very sharp function. To reproduce this kind of behavior we need a huge number of terms in the series expansion which is impractical. Having diagnosed the problem one can easily find a cure: instead of finding a Taylor expansion for the rapidly varying P, one finds an expansion for log P which varies much more smoothly.

Let us therefore expand $\log P$ about its maximum at m_0, the location of which is at this point still unknown. The first few terms are

$$\log P = \log P|_{m_0} + \left.\frac{d\log P}{dm}\right|_{m_0}(m-m_0) + \frac{1}{2}\left.\frac{d^2\log P}{dm^2}\right|_{m_0}(m-m_0)^2 + \ldots,$$

where

$$\log P = \log N! - \log m! - \log(N-m)! + m\log\theta + (N-m)\log(1-\theta)\,.$$

What is a derivative with respect to an integer? For large m the function $\log m!$ varies so slowly (relative to the huge value of $\log m!$ itself) that we may consider m to be a continuous variable. Then

$$\frac{d\log m!}{dm} \approx \frac{\log m! - \log(m-1)!}{1} = \log\frac{m!}{(m-1)!} = \log m\ . \tag{2.104}$$

Integrating one obtains a very useful approximation — called the Stirling approximation — for the logarithm of a large factorial

$$\log m! \approx \int_0^m \log x\,dx = (x\log\,x - x)|_0^m = m\log\,m - m\ .$$

A somewhat better expression which includes the next term in the Stirling expansion is

$$\log m! \approx m\log m - m + \frac{1}{2}\log 2\pi m + \ldots \tag{2.105}$$

Notice that the third term is much smaller than the first two: the first two terms are of order m while the last is of order $\log m$. For $m = 10^{23}$, $\log m$ is only 55.3.

The derivatives in the Taylor expansion are

$$\frac{d\log P}{dm} = -\log m + \log(n-m) + \log\theta - \log(1-\theta) = \log\frac{\theta(N-m)}{m(1-\theta)},$$

and

$$\frac{d^2\log P}{dm^2} = -\frac{1}{m} - \frac{1}{N-m} = \frac{-N}{m(N-m)}.$$

To find the value m_0 where P is maximum set $d\log P/dm = 0$. This gives $m_0 = N\theta = \langle m\rangle$, and substituting into the second derivative of $\log P$ we get

$$\left.\frac{d^2\log P}{dm^2}\right|_{\langle m\rangle} = -\frac{1}{N\theta\,(1-\theta)} = -\frac{1}{(\Delta m)^2}.$$

Therefore

$$\log P = \log P\,(\langle m\rangle) - \frac{(m-\langle m\rangle)^2}{2\,(\Delta m)^2} + \ldots$$

or

$$P(m) = P\left(\langle m\rangle\right) \exp\left[-\frac{\left(m-\langle m\rangle\right)^2}{2\left(\Delta m\right)^2}\right].$$

The remaining unknown constant $P\left(\langle m\rangle\right)$ can be evaluated by requiring that the distribution $P(m)$ be properly normalized, that is

$$1 = \sum_{m=0}^{N} P(m) \approx \int_0^N P(x)\,dx \approx \int_{-\infty}^{\infty} P(x)\,dx.$$

Using

$$\int_{-\infty}^{\infty} e^{-\alpha x^2} dx = \sqrt{\frac{\pi}{\alpha}},$$

we get

$$P\left(\langle m\rangle\right) = \frac{1}{\sqrt{2\pi\left(\Delta m\right)^2}}.$$

Thus, the expression for the Gaussian distribution with mean $\langle m\rangle$ and *rms* deviation Δm is

$$P(m) = \frac{1}{\sqrt{2\pi\left(\Delta m\right)^2}} \exp\left[-\frac{\left(m-\langle m\rangle\right)^2}{2\left(\Delta m\right)^2}\right]. \tag{2.106}$$

It can be rewritten as a probability for the frequency $f = m/N$ using $\langle m\rangle = N\theta$ and $\left(\Delta m\right)^2 = N\theta\left(1-\theta\right)$. The probability that f lies in the small range $df = 1/N$ is

$$p(f)df = \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left[-\frac{\left(f-\theta\right)^2}{2\,\sigma_N^2}\right] df\ , \tag{2.107}$$

where $\sigma_N^2 = \theta(1-\theta)/N$.

To appreciate the significance of the theorem consider a macroscopic variable x built up by adding a large number of small contributions, $x = \sum_{n=1}^{N} \xi_n$, where the ξ_n are statistically independent. We assume that each ξ_n takes the value ε with probability θ, and the value 0 with probability $1-\theta$. Then the probability that x takes the value $m\varepsilon$ is given by the binomial distribution $P(m|N,\theta)$. For large N the probability that x lies in the small range $m\varepsilon \pm dx/2$ where $dx = \varepsilon$ is

$$p(x)dx = \frac{1}{\sqrt{2\pi\left(\Delta x\right)^2}} \exp\left[-\frac{\left(x-\langle x\rangle\right)^2}{2\left(\Delta x\right)^2}\right] dx\ , \tag{2.108}$$

where $\langle x\rangle = N\theta\varepsilon$ and $\left(\Delta x\right)^2 = N\theta(1-\theta)\varepsilon^2$. Thus, *the Gaussian distribution arises whenever we have a quantity that is the result of adding a large number of small independent contributions*. The derivation above assumes that the microscopic contributions are discrete (either 0 or ε), and identically distributed but, as shown in the next section, both of these conditions can be relaxed.

2.8.2 The Central Limit Theorem

Consider the average

$$x = \frac{1}{N}\sum_{r=1}^{N} x_r \ , \tag{2.109}$$

of N independent variables $x_1, \ldots, x_N$. Our goal is to calculate the probability of x in the limit of large N. Let $p_r(x_r)$ be the probability distribution for the rth variable with

$$\langle x_r \rangle = \mu_r \quad \text{and} \quad (\Delta x_r)^2 = \sigma_r^2 \ . \tag{2.110}$$

The probability density for x is given by the integral

$$P\,(x) = \int dx_1 \ldots dx_N \ p_1(x_1) \ldots p_N(x_N)\, \delta\left(x - \frac{1}{N}\sum_{r=1}^{N} x_r\right) \ . \tag{2.111}$$

(This is just an exercise in the sum and product rules.) To calculate $P\,(x)$ introduce the averages

$$\bar{\mu} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{r=1}^{N} \mu_r \quad \text{and} \quad \bar{\sigma}^2 \stackrel{\text{def}}{=} \frac{1}{N}\sum_{r=1}^{N} \sigma_r^2 \ , \tag{2.112}$$

and consider the distribution for the variable $x - \bar{\mu}$ which is $\Pr(x - \bar{\mu}) = P(x)$. Its Fourier transform,

$$\begin{aligned} F(k) &= \int dx \ \Pr(x - \bar{\mu}) e^{ik(x-\bar{\mu})} = \int dx\, P(x) e^{ik(x-\bar{\mu})} \\ &= \int dx_1 \ldots dx_N \ p_1(x_1) \ldots p_N(x_N) \ \exp\left[\frac{ik}{N}\sum_{r=1}^{N} (x_r - \mu_r)\right], \end{aligned}$$

can be rearranged into a product

$$F(k) = \left[\int dx_1\, p_1(x_1) e^{i\frac{k}{N}(x_1-\mu_1)}\right] \ldots \left[\int dx_N\, p_N(x_N) e^{i\frac{k}{N}(x_N-\mu_N)}\right] \ . \tag{2.113}$$

The Fourier transform $f(k)$ of a distribution $p(\xi)$ has many interesting and useful properties. For example,

$$f(k) = \int d\xi\, p(\xi) e^{ik\xi} = \left\langle e^{ik\xi} \right\rangle \ , \tag{2.114}$$

and the series expansion of the exponential gives

$$f(k) = \left\langle \sum_{n=0}^{\infty} \frac{(ik\xi)^n}{n!} \right\rangle = \sum_{n=0}^{\infty} \frac{(ik)^n}{n!} \langle \xi^n \rangle \ . \tag{2.115}$$

In words, the coefficients of the Taylor expansion of $f(k)$ give all the moments of $p(\xi)$. The Fourier transform $f(k)$ is called the *moment generating function* and also the *characteristic function* of the distribution.

Going back to our calculation of $P(x)$, eq.(2.111), its Fourier transform, eq.(2.113) is,

$$F(k) = \prod_{r=1}^{N} f_r(\frac{k}{N}) \ , \tag{2.116}$$

where

$$\begin{aligned} f_r(\frac{k}{N}) &= \int dx_r \, p_r(x_r) e^{i\frac{k}{N}(x_r - \mu_r)} \\ &= 1 + i\frac{k}{N}\left\langle x_r - \mu_r \right\rangle - \frac{k^2}{2N^2}\left\langle (x_r - \mu_r)^2 \right\rangle + \dots \\ &= 1 - \frac{k^2\sigma_r^2}{2N^2} + O\left(\frac{k^3}{N^3}\right) \ . \end{aligned} \tag{2.117}$$

For a sufficiently large N this can be written as

$$f_r(\frac{k}{N}) \longrightarrow \exp\left(-\frac{k^2\sigma_r^2}{2N^2}\right) \ . \tag{2.118}$$

so that

$$F(k) = \exp\left(-\frac{k^2}{2N^2}\sum_{r=1}^{N}\sigma_r^2\right) = \exp\left(-\frac{k^2\bar{\sigma}^2}{2N}\right) \ . \tag{2.119}$$

Finally, taking the inverse Fourier transform, we obtain the desired result, which is called the central limit theorem

$$\Pr(x - \bar{\mu}) = P(x) = \frac{1}{\sqrt{2\pi\bar{\sigma}^2/N}} \exp\left(-\frac{(x-\bar{\mu})^2}{2\bar{\sigma}^2/N}\right) \ . \tag{2.120}$$

To conclude we comment on its significance. We have shown that almost independently of the form of the distributions $p_r\,(x_r)$ the distribution of the average x is Gaussian centered at $\bar{\mu}$ with standard deviation $\bar{\sigma}^2/N$. Not only the $p_r\,(x_r)$ need not be binomial, they do not even have to be equal to each other. This helps to explain the widespread applicability of the Gaussian distribution: it applies to almost any 'macro-variable' (such as x) that results from adding a large number of independent 'micro-variables' (such as x_r/N).

But there are restrictions; although very common, Gaussian distributions do not obtain always. A careful look at the derivation above shows the crucial step was taken in eqs.(2.117) and (2.119) where we neglected the contributions of the third and higher moments. Earlier we mentioned that the success of Gaussian distributions is due to the fact that they codify the information that happens to be relevant to the particular phenomenon under consideration. Now we see what that relevant information might be: it is contained in the first two moments, the mean and the variance — Gaussian distributions are successful when third and higher moments are irrelevant. (This can be stated more precisely in terms of the so-called Lyapunov condition.)

Later we shall approach this same problem from the point of view of the method of maximum entropy and there we will show that, indeed, the Gaussian distribution can be derived as the distribution that codifies information about the mean and the variance while remaining maximally ignorant about everything else.

2.9 Updating probabilities: Bayes' rule

Now that we have solved the problem of how to represent a state of knowledge as a consistent web of interconnected beliefs we can address the problem of updating from one consistent web of beliefs to another when new information becomes available. We will only consider those special situations where the information to be processed is in the form of data. The question of what else, beyond data, could possibly qualify as information will be addressed in later chapters.

Specifically the problem is to update our beliefs about θ (either a single parameter or many) on the basis of data x (either a single number or several) and of a known relation between θ and x. The updating consists of replacing the *prior* probability distribution $q(\theta)$ that represents our beliefs before the data is processed, by a *posterior* distribution $p(\theta)$ that applies after the data has been processed.

2.9.1 Formulating the problem

We must first describe the state of our knowledge before the data has been collected or, if the data has already been collected, before we have taken it into account. At this stage of the game not only we do not know θ, we do not know x either. As mentioned above, in order to infer θ from x we must also know how these two quantities are related to each other. Without this information one cannot proceed further. Fortunately we usually know enough about the physics of an experiment that if θ were known we would have a fairly good idea of what values of x to expect. For example, given a value θ for the charge of the electron, we can calculate the velocity x of an oil drop in Millikan's experiment, add some uncertainty in the form of Gaussian noise and we have a very reasonable estimate of the conditional distribution $q(x|\theta)$. The distribution $q(x|\theta)$ is called the *sampling* distribution and also (less appropriately) the *likelihood*. We will assume it is known.

We should emphasize that the crucial information about how x is related to θ is contained in the functional form of the distribution $q(x|\theta)$ —say, whether it is a Gaussian or a Cauchy distribution— and not in the actual values of the arguments x and θ which are, at this point, still unknown.

Thus, to describe the web of prior beliefs we must know the prior $q(\theta)$ and also the sampling distribution $q(x|\theta)$. This means that we must know the full joint distribution,

$$q(\theta, x) = q(\theta)q(x|\theta) \ . \tag{2.121}$$

This is very important: we must be clear about what we are talking about. The relevant universe of discourse is neither the space Θ of possible parameters θ nor the space $\mathcal{X}$ of possible data x. It is rather the product space $\Theta \times \mathcal{X}$ and the probability distributions that concern us are the joint distributions $q(\theta, x)$.

Next we collect data: the observed value turns out to be x'. Our goal is to use this information to update to a web of posterior beliefs represented by a

new joint distribution $p(\theta, x)$. How shall we choose $p(\theta, x)$? The new data tells us that the value of x is now known to be x'. Therefore, the new web of beliefs is constrained to satisfy

$$p(x) = \int d\theta\, p(\theta, x) = \delta(x - x') \ . \tag{2.122}$$

(For simplicity we have here assumed that x is a continuous variable; had x been discrete the Dirac δs would be replaced by Kronecker δs.) This is all we know and it is not sufficient to determine $p(\theta, x)$. Apart from the general requirement that the new web of beliefs must be internally consistent there is nothing in any of our previous considerations that induces us to prefer one consistent web over another. A new principle is needed.

2.9.2 Minimal updating: Bayes' rule

The basic updating principle that we adopt below reflects the conviction that what we have learned in the past, the prior knowledge, is a valuable resource that should not be squandered. Prior beliefs should be revised only to extent that the new information has rendered them obsolete and the updated web of beliefs should coincide with the old one as much as possible. We propose to adopt the following principle of parsimony,

> **Principle of Minimal Updating (PMU)** *The web of beliefs needs to be revised only to the extent required by the new data.*

This seems so reasonable and natural that an explicit statement may seem superfluous. The important point, however, is that *it is not logically necessary.* We could update in many other ways that preserve both internal consistency and consistency with the new information.

As we saw above the new data, eq.(2.122), does not fully determine the joint distribution

$$p(\theta, x) = p(x)p(\theta|x) = \delta(x - x')p(\theta|x) \, . \tag{2.123}$$

All distributions of the form

$$p(\theta, x) = \delta(x - x')p(\theta|x') \ , \tag{2.124}$$

where $p(\theta|x')$ is quite arbitrary, are compatible with the newly acquired data. We still need to assign $p(\theta|x')$. It is at this point that we invoke the PMU. We stipulate that, having updated $q(x)$ to $p(x) = \delta(x - x')$, no further revision is needed and we set

$$p(\theta|x') = q(\theta|x') \ . \tag{(PMU)}$$

Therefore, the web of posterior beliefs is described by

$$p(\theta, x) = \delta(x - x')q(\theta|x') \ . \tag{2.125}$$

To obtain the posterior probability for θ marginalize over x,

$$p(\theta) = \int dx\, p(\theta, x) = \int dx\, \delta(x - x')q(\theta|x') \ , \tag{2.126}$$

to get

$$p(\theta) = q(\theta|x') \ . \tag{2.127}$$

In words, the *posterior probability equals the prior conditional probability* of θ given x'. This result, known as Bayes' rule, is extremely reasonable: we *maintain* those beliefs about θ that are consistent with the data values x' that turned out to be true. Beliefs based on values of x that were not observed are discarded because they are now known to be false. 'Maintain' is the key word: it reflects the PMU in action.

Using the product rule

$$q(\theta, x') = q(\theta)q(x'|\theta) = q(x')q(\theta|x') \ , \tag{2.128}$$

Bayes' rule can be written as

$$p(\theta) = q(\theta)\frac{q(x'|\theta)}{q(x')} \ . \tag{2.129}$$

The interpretation of Bayes' rule is straightforward: according to eq.(2.129) the posterior distribution $p(\theta)$ gives preference to those values of θ that were previously preferred as described by the prior $q(\theta)$, but this is now modulated by the likelihood factor $q(x'|\theta)$ in such a way as to enhance our preference for values of θ that make the observed data more likely, less surprising. The factor in the denominator $q(x')$, which is the prior probability of the data, is given by

$$q(x') = \int q(\theta)q(x'|\theta)\, d\theta \ , \tag{2.130}$$

and plays the role of a normalization constant for the posterior distribution $p(\theta)$. It does not help to discriminate one value of θ from another because it affects all values of θ equally. As we shall see later in this chapter, $q(x')$ turns out to be important in problems of model selection.

Remark: Bayes' rule is often written in the form

$$q(\theta|x') = q(\theta)\frac{q(x'|\theta)}{q(x')} \ , \tag{2.131}$$

and called Bayes' theorem. This formula is very simple; perhaps it is too simple. It is just a restatement of the product rule, eq.(2.128), and therefore it is a simple consequence of the *internal* consistency of the *prior* web of beliefs. The drawback of this formula is that the left hand side is not the *posterior* but rather the *prior conditional* probability; it obscures the fact that an additional principle – the PMU – was needed for updating.

Neither the rule, eq.(6.100), nor the theorem, eq.(2.131), was ever actually written down by Bayes. The person who first explicitly stated the theorem and, more importantly, who first realized its deep significance was Laplace.

Example: Is there life on Mars?

Suppose we are interested in whether there is life on Mars or not. How is the probability that there is life on Mars altered by new data indicating the presence of water on Mars. Let θ = 'There is life on Mars'. The prior information includes the fact I = 'All known life forms require water'. The new data is that x' = 'There is water on Mars'. Let us look at Bayes' rule. We can't say much about $q(x'|I)$ but whatever its value it is definitely less than 1. On the other hand $q(x'|\theta I) \approx 1$. Therefore the factor multiplying the prior is larger than 1. Our belief in the truth of θ is strengthened by the new data x'. This is just common sense, but notice that this kind of probabilistic reasoning cannot be carried out if one adheres to a strictly frequentist interpretation — there is no set of trials. The name 'Bayesian probabilities' given to 'degrees of belief' originates in the fact that it is only under this interpretation that the full power of Bayes' rule can be exploited. Everybody can prove Bayes' theorem; only Bayesians can reap the advantages of Bayes' rule.

Example: Testing positive for a rare disease

Suppose you are tested for a disease, say cancer, and the test turns out to be positive. Suppose further that the test is said to be 99% accurate. Should you panic? It may be wise to proceed with caution.

One should start by explaining that '99% accurate' means that when the test is applied to people known to have cancer the result is positive 99% of the time, and when applied to people known to be healthy, the result is negative 99% of the time. We express this accuracy as $q(y|c) = A = 0.99$ and $q(n|\tilde{c}) = A = 0.99$ (y and n stand for 'positive' and 'negative', c and $\tilde{c}$ stand for 'cancer' or 'no cancer'). There is a 1% probability of false positives, $q(y|\tilde{c}) = 1 - A$, and a 1% probability of false negatives, $q(n|c) = 1 - A$.

On the other hand, what we really want to know is $p(c) = q(c|y)$, the probability of having cancer given that you tested positive. This is not the same as the probability of testing positive given that you have cancer, $q(y|c)$; the two probabilities are not the same thing! So there might be some hope. The connection between what we want, $q(c|y)$, and what we know, $q(y|c)$, is given by Bayes' theorem,

$$q(c|y) = \frac{q(c)q(y|c)}{q(y)} \ .$$

An important virtue of Bayes' rule is that it doesn't just tell you how to process information; it also tells you what information you should seek. In this case one should find $q(c)$, the probability of having cancer irrespective of being tested positive or negative. Suppose you inquire and find that the incidence of cancer in the general population is 1%; this means that $q(c) = 0.01$. Thus,

$$q(c|y) = \frac{q(c)A}{q(y)}$$

One also needs to know $q(y)$, the probability of the test being positive irrespective of whether the person has cancer or not. To obtain $q(y)$ use

$$q(\tilde{c}|y) = \frac{q(\tilde{c})q(y|\tilde{c})}{q(y)} = \frac{(1-q(c))\,(1-A)}{q(y)} \ ,$$

and $q(c|y) + q(\tilde{c}|y) = 1$ which leads to our final answer

$$q(c|y) = \frac{q(c)A}{q(c)A + (1-q(c))\,(1-A)} \ . \tag{2.132}$$

For an accuracy $A = 0.99$ and an incidence $q(c) = 0.01$ we get $q(c|y) = 50\%$ which is not nearly as bad as one might have originally feared. Should one dismiss the information provided by the test as misleading? No. Note that the probability of having cancer prior to the test was 1% and on learning the test result this was raised all the way up to 50%. Note also that when the disease is really rare, $q(c) \to 0$, we still get $q(c|y) \to 0$ even when the test is quite accurate. This means that for rare diseases most positive tests turn out to be false positives.

We conclude that both the prior and the data contain important information; neither should be neglected.

Remark: The previous discussion illustrates a mistake that is common in verbal discussions: if h denotes a hypothesis and e is some evidence, it is quite obvious that we should not confuse $q(e|h)$ with $q(h|e)$. However, when expressed verbally the distinction is not nearly as obvious. For example, in a criminal trial jurors might be told that if the defendant were guilty (the hypothesis) the probability of some observed evidence would be large, and the jurors might easily be misled into concluding that given the evidence the probability is high that the defendant is guilty. Lawyers call this the *prosecutor's fallacy.*

Example: Uncertain data and Jeffrey's rule

As before we want to update from a prior joint distribution $q(\theta, x) = q(x)q(\theta|x)$ to a posterior joint distribution $p(\theta, x) = p(x)p(\theta|x)$ when information becomes available. When the information is data x' that precisely fixes the value of x, we impose that $p(x) = \delta(x - x')$. The remaining unknown $p(\theta|x)$ is determined by invoking the PMU: no further updating is needed. This fixes $p(\theta|x')$ to be the old $q(\theta|x')$ and yields Bayes' rule.

It may happen, however, that there is a measurement error. The data x' that was actually observed does not constrain the value of x completely. To be explicit let us assume that the remaining uncertainty in x is well understood: the observation x' constrains our beliefs about x to a distribution $P_{x'}(x)$ that happens to be known. $P_{x'}(x)$ could, for example, be a Gaussian distribution centered at x', with some known standard deviation σ.

This information is incorporated into the posterior distribution, $p(\theta, x) = p(x)p(\theta|x)$, by imposing that $p(x) = P_{x'}(x)$. The remaining conditional distribution is, as before, determined by invoking the PMU,

$$p(\theta|x) = q(\theta|x) \ , \tag{2.133}$$

and therefore, the joint posterior is

$$p(\theta, x) = P_{x'}(x)q(\theta|x) \ . \tag{2.134}$$

Marginalizing over the uncertain x yields the new posterior for θ,

$$p(\theta) = \int dx \, P_{x'}(x)q(\theta|x) \ . \tag{2.135}$$

This generalization of Bayes' rule is sometimes called Jeffrey's conditionalization rule [Jeffrey 2004].

Incidentally, this is an example of updating that shows that *it is not always the case that information comes purely in the form of data x'.* In the derivation above there clearly is some information in the observed value x' and some information in the particular functional form of the distribution $P_{x'}(x)$, whether it is a Gaussian or some other distribution.

The common element in our previous derivation of Bayes' rule and in the present derivation of Jeffrey's rule is that in both cases the information being processed is a constraint on the allowed posterior marginal distributions $p(x)$. Later we shall see (chapter 5) how the updating rules can be generalized still further to apply to even more general constraints.

2.9.3 Multiple experiments, sequential updating

The problem here is to update our beliefs about θ on the basis of data $x_1, x_2, \ldots$ obtained in a sequence of experiments. The relations between θ and the variables x_i are given through known sampling distributions. We will assume that the experiments are independent but they need not be identical. When the experiments are not independent it is more appropriate to refer to them as being performed is a single more complex experiment the outcome of which is a collection of numbers $\{x_1, \ldots, x_n\}$.

For simplicity we deal with just two identical experiments. The prior web of beliefs is described by the joint distribution,

$$q(x_1, x_2, \theta) = q(\theta)q(x_1|\theta)q(x_2|\theta) = q(x_1)q(\theta|x_1)q(x_2|\theta) \ , \tag{2.136}$$

where we have used independence, $q(x_2|\theta, x_1) = q(x_2|\theta)$.

The first experiment yields the data $x_1 = x_1'$. Bayes' rule gives the updated distribution for θ as

$$p_1(\theta) = q(\theta|x_1') = q(\theta)\frac{q(x_1'|\theta)}{q(x_1')} \ . \tag{2.137}$$

The second experiment yields the data $x_2 = x_2'$ and requires a second application of Bayes' rule. The posterior $p_1(\theta)$ in eq.(2.137) now plays the role of the prior and the new posterior distribution for θ is

$$p_{12}(\theta) = p_1(\theta|x_2') = p_1(\theta)\frac{q(x_2'|\theta)}{p_1(x_2')} \ , \tag{2.138}$$

therefore

$$p_{12}(\theta) \propto q(\theta)q(x'_1|\theta)q(x'_2|\theta) \ . \tag{2.139}$$

We have explicitly followed the update from $q(\theta)$ to $p_1(\theta)$ to $p_{12}(\theta)$. The same result is obtained if the data from both experiments were processed simultaneously,

$$p_{12}(\theta) = q(\theta|x'_1, x'_2) \propto q(\theta)q(x'_1, x'_2|\theta) \ . \tag{2.140}$$

From the symmetry of eq.(2.139) it is clear that the same posterior $p_{12}(\theta)$ is obtained irrespective of the order that the data x'_1 and x'_2 are processed. The commutativity of Bayesian updating follows from the special circumstance that the information conveyed by one experiment does not revise or render obsolete the information conveyed by the other experiment. As we generalize our methods of inference for processing other kinds of information that do interfere with each other (and therefore one may render the other obsolete) we should not expect, much less demand, that commutativity will continue to hold.

2.9.4 Remarks on priors

Let us return to the question of the extent to which probabilities incorporate subjective and objective elements. We have seen that Bayes' rule allows us to update from prior to posterior distributions. The posterior distributions incorporate the presumably objective information contained in the data plus whatever earlier beliefs had been codified into the prior. To the extent that the Bayes updating rule is itself unique one can claim that the posterior is "more objective" than the prior. As we update more and more we should expect that our probabilities should reflect more and more the input data and less and less the original subjective prior distribution. In other words, some subjectivity is unavoidable at the beginning of an inference chain, but it can be gradually suppressed as more and more information is processed.

The problem of choosing the first prior in the inference chain is a difficult one. We will tackle it in several different ways. Later in this chapter, as we introduce some elementary notions of data analysis, we will address it in the standard way: just make a "reasonable" guess — whatever that might mean. With experience and intuition this seems to work well. But when addressing new problems we have neither experience nor intuition and guessing is risky. We would like to develop more systematic ways to proceed. Indeed it can be shown that certain types of prior information (for example, symmetries and/or other constraints) can be objectively translated into a prior once we have developed the appropriate tools — entropy and geometry. (See *e.g.* [Caticha Preuss 2004] and references therein.)

Our immediate goal here is, first, to remark on the dangerous consequences of extreme degrees of belief, and then to prove our previous intuitive assertion that the accumulation of data will swamp the original prior and render it irrelevant.

Dangerous extremes: the prejudiced mind

The consistency of Bayes' rule can be checked for the extreme cases of certainty and impossibility: Let B describe any background information. If $q\,(\theta|B) = 1$, then $\theta B = B$ and $q(x|\theta B) = q(x|B)$, so that Bayes' rule gives

$$p(\theta|B) = q(\theta|B)\frac{q(x|\theta B)}{q(x|B)} = 1\ . \tag{2.141}$$

A similar argument can be carried through in the case of impossibility: If $q\,(\theta|B) = 0$, then $p\,(\theta|B) = 0$. The conclusion is that if we are absolutely certain about the truth of θ, acquiring data x will have absolutely no effect on our opinions; the new data is worthless.

This should serve as a warning to the dangers of erroneously assigning a probability of 1 or of 0: since no amount of data could sway us from our prior beliefs we may decide we did not need to collect the data in the first place. If you are absolutely sure that Jupiter has no moons, you may either decide that it is not necessary to look through the telescope, or, if you do look and you see some little bright spots, you will probably decide the spots are mere optical illusions. Extreme degrees of belief are dangerous: a truly prejudiced mind does not, and indeed, *cannot* question its own beliefs.

Lots of data overwhelms the prior

As more and more data is accumulated according to the sequential updating described earlier one would expect that the continuous inflow of information will eventually render irrelevant whatever prior information we might have had at the start. We will now show that this is indeed the case: unless we have assigned a pathological prior after a large number of experiments the posterior becomes essentially independent of the prior.

Consider N independent repetitions of a certain experiment that yield the data $x = \{x_1 \ldots x_N\}$. (For simplicity we omit all primes on the observed data.) The corresponding likelihood is

$$q(x|\theta) = \prod_{r=1}^{N} q(x_r|\theta)\ , \tag{2.142}$$

and the posterior distribution $p(\theta)$ is

$$p(\theta|x) = \frac{q(\theta)}{q(X)}q(x|\theta) = \frac{q(\theta)}{q(x)}\prod_{r=1}^{N} q(x_r|\theta)\ . \tag{2.143}$$

To investigate the extent to which the data x supports a particular value θ_1 rather than any other value θ_2 it is convenient to study the ratio

$$\frac{p(\theta_1|x)}{p(\theta_2|x)} = \frac{q(\theta_1)}{q(\theta_2)}R(x)\ , \tag{2.144}$$

where we introduced the likelihood ratios

$$R(x) \stackrel{\text{def}}{=} \prod_{r=1}^{N} R_r(x_r) \quad \text{and} \quad R_r(x_r) \stackrel{\text{def}}{=} \frac{q(x_r|\theta_1)}{q(x_r|\theta_2)} \ . \tag{2.145}$$

We want to prove the following theorem: Barring two trivial exceptions, for any arbitrarily large positive Λ, we have

$$\lim_{N\to\infty} P\left(R(x) > \Lambda|\theta_1\right) = 1 \tag{2.146}$$

or, in other words,

$$\text{given } \theta_1, \quad R(x) \longrightarrow \infty \quad \textit{in probability.} \tag{2.147}$$

The significance of the theorem is that as data accumulates a rational agent becomes more and more convinced of the truth — in this case the true value is θ_1 — and this happens essentially irrespective of the prior $p(\theta)$.

The theorem fails in two cases: first, when the prior $q(\theta_1)$ vanishes, and second, when $q(x_r|\theta_1) = q(x_r|\theta_2)$ for all x_r which means that the experiment was poorly designed because it cannot distinguish between θ_1 and θ_2.

The proof of the theorem is an application of the law of large numbers. Consider the quantity

$$\frac{1}{N}\log R(x) = \frac{1}{N}\sum_{r=1}^{N} \log R_r(x_r) \ . \tag{2.148}$$

Since the variables $\log R_r(x_r)$ are independent, eq.(2.102) gives

$$\lim_{N\to\infty} P\left(\left|\frac{1}{N}\log R(x) - K(\theta_1,\theta_2)\right| \leq \varepsilon|\theta_1\right) = 1 \tag{2.149}$$

where ε is any small positive number and

$$\begin{aligned} K(\theta_1,\theta_2) &= \left\langle \frac{1}{N}\log R(x)|\theta_1 \right\rangle \\ &= \sum_{x_r} q(x_r|\theta_1)\log R_r(x_r) \ . \end{aligned} \tag{2.150}$$

In other words,

$$\text{given } \theta_1, \quad e^{N(K-\varepsilon)} \leq R(x) \leq e^{N(K+\varepsilon)} \quad \textit{in probability.} \tag{2.151}$$

In Chapter 4 we will prove the Gibbs inequality, $K(\theta_1,\theta_2) \geq 0$. The equality holds if and only if the two distributions $q(x_r|\theta_1)$ and $q(x_r|\theta_2)$ are identical, which is precisely the second of the two trivial exceptions we explicitly avoid. Thus $K(\theta_1,\theta_2) > 0$, and this concludes the proof.

We see here the first appearance of a quantity,

$$K(\theta_1,\theta_2) = +\sum_{x_r} q(x_r|\theta_1)\log\frac{q(x_r|\theta_1)}{q(x_r|\theta_2)} \ , \tag{2.152}$$

that will prove to be central in later discussions. When multiplied by -1, the quantity $-K(\theta_1, \theta_2)$ is called the *relative entropy*,[6] that is the entropy of $q(x_r|\theta_1)$ *relative* to $q(x_r|\theta_2)$. It can be interpreted as a measure of the extent that the distribution $q(x_r|\theta_1)$ can be distinguished from $q(x_r|\theta_2)$.

2.10 Hypothesis testing and confirmation

The basic goal of statistical inference is to update our opinions about the truth of a particular theory or hypothesis θ on the basis of evidence provided by data E. The update proceeds according to Bayes rule,[7]

$$p(\theta|E) = p(\theta)\frac{p(E|\theta)}{p(E)} \ , \tag{2.153}$$

and one can say that the hypothesis θ is confirmed or corroborated by the evidence E when $p(\theta|E) > p(\theta)$.

Sometimes one wishes to compare two hypothesis, θ_1 and θ_2, and the comparison is conveniently done using the ratio

$$\frac{p(\theta_1|E)}{p(\theta_2|E)} = \frac{p(\theta_1)}{p(\theta_2)}\frac{p(E|\theta_1)}{p(E|\theta_2)} \ . \tag{2.154}$$

The relevant quantity is the "likelihood ratio" or "Bayes factor"

$$R(\theta_1 : \theta_2) \stackrel{\text{def}}{=} \frac{p(E|\theta_1)}{p(E|\theta_2)} \ . \tag{2.155}$$

When $R(\theta_1 : \theta_2) > 1$ one says that the evidence E provides support in favor of θ_1 against θ_2.

The question of the testing or confirmation of a hypothesis is so central to the scientific method that it pays to explore it. First we introduce the concept of weight of evidence, a variant of the Bayes factor, that has been found particularly useful in such discussions. Then, to explore some of the subtleties and potential pitfalls we discuss the paradox associated with the name of Hempel.

Weight of evidence

A useful variant of the Bayes factor is its logarithm,

$$w(\theta_1 : \theta_2) \stackrel{\text{def}}{=} \log \frac{p(E|\theta_1)}{p(E|\theta_2)} \ . \tag{2.156}$$

[6] Other names include relative information, directed divergence, and Kullback-Leibler distance.

[7] From here on we revert to the usual notation p for probabilities. Whether p refers to a prior or a posterior will, as is usual in this field, have to be inferred from the context.

This is called the *weight of evidence* for θ_1 against θ_2 [Good 1950].[8] A useful special case is when the second hypothesis θ_2 is the negation of the first. Then

$$w(\theta : E) \stackrel{\text{def}}{=} \log \frac{p(E|\theta)}{p(E|\tilde{\theta})} \ , \tag{2.157}$$

is called the *weight of evidence in favor of the hypothesis* θ *provided by the evidence* E. The change to a logarithmic scale is convenient because it confers useful additive properties upon the weight of evidence — which justifies calling it a 'weight'. Consider, for example, the odds in favor of θ given by the ratio

$$\text{Odds}(\theta) \stackrel{\text{def}}{=} \frac{p(\theta)}{p(\tilde{\theta})} \ . \tag{2.158}$$

The posterior and prior odds are related by

$$\frac{p(\theta|E)}{p(\tilde{\theta}|E)} = \frac{p(\theta)}{p(\tilde{\theta})} \frac{p(E|\theta)}{p(E|\tilde{\theta})} \ , \tag{2.159}$$

and taking logarithms we have

$$\log \text{Odds}(\theta|E) = \log \text{Odds}(\theta) + w(\theta : E) \ . \tag{2.160}$$

The weight of evidence can be positive and confirm the hypothesis by increasing its odds, or it can be negative and refute it. Furthermore, when we deal with two pieces of evidence and E consists of E_1 and E_2, we have

$$\log \frac{p(E_1 E_2|\theta)}{p(E_1 E_2|\tilde{\theta})} = \log \frac{p(E_1|\theta)}{p(E_1|\tilde{\theta})} + \log \frac{p(E_2|E_1\theta)}{p(E_2|E_1\tilde{\theta})}$$

so that

$$w(\theta : E_1 E_2) = w(\theta : E_1) + w(\theta : E_2|E_1) \ . \tag{2.161}$$

Hempel's paradox

Here is the paradox: "A case of a hypothesis supports the hypothesis. Now, the hypothesis that all crows are black is logically equivalent to the contrapositive that all non-black things are non-crows, and this is supported by the observation of a white shoe." [Hempel 1967]

The premise that "a case of a hypothesis supports the hypothesis" seems reasonable enough. After all, how else but by observing black crows can one ever expect to confirm that "all crows are black"? But to assert that a white shoe confirms that all crows are black seems a bit too much. If so then the very same white shoe would equally well confirm the hypotheses that all crows are green, or that all swans are black. We have a paradox.

[8] According to [Good 1983] the concept was known to H. Jeffreys and A. Turing around 1940-41 and C. S. Peirce had proposed the name weight of evidence for a similar concept already in 1878.

Let us consider the starting premise that the observation of a black crow supports the hypothesis $\theta =$ "All crows are black" more carefully. Suppose we observe a crow (C) and it turns out to be black (B). The evidence is $E = B|C$, and the corresponding weight of evidence is positive,

$$w(\theta : B|C) = \log \frac{p(B|C\theta)}{p(B|C\tilde{\theta})} = \log \frac{1}{p(B|C\tilde{\theta})} \geq 0 \ , \tag{2.162}$$

as expected. It is this result that justifies our intuition that "a case of a hypothesis supports the hypothesis"; the question is whether there are limitations. [Good 1983]

The reference to the possibility of white shoes points to an uncertainty about whether the observed object will turn out to be a crow or something else. Then the relevant weight of evidence concerns the joint probability of B and C,

$$w(\theta : BC) = w(\theta : C) + w(\theta : B|C) \ , \tag{2.163}$$

which is also positive because the second term on the right is positive while the first vanishes. Indeed, using Bayes' theorem,

$$w(\theta : C) = \log \frac{p(C|\theta)}{p(C|\tilde{\theta})} = \log \left(\frac{p(C)p(\theta|C)}{p(\theta)} \frac{p(\tilde{\theta})}{p(C)p(\tilde{\theta}|C)} \right) \ . \tag{2.164}$$

Now, *in the absence of any background information about crows* the observation that a certain object turns out to be a crow tells us nothing about its color and therefore $p(\theta|C) = p(\theta)$ and $p(\tilde{\theta}|C) = p(\tilde{\theta})$. Therefore

$$w(\theta : C) = 0 \quad \text{so that} \quad w(\theta : BC) \geq 0 \ . \tag{2.165}$$

A similar conclusion holds if the evidence consists in the observation of a white shoe. Does a non-black non-crow support all crows are black? In this case

$$w(\theta : \tilde{B}\tilde{C}) = w(\theta : \tilde{B}) + w(\theta : \tilde{C}|\tilde{B}) \geq 0$$

because

$$w(\theta : \tilde{C}|\tilde{B}) = \log \frac{p(\tilde{C}|\tilde{B}\theta)}{p(\tilde{C}|\tilde{B}\tilde{\theta})} = \log \frac{1}{p(\tilde{C}|\tilde{B}\tilde{\theta})} \geq 0 \tag{2.166}$$

while

$$w(\theta : \tilde{B}) = \log \frac{p(\tilde{B}|\theta)}{p(\tilde{B}|\tilde{\theta})} = \log \left(\frac{p(\tilde{B})p(\theta|\tilde{B})}{p(\theta)} \frac{p(\tilde{\theta})}{p(\tilde{B})p(\tilde{\theta}|\tilde{B})} \right) = 0 \ . \tag{2.167}$$

Indeed, just as before *in the absence of any background information about crows* the observation of some non-black object tells us nothing about crows, thus $p(\theta|\tilde{B}) = p(\theta)$.

But it is quite conceivable that additional background information that establishes a connection between θ and C is available. One possible scenario is the

following: There are two worlds. In one world, denoted θ_1 there are a million birds of which one hundred are crows and all of them are black; in the other world, denoted θ_2, there also are a million birds among which there is one white and 999 black crows. We pick a bird at random and it turns out to be a black crow. Which world is it, θ_1 or $\theta_2 = \tilde{\theta}_1$? The weight of evidence is

$$w(\theta_1 : BC) = w(\theta_1 : C) + w(\theta_1 : B|C) \ .$$

The relevant probabilities are $p(B|C\theta_1) = 1$ and $p(B|C\theta_2) = 0.999$. Therefore

$$w(\theta_1 : B|C) = \log \frac{p(B|C\theta_1)}{p(B|C\theta_2)} = \log \frac{1}{1 - 10^{-3}} \approx 10^{-3} \tag{2.168}$$

while $p(C|\theta_1) = 10^{-4}$ and $p(C|\theta_2) = 10^{-3}$ so that

$$w(\theta_1 : C) = \log \frac{p(C|\theta_1)}{p(C|\theta_2)} = \log 10^{-1} \approx -2.303 \ . \tag{2.169}$$

Therefore $w(\theta_1 : BC) = -2.302 < 0$. In this scenario the observation of a black crow is evidence for the opposite conclusion that not all crows are black.

We conclude that just like any other form of induction the principle that "a case of a hypothesis supports the hypothesis" involves considerable risk. Whether it is justified or not depends to a large extent on the nature of the available background information. When confronted with a situation in which we are completely ignorant about the relation between two variables the prudent way to proceed is, of course, to try to find out whether a relevant connection exists and what it might be. But this is not always possible and in these cases the default assumption should be that they are a priori independent.

The justification the assumption of independence a priori is purely pragmatic. Indeed the universe contains an infinitely large number of other variables about which we know absolutely nothing and that could in principle affect our inferences. Seeking information about all those other variables is clearly out of the question: waiting to make an inference until after all possible information has been collected amounts to being paralyzed into making no inferences at all. On the positive side, however, the assumption that the vast majority of those infinitely many other variables are completely irrelevant actually works — at least most of the time.

There is one final loose end that we must revisit: our arguments above indicate that, in the absence of any other background information, the observation of a white shoe not only supports the hypothesis that "all crows are black", but it also supports the hypothesis that "all swans are black". Two questions arise: is this reasoning correct? and, if so, why is it so disturbing? The answer to the first question is that it is indeed correct. The answer to the second question is that confirming the hypothesis "all swans are black" is disturbing because we happen to have background information about the color of swans which we failed to include in the analysis. Had we not known anything about swans there would have been no reason to feel any discomfort at all. This is just one more example of the fact that inductive arguments are not infallible; a positive weight of evidence provides mere support not absolute certainty.

2.11 Examples from data analysis

To illustrate the use of Bayes' theorem as a tool to process information when the information is in the form of data we consider some elementary examples from the field of data analysis. (For more detailed treatments that are friendly to physicists see e.g. [Bretthorst 1988, Sivia Skilling 2006, Gregory 2005].)

2.11.1 Parameter estimation

Suppose the probability for the quantity x depends on certain parameters θ, $p = p(x|\theta)$. Although most of the discussion here can be carried out for an arbitrary function p it is best to be specific and focus on the important case of a Gaussian distribution,

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \ . \tag{2.170}$$

The objective is to estimate the parameters $\theta = (\mu, \sigma)$ on the basis of a set of data $x = (x_1, \ldots x_N)$. We assume the measurements are statistically independent of each other and use Bayes' theorem to get

$$p(\mu,\sigma|x) = \frac{p(\mu,\sigma)}{p\,(X)} \prod_{i=1}^{N} p(x_i|\mu,\sigma) \ . \tag{2.171}$$

Independence is important in practice because it leads to considerable practical simplifications but it is not essential: instead of N independent measurements each providing a single datum we would have a single complex experiment that provides N non-independent data.

Looking at eq.(2.171) we see that a more precise formulation of the same problem is the following. We want to estimate certain parameters θ, in our case μ and σ, from repeated measurements of the quantity x on the basis of *several* pieces of information. The most obvious is

1. The information contained in the actual values of the collected data x.

Almost equally obvious (at least to those who are comfortable with the Bayesian interpretation of probabilities) is

2. The information about the parameters that is codified into the prior distribution $p(\theta)$.

Where and how this prior information was obtained is not relevant at this point; it could have resulted from previous experiments, or from other background knowledge about the problem. The only relevant part is whatever ended up being distilled into $p(\theta)$.

The last piece of information is not always explicitly recognized; it is

3. The information that is codified into the functional form of the 'sampling' distribution $p(x|\theta)$.

If we are to estimate parameters θ on the basis of measurements of a quantity x it is clear that we must know how θ and x are related to each other. Notice that item 3 refers to the *functional form* – whether the distribution is Gaussian as opposed to Poisson or binomial or something else – and not to the actual values of the data x which is what is taken into account in item 1. The nature of the relation in $p(x|\theta)$ is in general statistical but it could also be completely deterministic. For example, when x is a known function of θ, say $x = f(\theta)$, we have $p(x|\theta) = \delta\left[x - f(\theta)\right]$. In this latter case there is no need for Bayes' rule.

Eq. (2.171) is rewritten as

$$p(\mu, \sigma|x) = \frac{p(\mu, \sigma)}{p\left(x\right)} \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left[-\sum_{i=1}^{N} \frac{\left(x_i - \mu\right)^2}{2\sigma^2}\right] \tag{2.172}$$

Introducing the sample average $\bar{x}$ and sample variance s^2,

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad s^2 = \frac{1}{N}\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2 \ , \tag{2.173}$$

eq.(2.172) becomes

$$p(\mu, \sigma|x) = \frac{p(\mu, \sigma)}{p\left(x\right)} \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left[-\frac{\left(\mu - \bar{x}\right)^2 + s^2}{2\sigma^2/N}\right] \ . \tag{2.174}$$

It is interesting that the data appears here only in the particular combination in eq.(2.173) – different sets of data characterized by the same $\bar{x}$ and s^2 lead to the same inference about μ and σ. (As discussed earlier the factor $p\left(x\right)$ is not relevant here since it can be absorbed into the normalization of the posterior $p(\mu, \sigma|x)$.)

Eq. (2.174) incorporates the information described in items 1 and 3 above. The prior distribution, item 2, remains to be specified. Let us start by considering the simple case where the value of σ is actually known. Then $p(\mu, \sigma) = p(\mu)\delta(\sigma - \sigma_0)$ and the goal is to estimate μ. Bayes' theorem is now written as

$$p(\mu|x) = \frac{p(\mu)}{p\left(x\right)} \frac{1}{\left(2\pi\sigma_0^2\right)^{N/2}} \exp\left[-\sum_{i=1}^{N} \frac{\left(x_i - \mu\right)^2}{2\sigma_0^2}\right] \tag{2.175}$$

$$= \frac{p(\mu)}{p\left(x\right)} \frac{1}{\left(2\pi\sigma_0^2\right)^{N/2}} \exp\left[-\frac{\left(\mu - \bar{x}\right)^2 + s^2}{2\sigma_0^2/N}\right]$$

$$\propto p(\mu) \exp\left[-\frac{\left(\mu - \bar{x}\right)^2}{2\sigma_0^2/N}\right] \ . \tag{2.176}$$

Suppose further that we know nothing about μ; it could have any value. This state of extreme ignorance is represented by a very broad distribution that we take as essentially uniform within some large range; μ is just as likely to have one value as another. For $p(\mu) \sim$ const the posterior distribution is Gaussian, with

mean given by the sample average $\bar{x}$, and variance σ_0^2/N. The best estimate for the value of μ is the sample average and the uncertainty is the standard deviation. This is usually expressed in the form

$$\mu = \bar{x} \pm \frac{\sigma_0}{\sqrt{N}} \ . \tag{2.177}$$

Note that the estimate of μ from N measurements has a much smaller error than the estimate from just one measurement; the individual measurements are plagued with errors but they tend to cancel out in the sample average.

In the case of very little prior information — the uniform prior — we have recovered the same results as in the standard non-Bayesian data analysis approach. The real difference arises when prior information is available: the non-Bayesian approach can't deal with it and can only proceed by ignoring it. On the other hand, within the Bayesian approach prior information is easily taken into account. For example, if we know on the basis of other physical considerations that μ has to be positive we assign $p(\mu) = 0$ for $\mu < 0$ and we calculate the estimate of μ from the truncated Gaussian in eq.(2.176).

A slightly more complicated case arises when the value of σ is not known. Let us assume again that our ignorance of both μ and σ is quite extreme and choose a uniform prior,

$$p(\mu, \sigma) \propto \begin{cases} C & \text{for} \quad \sigma > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.178}$$

Another popular choice is a prior that is uniform in μ and in $\log \sigma$. When there is a considerable amount of data the two choices lead to practically the same conclusions but we see that there is an important question here: what do we mean by the word 'uniform'? Uniform in terms of which variable? σ, or σ^2, or $\log \sigma$? Later we shall have much more to say about this misleadingly innocuous question.

To estimate μ we return to eq.(2.172) or (2.174). For the purpose of estimating μ the variable σ is an uninteresting nuisance which, as discussed in section 2.5.4, is eliminated through marginalization,

$$p(\mu|x) = \int_0^\infty d\sigma \, p(\mu, \sigma|x) \tag{2.179}$$

$$\propto \int_0^\infty d\sigma \, \frac{1}{\sigma^N} \exp\left[-\frac{(\mu - \bar{x})^2 + s^2}{2\sigma^2/N} \right] \ . \tag{2.180}$$

Change variables to $t = 1/\sigma$, then

$$p(\mu|x) \propto \int_0^\infty dt \, t^{N-2} \exp\left[-\frac{t^2}{2} N \left((\mu - \bar{x})^2 + s^2 \right) \right] \ . \tag{2.181}$$

Repeated integrations by parts lead to

$$p(\mu|x) \propto \left[N \left((\mu - \bar{x})^2 + s^2 \right) \right]^{-\frac{N-1}{2}} \ , \tag{2.182}$$

which is called the *Student-t* distribution. Since the distribution is symmetric the estimate for μ is easy to get,

$$\langle \mu \rangle = \bar{x} \ . \tag{2.183}$$

The posterior $p(\mu|x)$ is a Lorentzian-like function raised to some power. As the number of data grows, say $N \gtrsim 10$, the tails of the distribution are suppressed and $p(\mu|x)$ approaches a Gaussian. To obtain an error bar in the estimate $\mu = \bar{x}$ we can estimate the variance of μ using the following trick. Note that for the Gaussian in eq.(2.170),

$$\frac{d^2}{dx^2} \log \, p(x|\mu,\sigma)\Big|_{x_{\max}} = -\frac{1}{\sigma^2} \ . \tag{2.184}$$

Therefore, to the extent that eq.(2.182) approximates a Gaussian, we can write

$$(\Delta\mu)^2 \approx \left[-\frac{d^2}{d\mu^2} \log \, p(\mu|x)\Big|_{\mu_{\max}} \right]^{-1} = \frac{s^2}{N-1} \ . \tag{2.185}$$

(This explains the famous factor of $N-1$. As we can see it is not a particularly fundamental result; it follows from approximations that are meaningful only for large N.)

We can also estimate σ directly from the data. This requires that we marginalize over μ,

$$p(\sigma|x) = \int_{-\infty}^{\infty} d\mu \, p(\mu,\sigma|x) \tag{2.186}$$

$$\propto \frac{1}{\sigma^N} \exp\left[-\frac{N s^2}{2\sigma^2} \right] \int_{-\infty}^{\infty} d\mu \, \exp\left[-\frac{(\mu-\bar{x})^2}{2\sigma^2/N} \right] \ . \tag{2.187}$$

The Gaussian integral over μ is $\left(2\pi\sigma^2/N\right)^{1/2} \propto \sigma$ and therefore

$$p(\sigma|X) \propto \frac{1}{\sigma^{N-1}} \exp\left[-\frac{N s^2}{2\sigma^2} \right] \ . \tag{2.188}$$

As an estimate for σ we can use the value where the distribution is maximized,

$$\sigma_{\max} = \sqrt{\frac{N}{N-1} s^2} \ , \tag{2.189}$$

which agrees with our previous estimate of $(\Delta\mu)^2$,

$$\frac{\sigma_{\max}^2}{N} = \frac{s^2}{N-1} \ . \tag{2.190}$$

An error bar for σ itself can be obtained using the previous trick (provided N is large enough) of taking a second derivative of $\log p$. The result is

$$\sigma = \sigma_{\max} \pm \frac{\sigma_{\max}}{\sqrt{2\,(N-1)}} \ . \tag{2.191}$$

2.11.2 Curve fitting

The problem of fitting a curve to a set of data points is a problem of parameter estimation. There are no new issues of principle to be resolved. In practice, however, it can be considerably more complicated than the simple cases discussed in the previous paragraphs.

The problem is as follows. The observed data is in the form of pairs (x_i, y_i) with $i = 1, \ldots N$ and we believe that the true ys are related to the xs through a function $y = f_\theta(x)$ which depends on several parameters θ. The goal is to estimate the parameters θ and the complication is that the measured values of y are afflicted by experimental errors,

$$y_i = f_\theta(x_i) + \varepsilon_i \ . \tag{2.192}$$

For simplicity we assume that the probability of the error ε_i is Gaussian with mean $\langle \varepsilon_i \rangle = 0$ and that the variances $\langle \varepsilon_i^2 \rangle = \sigma^2$ are known and the same for all data pairs. We also assume that there are no errors affecting the xs. A more realistic account might have to reconsider these assumptions.

The sampling distribution is

$$p(y|\theta) = \prod_{i=1}^{N} p(y_i|\theta) \ , \tag{2.193}$$

where

$$p(y_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f_\theta(x_i))^2}{2\sigma^2} \right) \ . \tag{2.194}$$

Bayes' theorem gives,

$$p(\theta|y) \propto p(\theta) \exp\left[-\sum_{i=1}^{N} \frac{(y_i - f_\theta(x_i))^2}{2\sigma^2} \right] \ . \tag{2.195}$$

As an example, suppose that we are trying to fit a straight line through data points

$$f(x) = a + bx \ , \tag{2.196}$$

and suppose further that we are quite ignorant about the values of $\theta = (a, b)$ and $p(\theta) = p(a, b) \sim \text{const}$, then

$$p(a, b|y) \propto \exp\left[-\sum_{i=1}^{N} \frac{(y_i - a - bx_i)^2}{2\sigma^2} \right] \ . \tag{2.197}$$

A good estimate of a and b is the value that maximizes the posterior distribution, which as we see, is equivalent to using the method of least squares. But this Bayesian analysis,simple as it is, can already give us more: from $p(a, b|Y)$ we can also estimate the uncertainties Δa and Δb which lies beyond the scope of least squares.

2.11.3 Model selection

Suppose we are trying to fit a curve $y = f_\theta(x)$ through data points (x_i, y_i), $i = 1, \ldots N$. How do we choose the function f_θ? To be specific let f_θ be a polynomial of order n,

$$f_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_n x^n \ , \tag{2.198}$$

the techniques of the previous section allow us to estimate the parameters $\theta_0, \ldots, \theta_n$ but how do we decide the order n? Should we fit a straight or a quadratic line? It is not obvious. Having more parameters means that we will be able to achieve a closer fit to the data, which is good, but we might also be fitting the noise, which is bad. The same problem arises when the data shows peaks and we want to estimate their location, their width, and *their number*; could there be an additional peak hiding in the noise? Are we just fitting noise, or does the data really support one additional peak?

We say these are problems of model selection. To appreciate how important they can be consider replacing the modestly unassuming word 'model' by the more impressive sounding word 'theory'. Given two competing theories, which one does the data support best? What is at stake is nothing less than the foundation of experimental science.

On the basis of data x we want to select one model among several competing candidates labeled by $m = 1, 2, \ldots$ Suppose model m is defined in terms of some parameters $\theta_m = \{\theta_{m1}, \theta_{m2}, \ldots\}$ and their relation to the data x is contained in the sampling distribution $p(x|m, \theta_m)$. The extent to which the data supports model m, *i.e.*, the probability of model m given the data, is given by Bayes' theorem,

$$p(m|x) = \frac{p(m)}{p(x)} p(x|m) \ , \tag{2.199}$$

where $p(m)$ is the prior for the model. The factor $p(x|m)$, which is the *prior probability for the data* given the model, and plays the role of a *likelihood function* is often called the 'evidence'. The evidence is calculated from

$$p(x|m) = \int d\theta_m \ p(x, \theta_m|m) = \int d\theta_m \ p(\theta_m|m) \ p(x|m, \theta_m) \ . \tag{2.200}$$

Therefore

$$p(m|x) \propto p(m) \int d\theta_m \ p(\theta_m|m) p(x|m, \theta_m) \ . \tag{2.201}$$

Thus, the problem is solved, at least in principle, once the priors $p(m)$ and $p(\theta_m|m)$ are assigned. Of course, the practical problem of calculating the multi-dimensional integrals can be quite formidable.

No further progress is possible without making specific choices for the various functions in eq.(2.201) but we can offer some qualitative comments. When comparing two models, m_1 and m_2, it is fairly common to argue that a priori we have no reason to prefer one over the other and therefore we assign the same prior probability $p(m_1) = p(m_2)$. (Of course this is not always justified.

Particularly in the case of theories that claim to be fundamental people usually have very strong prior prejudices favoring one theory against the other. Be that as it may, let us proceed.)

Suppose the prior $p(\theta_m|m)$ represents a uniform distribution over the parameter space. Since

$$\int d\theta_m \; p(\theta_m|m) = 1 \quad \text{then} \quad p(\theta_m|m) \approx \frac{1}{V_m} \; , \tag{2.202}$$

where V_m is the 'volume' of the parameter space. Suppose further that $p(x|m, \theta_m)$ has a single peak of height $L_{\max}$ spread out over a region of 'volume' $\delta\theta_m$. The value θ_m where $p(x|m, \theta_m)$ attains its maximum can be used as an estimate for θ_m and the 'volume' $\delta\theta_m$ is then interpreted as an uncertainty. Then the integral of $p(x|m, \theta_m)$ can be approximated by the product $L_{\max} \times \delta\theta_m$. Thus, in a very rough and qualitative way the probability for the model given the data is

$$p(m|x) \propto \frac{L_{\max} \times \delta\theta_m}{V_m} \; . \tag{2.203}$$

We can now interpret eq.(2.201) as follows. Our preference for a model will be dictated by how well the model fits the data; this is measured by $[p(x|m, \theta_m)]_{\max} = L_{\max}$. The volume of the region of uncertainty $\delta\theta_m$ also contributes: if more values of the parameters are consistent with the data, then there are more ways the model agrees with the data, and the model is favored. Finally, the larger the volume of possible parameter values V_m the more the model is penalized. Since a larger volume V_m means a more complex model the $1/V_m$ factor penalizes complexity. The preference for simpler models is said to implement Occam's razor. This is a reference to the principle, stated by William of Occam, a 13th century Franciscan monk, that one should not seek a more complicated explanation when a simpler one will do. Such an interpretation is satisfying but ultimately it is quite unnecessary. Occam's principle does not need not be put in by hand: Bayes' theorem takes care of it automatically in eq.(2.201)!

2.11.4 Maximum Likelihood

If one adopts the frequency interpretation of probabilities then most uses of Bayes' theorem are not allowed. The reason is simple: it makes sense to assign a probability distribution $p(x|\theta)$ to the data $x = \{x_i\}$ because the x are random variables but it is absolutely meaningless to talk about probabilities for the parameters θ because they have no frequency distributions, they are not *random* variables, they are merely *unknown*. This means that many problems in science lie beyond the reach of a frequentist probability theory.

To overcome this difficulty a new subject was invented: statistics. Within the Bayesian approach the two subjects, statistics and probability theory, are unified into the single field of inductive inference. In the frequentist approach to statistics in order to infer an unknown quantity θ on the basis of measurements of another quantity, the data x, one postulates the existence of some function

of the data, $\hat{\theta}(x)$, called the 'statistic', that relates the two: the estimate for θ is $\hat{\theta}(x)$. Being afflicted by experimental errors the data x are deemed to be legitimate random variables to which frequentist probability concepts can be applied. The problem is to estimate the unknown θ when what is known is the sampling distribution $p(x|\theta)$. The solution proposed by Fisher was to select as estimator $\hat{\theta}(x)$ that value of θ that maximizes the probability of the observed data x. Since $p(x|\theta)$ is a function of the variable x and θ appears as a fixed parameter, Fisher introduced a function of θ, which he called the likelihood, where the observed data x appear as fixed parameters,

$$L(\theta|x) \stackrel{\text{def}}{=} p(x|\theta) \ . \tag{2.204}$$

Thus, this method of parameter estimation is called the method of 'maximum likelihood'. The likelihood function $L(\theta|x)$ is not a probability, it is not normalized in any way, and it makes no sense to use it compute an average or a variance, but the same intuition that leads one to propose maximization of the likelihood to estimate θ also leads one to use the width of the likelihood function as to estimate an error bar.

The Bayesian approach agrees with the method of maximum likelihood in the special case where of prior is uniform,

$$p(\theta) = \text{const} \Rightarrow p(\theta|x) \propto p(\theta)p(x|\theta) \propto p(x|\theta) \ . \tag{2.205}$$

This explains why the Bayesian discussion of this section has reproduced so many of the standard results of the 'orthodox' theory. But then the Bayesian approach has many other advantages. Unlike the likelihood, the posterior is a true probability distribution that allows estimation not just of θ but of any one of its moments. And, most important, there is no limitation to uniform priors. If there is additional prior information that is relevant to a problem the prior distribution provides a mechanism to take it into account.

Chapter 3

Entropy I: The Evolution of Carnot's Principle

An important problem that occupied the minds of many scientists in the 18th century was either to devise a perpetual motion machine, or to prove its impossibility from the established principles of mechanics. Both attempts failed. Ever since the most rudimentary understanding of the laws of thermodynamics was achieved in the 19th century no competent scientist would waste time considering perpetual motion. The other goal has also proved elusive; there exist no derivations the Second Law from purely mechanical principles. It took a long time, and for many the subject is still controversial, but the reason has gradually become clear: entropy is not a physical quantity, it is a tool for inference, a tool for reasoning in situations of incomplete information. It is quite impossible that such a non-mechanical quantity could emerge from a combination of mechanical notions. If anything it should be the other way around.

Much of the material for this chapter (including the title) is inspired by a beautiful article by E. T. Jaynes [Jaynes 1988]. I have also borrowed from the historical papers [Klein 1970, 1973] and [Uffink 2004].

3.1 Carnot: reversible engines

Sadi Carnot was interested in improving the efficiency of steam engines, that is, of maximizing the amount of useful work that can be extracted from an engine per unit of burnt fuel. His work, published in 1824, was concerned with whether appropriate choices of a working substance other than steam and of the operating temperatures and pressures would improve the efficiency.

Carnot was convinced that perpetual motion was impossible but this was not a fact that he could prove. Indeed, he could not have had a proof: thermodynamics had not been invented yet. His conviction derived instead from the long list of previous attempts (including those by his father Lazare Carnot) that had ended in failure. Carnot's brilliant idea was to proceed anyway and

use what he knew was true but could not prove as the postulate from which he would draw all sorts of other conclusions about engines.[1]

At the time Carnot did his work the nature of heat as a form of energy had not yet been understood. He adopted a model that was fashionable at the time – the caloric model – according to which heat is a substance that could be transferred but neither created nor destroyed. For Carnot an engine used heat to produce work in much the same way that falling water can turn a waterwheel and produce work: the caloric would "fall" from a higher temperature to a lower temperature thereby making the engine turn. What was being transformed into work was not the caloric itself but the energy acquired in the fall.

According to the caloric model the amount of heat extracted from the high temperature source should be the same as the amount of heat discarded into the low temperature sink. Later measurements showed that this was not true, but Carnot was quite lucky. Although the model was seriously wrong, it did have a great virtue: it suggested that the generation of work in a heat engine should include not just the high temperature source from which heat is extracted (the boiler) but also a low temperature sink (the condenser) into which heat is discarded. Later, when heat was interpreted as a form of energy transfer it was understood that for continued operation it was necessary that excess heat be discarded into a low temperature sink so that the engine could complete each cycle by returning to same initial state.

Carnot's caloric-waterwheel model was fortunate in yet another respect – he was not just lucky, he was very lucky – a waterwheel engine can be operated in reverse and used as a pump. This led him to consider a reversible heat engine in which work would be used to draw heat from a cold source and 'pump it up' to deliver heat to the hot reservoir. The analysis of such reversible heat engines led Carnot to the important conclusion

Carnot's Principle: "*No heat engine E can be more efficient than a reversible engine E_R operating between the same temperatures.*"

The proof of Carnot's principle is quite straightforward but because he used the caloric model it was not strictly correct – the necessary revisions were later supplied by Clausius in 1850. As a side remark, it is interesting that Carnot's notebooks, which were made public long after his death by his family in about 1870, indicate that soon after 1824 Carnot came to reject the caloric model and that he achieved the modern understanding of heat as a form of energy transfer. This work – which preceded Joule's experiments by about fifteen years – was not published and therefore had no influence on the history of thermodynamics [Wilson 1981].

The following is Clausius' proof. Figure (3.1a) shows a heat engine E that

[1] In his attempt to understand the undetectability of the ether Einstein faced a similar problem: he knew that it was hopeless to seek an understanding of the constancy of the speed of light on the basis of the primitive physics of the atomic structure of solid rods that was available at the time. Inspired by Carnot he deliberately followed the same strategy – to give up and declare victory – and postulated the constancy of the speed of light as the unproven but known truth which would serve as the foundation from which other conclusions could be derived.

draws heat q_1 from a source at high temperature t_1, delivers heat q_2 to a sink at low temperature t_2, and generates work $w = q_1 - q_2$. Next consider an engine E_S that is more efficient than a reversible one, E_R. In figure (3.1b) we show the super-efficient engine E_S coupled to the reversible E_R. Then for the same heat q_1 drawn from the hot source the super-efficient engine E_S would deliver more work than E_R, $w > w_R$. One could split the work w generated by E_S into two parts w_R and $w - w_R$. The first part w_R could be used to drive E_R in reverse and pump heat q_1 back up to the hot source, which is thus left unchanged. The remaining work $w - w_R$ could then be used for any other purposes. The net result is to extract heat $q_{2R} - q_2 > 0$ from the cold reservoir and convert it to work without any need for fuel. The conclusion is that the existence of a super-efficient heat engine would allow the construction of a perpetual motion engine. Assuming that the latter do not exist implies Carnot's principle: heat engines that are more efficient than reversible ones do not exist.

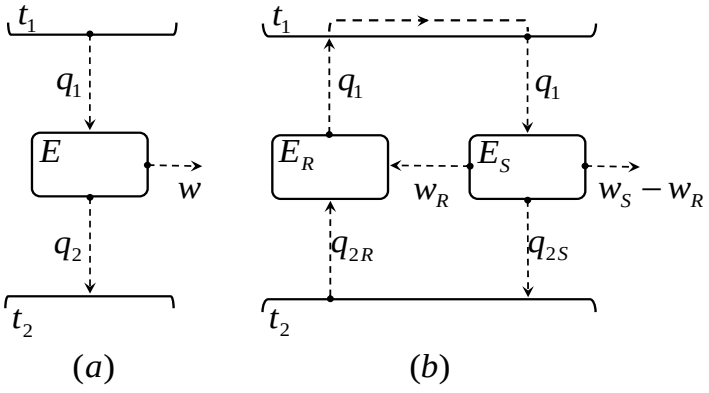


Figure 3.1: (a) An engine E operates between heat reservoirs at temperatures t_1 and t_2. (b) A perpetual motion machine can be built by coupling a super-efficient engine E_S to a reversible engine E_R.

A blank statement of the principle that perpetual motion is not possible is true but it is incomplete. It blurs the important distinction between perpetual motion engines of the *first kind* which operate by violating energy conservation and perpetual motion engines of the *second kind* which do not violate energy conservation. Carnot's conclusion deserves to be singled out as a new principle because it is specific to the second kind of machine.

Other important conclusions obtained by Carnot are that all reversible engines operating between the same temperatures are equally efficient; their efficiency is a function of the temperatures only,

$$e \stackrel{\text{def}}{=} \frac{w}{q_1} = e(t_1, t_2) \ , \tag{3.1}$$

and is therefore independent of any and all other details of how the engine is constructed and operated; that efficiency increases with the temperature difference [see eq.(3.5) below]. Furthermore, the most efficient heat engine cycle, now called the Carnot cycle, is one in which all heat is absorbed at the high t_1 and all heat is discharged at the low t_2. Thus, the Carnot cycle is defined by two isotherms and two adiabats.

The next important step, the determination of the universal function $e(t_1, t_2)$, was accomplished by Kelvin.

3.2 Kelvin: temperature

After Joule's experiments in the 1840's on the conversion of work into heat the caloric model had to be abandoned. Heat was finally recognized as a form of energy and the additional relation $w = q_1 - q_2$ was the ingredient that, in the hands of Kelvin and Clausius, allowed Carnot's principle to be developed into the next stage.

Suppose two reversible engines E_a and E_b are linked in series to form a single more complex reversible engine E_c. E_a operates between temperatures t_1 and t_2, and E_b between t_2 and t_3. E_a draws heat q_1 and discharges q_2, while E_b uses q_2 as input and discharges q_3. The efficiencies of the three engines are

$$e_a = e\,(t_1, t_2) = \frac{w_a}{q_1} \ , \quad e_b = e\,(t_2, t_3) = \frac{w_b}{q_2} \ , \tag{3.2}$$

and

$$e_c = e\,(t_1, t_3) = \frac{w_a + w_b}{q_1} \ . \tag{3.3}$$

They are related by

$$e_c = e_a + \frac{w_b}{q_2}\frac{q_2}{q_1} = e_a + e_b\left(1 - \frac{w_a}{q_1}\right) \ , \tag{3.4}$$

or

$$e_c = e_a + e_b(1 - e_a) \ , \tag{3.5}$$

which is a functional equation for $e = e\,(t_1, t_2)$. Before we proceed to find the solution we note that since $0 \leq e \leq 1$ it follows that $e_c \geq e_a$. Similarly, writing

$$e_c = e_b + e_a(1 - e_b) \ , \tag{3.6}$$

implies $e_c \geq e_b$. Therefore the efficiency $e\,(t_1, t_2)$ can be increased either by increasing the higher temperature or by lowering the lower temperature.

To find the solution of eq.(3.5) change variables to $x = \log(1-e)$,

$$x_c(t_1,t_3) = x_a(t_1,t_2) + x_b(t_2,t_3) \ , \tag{3.7}$$

and then differentiate with respect to t_2 to get

$$\frac{\partial}{\partial t_2} x_a(t_1,t_2) = -\frac{\partial}{\partial t_2} x_b(t_2,t_3) \ . \tag{3.8}$$

The left hand side is independent of t_3 while the second is independent of t_1, therefore $\partial x_a/\partial t_2$ must be some function g of t_2 only,

$$\frac{\partial}{\partial t_2} x_a(t_1,t_2) = g(t_2) \ . \tag{3.9}$$

Integrating gives $x(t_1,t_2) = F(t_1) + G(t_2)$ where the two functions F and G are at this point unknown. The boundary condition $e(t,t) = 0$ or equivalently $x(t,t) = 0$ implies that we deal with merely one unknown function: $G(t) = -F(t)$. Therefore

$$x(t_1,t_2) = F(t_1) - F(t_2) \quad \text{or} \quad e(t_1,t_2) = 1 - \frac{f(t_2)}{f(t_1)} \ , \tag{3.10}$$

where $f = e^{-F}$. Since $e(t_1,t_2)$ increases with t_1 and decreases with t_2 the function $f(t)$ must be monotonically increasing.

Kelvin recognized that there is nothing fundamental about the original temperature scale t. It may depend, for example, on the particular materials employed to construct the thermometer. He realized that the freedom in eq.(3.10) in the choice of the function f corresponds to the freedom of changing temperature scales by using different thermometric materials. The only feature common to all thermometers that claim to rank systems according to their 'degree of hotness' is that they must agree that if A is hotter than B, and B is hotter than C, then A is hotter than C. One can therefore *regraduate* any old inconvenient t scale by a monotonic function to obtain a new scale T chosen for the purely pragmatic reason that it leads to a more elegant formulation of the theory. Inspection of eq.(3.10) immediately suggests the optimal regraduating function choice, which leads to Kelvin's definition of absolute temperature,

$$T = C f(t) \ . \tag{3.11}$$

The scale factor C reflects the still remaining freedom to choose the units. In the absolute scale the efficiency for the ideal reversible heat engine is very simple,

$$e(t_1,t_2) = 1 - \frac{T_2}{T_1} \ . \tag{3.12}$$

In short, what Kelvin proposed was to use an ideal reversible engine as a thermometer with its efficiency playing the role of the thermometric variable.

Carnot's principle that any heat engine E' must be less efficient than the reversible one, $e' \leq e$, is rewritten as

$$e' = \frac{w}{q_1} = 1 - \frac{q_2}{q_1} \leq e = 1 - \frac{T_2}{T_1} \ , \tag{3.13}$$

or,

$$\frac{q_1}{T_1} - \frac{q_2}{T_2} \leq 0 \ . \tag{3.14}$$

It is convenient to redefine heat so that inputs are positive, $Q_1 = q_1$, and outputs are negative, $Q_2 = -q_2$. Then,

$$\frac{Q_1}{T_1} + \frac{Q_2}{T_2} \leq 0 \ , \tag{3.15}$$

where the equality holds when and only when the engine is reversible.

The generalization to an engine or any system that undergoes a cyclic process in which heat is exchanged with more than two reservoirs is straightforward. If heat Q_i is absorbed from the reservoir at temperature T_i we obtain the Kelvin form (1854) of Carnot's principle,

$$\sum_i \frac{Q_i}{T_i} \leq 0 \ . \tag{3.16}$$

which, in the hands of Clausius, led to the next non-trivial step, the introduction of the concept of entropy.

3.3 Clausius: entropy

By about 1850 both Kelvin and Clausius had realized that two laws were necessary as a foundation for thermodynamics. The somewhat awkward expressions for the second law that they had adopted at the time were reminiscent of Carnot's; they stated the impossibility of heat engines whose sole effect would be to transform heat from a single source into work, or of refrigerators that could pump heat from a cold to a hot reservoir without the input of external work. It took Clausius until 1865 – this is some fifteen years later, which indicates that the breakthrough was not at all trivial – before he came up with a new compact statement of the second law that allowed substantial further progress [Cropper 1986].

Clausius rewrote Kelvin's eq.(3.16) for a cycle where the system absorbs infinitesimal (positive or negative) amounts of heat dQ from a continuous sequence of reservoirs,

$$\oint \frac{dQ}{T} \leq 0 \ , \tag{3.17}$$

where T is the temperature of each reservoir. The equality is attained for a reversible process which is achieved when the system is slowly taken through a continuous sequence of equilibrium states and T is also the temperature of the

system as well as that of the reservoirs. The equality implies that the integral from any state A to any other state B is independent of the path taken,

$$\oint \frac{dQ}{T} = 0 \Rightarrow \int\limits_{R_1(A,B)} \frac{dQ}{T} = \int\limits_{R_2(A,B)} \frac{dQ}{T} \ , \tag{3.18}$$

where $R_1(A,B)$ and $R_2(A,B)$ denote any two reversible paths linking the same initial state A and final state B. Clausius saw that eq.(3.18) implies the existence of a function of the thermodynamic state. This function, which he called entropy, is defined up to an additive constant by

$$S_B = S_A + \int\limits_{R(A,B)} \frac{dQ}{T} \ . \tag{3.19}$$

This first notion of entropy we will call the Clausius entropy or the thermodynamic entropy. Note that it is of rather limited applicability as it is defined only for states of thermal equilibrium.

Eq.(3.19) seems like a mere reformulation of eqs.(3.16) and (3.17) but it represents a major advance because it allowed thermodynamics to reach beyond the study of cyclic processes. Consider a possibly irreversible process in which a system is taken from an initial state A to a final state B, and suppose the system is returned to the initial state along some other reversible path. Then, the more general eq.(3.17) gives

$$\int\limits_{A,\text{irrev}}^{B} \frac{dQ}{T} + \int\limits_{R(B,A)} \frac{dQ}{T} \leq 0 \ . \tag{3.20}$$

From eq.(3.19) the second integral is $S_A - S_B$. In the first integral $-dQ$ is the amount is the amount of heat absorbed by the reservoirs at temperature T and therefore it represents minus the change in the entropy of the reservoirs which in this case represent the rest of the universe,

$$(S^{\text{res}}_A - S^{\text{res}}_B) + (S_A - S_B) \leq 0 \quad \text{or} \quad S^{\text{res}}_B + S_B \geq S^{\text{res}}_A + S_A \ . \tag{3.21}$$

Thus the second law can be stated in terms of the total entropy $S^{\text{total}} = S^{\text{res}} + S$ as

$$S^{\text{total}}_{\text{final}} \geq S^{\text{total}}_{\text{initial}} \ , \tag{3.22}$$

and Clausius could then summarize the laws of thermodynamics as "*The energy of the universe is constant. The entropy of the universe tends to a maximum.*" All restrictions to cyclic processes have disappeared.

Clausius was also responsible for initiating another independent line of research in this subject. His paper "On the kind of motion we call heat" (1857) was the first (failed!) attempt to deduce the second law from purely mechanical principles applied to molecules. His results referred to averages taken over all molecules, for example the kinetic energy per molecule, and involved theorems in mechanics such as the virial theorem. For him the increase of entropy was meant to be an absolute law and not just a matter of overwhelming probability.

3.4 Maxwell: probability

We owe to Maxwell the introduction of probabilistic notions into fundamental physics (1860). Before him probabilities had been used by Laplace and by Gauss as a tool in the analysis of experimental data. Maxwell realized the practical impossibility of keeping track of the exact motion of all the molecules in a gas and pursued a less detailed description in terms of the distribution of velocities. (Perhaps he was inspired by his earlier study of the rings of Saturn which required reasoning about particles undergoing very complex trajectories.)

Maxwell interpreted his distribution function as the number of molecules with velocities in a certain range, and also as the probability $P(\vec{v})d^3v$ that a molecule has a velocity $\vec{v}$ in a certain range d^3v. It would take a long time to achieve a clearer understanding of the meaning of the term 'probability'. In any case, Maxwell concluded that "velocities are distributed among the particles according to the same law as the errors are distributed in the theory of the 'method of least squares'," and on the basis of this distribution he obtained a number of significant results on the transport properties of gases.

Over the years he proposed several derivations of his velocity distribution function. The earlier one (1860) is very elegant. It involves two assumptions: the first is a symmetry requirement, the distribution should only depend on the actual magnitude $|\vec{v}| = v$ of the velocity and not on its direction,

$$P(\vec{v})d^3v = f(v)d^3v = f\left(\sqrt{v_x^2 + v_y^2 + v_z^2}\right) d^3v \ . \tag{3.23}$$

The second assumption is that velocities along orthogonal directions should be independent

$$f(v)d^3v = p(v_x)p(v_y)p(v_z)d^3v \ . \tag{3.24}$$

Therefore

$$f\left(\sqrt{v_x^2 + v_y^2 + v_z^2}\right) = p(v_x)p(v_y)p(v_z) \ . \tag{3.25}$$

Setting $v_y = v_z = 0$ we get

$$f\left(v_x\right) = p(v_x)p(0)p(0) \ , \tag{3.26}$$

so that we obtain a functional equation for p,

$$p\left(\sqrt{v_x^2 + v_y^2 + v_z^2}\right) p(0)p(0) = p(v_x)p(v_y)p(v_z) \ , \tag{3.27}$$

or

$$\log\left[\frac{p\left(\sqrt{v_x^2 + v_y^2 + v_z^2}\right)}{p(0)}\right] = \log\left[\frac{p(v_x)}{p(0)}\right] + \log\left[\frac{p(v_y)}{p(0)}\right] + \log\left[\frac{p(v_z)}{p(0)}\right] \ , \tag{3.28}$$

or, introducing the function $G = \log[p/p(0)]$,

$$G\left(\sqrt{v_x^2 + v_y^2 + v_z^2}\right) = G(v_x) + G(v_y) + G(v_z). \tag{3.29}$$

The solution is straightforward. Differentiate with respect to v_x and to v_y to get

$$\frac{G'\left(\sqrt{v_x^2+v_y^2+v_z^2}\right)}{\sqrt{v_x^2+v_y^2+v_z^2}}v_x = G'(v_x) \quad \text{and} \quad \frac{G'\left(\sqrt{v_x^2+v_y^2+v_z^2}\right)}{\sqrt{v_x^2+v_y^2+v_z^2}}v_y = G'(v_y) \ . \tag{3.30}$$

Therefore

$$\frac{G'(v_x)}{v_x} = \frac{G'(v_y)}{v_y} = -2\alpha \ , \tag{3.31}$$

where -2α is a constant. Integrating gives

$$\log\left[\frac{p(v_x)}{p(0)}\right] = G(v_x) = -\alpha v_x^2 + \text{const} \ , \tag{3.32}$$

so that

$$P(\vec{v}) = f(v) = \left(\frac{\alpha}{\pi}\right)^{3/2} \exp\left[-\alpha\left(v_x^2+v_y^2+v_z^2\right)\right] \ , \tag{3.33}$$

the same distribution as "errors in the method of least squares".

Maxwell's distribution applies whether the molecule is part of a gas, a liquid, or a solid and, with the benefit of hindsight, the reason is quite easy to see. The probability that a molecule have velocity $\vec{v}$ and position $\vec{x}$ is given by the Boltzmann distribution $\propto \exp -H/kT$. For a large variety of situations the Hamiltonian for one molecule is of the form $H = mv^2/2 + V(\vec{x})$ where the potential $V(\vec{x})$ includes the interactions, whether they be weak or strong, with all the other molecules. If the potential $V(\vec{x})$ is independent of $\vec{v}$, then the distribution for $\vec{v}$ and $\vec{x}$ factorizes. Velocity and position are statistically independent, and the velocity distribution is Maxwell's.

Maxwell was the first to realize that the second law is not an absolute law (this was expressed in his popular textbook "Theory of Heat" in 1871), that it "has only statistical certainty" and indeed, that in fluctuation phenomena "the second law is continually being violated". Such phenomena are not rare: just look out the window and you can see that the sky is blue – a consequence of the scattering of light by density fluctuations in the atmosphere.

Maxwell introduced the notion of probability, but what did he actually mean by the word 'probability'? He used his distribution function as a velocity distribution, the number of molecules with velocities in a certain range, which betrays a frequentist interpretation. These probabilities are ultimately mechanical properties of the gas. But he also used his distribution to represent the lack of information we have about the precise microstate of the gas. This latter interpretation is particularly evident in a letter he wrote in 1867 where he argues that the second law could be violated by "a finite being who knows the paths and velocities of all molecules by simple inspection but can do no work except open or close a hole." Such a "demon" could allow fast molecules to pass through a hole from a vessel containing hot gas into a vessel containing cold gas, and could allow slow molecules pass in the opposite direction. The net

effect being the transfer of heat from a low to a high temperature, a violation of the second law. All that was required was that the demon "know" the right information. [Klein 1970]

3.5 Gibbs: beyond heat

Gibbs generalized the second law in two directions: to open systems and to inhomogeneous systems. With the introduction of the concept of the chemical potential, a quantity that regulates the transfer of particles in much the same way that temperature regulates the transfer of heat, he could apply the methods of thermodynamics to phase transitions, mixtures and solutions, chemical reactions, and much else. His paper "On the Equilibrium of Heterogeneous Systems" [Gibbs 1875-78] is formulated as the purest form of thermodynamics – a phenomenological theory of extremely wide applicability because its foundations do not rest on particular models about the structure and dynamics of the microscopic constituents.

And yet, Gibbs was keenly aware of the significance of the underlying molecular constitution – he was familiar with Maxwell's writings and in particular with his "Theory of Heat". His discussion of the process of mixing gases led him to analyze the paradox that bears his name. The entropy of two different gases increases when the gases are mixed; but does the entropy also increase when two gases of the same molecular species are mixed? Is this an irreversible process?

For Gibbs there was no paradox, much less one that would require some esoteric new (quantum) physics for its resolution. For him it was quite clear that thermodynamics was not concerned with microscopic details but rather with the changes from one macrostate to another. He explained that the mixing of two gases of the same molecular species does not lead to a different macrostate. Indeed: by "thermodynamic" state

> "...we do not mean a state in which each particle shall occupy more or less exactly the same position as at some previous epoch, but only a state which shall be indistinguishable from the previous one in its sensible properties. It is to states of systems thus incompletely defined that the problems of thermodynamics relate." [Gibbs 1875-78]

Thus, there is no entropy increase because there is no change of thermodynamic state. Gibbs' resolution of the non-paradox hinges on distinguishing two kinds of reversibility. One is the microscopic or mechanical reversibility in which the velocities of each individual particle is reversed and the system retraces the sequence of microstates. The other is macroscopic or Carnot reversibility in which the system retraces the sequence of macrostates.

Gibbs understood, as had Maxwell before him, that the explanation of the second law cannot rest on purely mechanical arguments. Since the second law applies to "incompletely defined" descriptions any explanation must also involve probabilistic concepts that are foreign to mechanics. This led him to conclude: "In other words, the impossibility of an uncompensated decrease of entropy

seems to be reduced to improbability," a sentence that Boltzmann adopted as the motto for the second volume of his "Lectures on the Theory of Gases." (For a modern discussion of the Gibbs' paradox see section 5.10.)

Remarkably neither Maxwell nor Gibbs established a connection between probability and entropy. Gibbs was very successful at showing what one can accomplish by maximizing entropy but he did not address the issue of what entropy is or what it means. The crucial steps in this direction were taken by Boltzmann.

But Gibbs' contributions did not end here. The ensemble theory introduced in his "Principles of Statistical Mechanics" in 1902 (it was Gibbs who coined the term 'statistical mechanics') represent a practical and conceptual step beyond Boltzmann's understanding of entropy.

3.6 Boltzmann: entropy and probability

It was Boltzmann who found the connection between entropy and probability, but his path was long and tortuous [Klein 1973, Uffink 2004]. Over the years he adopted several different interpretations of probability and, to add to the confusion, he was not always explicit about which one he was using, sometimes mixing them within the same paper, and even within the same equation. At first, he defined the probability of a molecule having a velocity $\vec{v}$ within a small cell d^3v as being proportional to the amount of time that the particle spent within that particular cell, but he also defined that same probability as the fraction of particles within the cell.

By 1868 he had managed to generalize the Maxwell distribution for point particles and the theorem of equipartition of energy to complex molecules in the presence of an external field. The basic argument, which led him to the Boltzmann distribution, was that in equilibrium the distribution should be stationary, that it should not change as a result of collisions among particles.

The collision argument gave the distribution for individual molecules; it did not keep track of information about correlations among molecules. It was also in 1868 that Boltzmann first applied probability to the microstate of the system as a whole rather than to the individual molecules. This led him to the microcanonical distribution in which microstates are uniformly distributed over the hypersurface of constant energy. Boltzmann identified the probability of the system being in some region of the N-particle phase space (rather than the one-particle space of molecular velocities) with the relative time the system would spend in that region – the so-called "time" ensemble. Alternatively, probability was also defined at a given instant in time as being proportional to the volume of the region. At first Boltzmann did not think it was necessary to comment on whether the two definitions are equivalent or not, but eventually he realized that their assumed equivalence should be explicitly stated. Later this came to be known as the 'ergodic' hypothesis, namely, that over a long time the trajectory of the system would cover the whole region of phase space consistent with the given value of the energy. Throughout this period Boltzmann's various notions

of probability were all still conceived as mechanical properties of the gas.

In 1871 Boltzmann achieved a significant success in establishing a connection between thermodynamic entropy and microscopic concepts such as the probability distribution in the N-particle phase space. In modern notation his argument was as follows. The energy of N interacting particles is given by

$$H = \sum_i^N \frac{p_i^2}{2m} + U\left(x_1, \ldots, x_N; V\right) \ , \tag{3.34}$$

where V stands for additional parameters that can be externally controlled such as, for example, the volume of the gas. The first non-trivial decision was to propose a quantity defined in purely microscopic terms that would correspond to the macroscopic internal energy. He opted for the "expectation"

$$E = \langle H \rangle = \int dz_N \, P_N \, H \ , \tag{3.35}$$

where $dz_N = d^{3N}x d^{3N}p$ is the volume element in the N-particle phase space, and P_N is the N-particle distribution function,

$$P_N = \frac{\exp\left(-\beta H\right)}{Z} \quad \text{where} \quad Z = \int dz_N \ e^{-\beta H} \ , \tag{3.36}$$

and $\beta = 1/kT$, so that,

$$E = \frac{3}{2} NkT + \langle U \rangle \ . \tag{3.37}$$

The connection to the thermodynamic entropy requires a clear idea of the nature of heat and how it differs from work. One needs to express heat in purely microscopic terms, and this is quite subtle because at the molecular level there is no distinction between a motion that is supposedly of a "thermal" type as opposed to other types of motion such as plain displacements or rotations. The distribution function is the crucial ingredient. In any infinitesimal transformation the change in the internal energy separates into two contributions,

$$\delta E = \int dz_N \ H \delta P_N \ + \int dz_N \, P_N \delta H \ . \tag{3.38}$$

The second integral, which can be written as $\langle \delta H \rangle = \langle \delta U \rangle$, arises purely from changes in the potential function U that are induced by manipulating parameters such as volume. Such a change in the potential is precisely what one means by mechanical work δW, therefore, since $\delta E = \delta Q + \delta W$, the first integral must represent the transferred heat δQ,

$$\delta Q = \delta E - \langle \delta U \rangle \ . \tag{3.39}$$

On the other hand, substituting δE from eq.(3.37), one gets

$$\delta Q = \frac{3}{2} Nk\delta T + \delta \left\langle U \right\rangle - \langle \delta U \rangle \ . \tag{3.40}$$

This is not a complete differential, but dividing by the temperature yields (after some algebra)

$$\frac{\delta Q}{T} = \delta \left[\frac{3}{2} N k \log T + \frac{\langle U \rangle}{T} + k \log \left(\int d^{3N} x \, e^{-\beta U} \right) + \text{const} \right] . \tag{3.41}$$

If the identification of δQ with heat is correct then this strongly suggests that the expression in brackets should be identified with the Clausius entropy S. Further rewriting leads to

$$S = \frac{E}{T} + k \log Z + \text{const} \, , \tag{3.42}$$

which is recognized as the correct modern expression.

Boltzmann's path towards understanding the second law was guided by one notion from which he never wavered: matter is an aggregate of molecules. Apart from this the story of his progress is the story of the increasingly more important role played by probabilistic notions, and ultimately, it is the story of the evolution of his understanding of the notion of probability itself. By 1877 Boltzmann achieves his final goal and explains entropy purely in terms of probability – mechanical notions were by now reduced to the bare minimum consistent with the subject matter: we are, after all, talking about collections of molecules with positions and momenta and their total energy is conserved. His final achievement hinges on the introduction of yet another way of thinking about probabilities involving the notion of the multiplicity of the macrostate.

He considered an idealized system consisting of N particles whose single-particle phase space is divided into m cells each with energy ε_n, $n = 1, ..., m$. The number of particles in the nth cell is denoted w_n, and the distribution function is given by the set of numbers $w_1, \ldots, w_m$. In Boltzmann's previous work the determination of the distribution function had been based on figuring out its time evolution from the mechanics of collisions. Here he used a purely combinatorial argument. A completely specified state, what he called a complexion and we call a microstate, is defined by specifying the cell of each individual molecule. A macrostate is less completely specified by the distribution function, $w_1, \ldots, w_m$. The number of microstates compatible with a given macrostate, which Boltzmann called the 'permutability', and we call the 'multiplicity' is

$$W = \frac{N!}{w_1! \ldots w_m!} \, . \tag{3.43}$$

Boltzmann proposed that the probability of the macrostate was proportional to its multiplicity, to the number of ways in which it could be achieved, which assumes each microstate is as likely as any other – the 'equal a priori probability postulate'.

The most probable macrostate is that which maximizes W subject to the constraints of a fixed total number of particles N and a fixed total energy E,

$$\sum_{n=1}^{m} w_n = N \quad \text{and} \quad \sum_{n=1}^{m} w_n \varepsilon_n = E. \tag{3.44}$$

When the numbers w_n are large enough that one can use Stirling's approximation for the factorials, we have

$$\log W = N \log N - N - \sum_{n=1}^{m} (w_n \log w_n - w_n) \tag{3.45}$$

$$= -\sum_{n=1}^{m} w_n \log w_n + \text{const} \ , \tag{3.46}$$

or perhaps better

$$\log W = -N \sum_{n=1}^{m} \frac{w_n}{N} \log \frac{w_n}{N} \tag{3.47}$$

so that

$$\log W = -N \sum_{n=1}^{m} f_n \log f_n \tag{3.48}$$

where $f_n = w_n/N$ is the fraction of molecules in the nth cell with energy ε_n, or, alternatively the probability that a molecule is in its nth state. As we shall later derive in detail, the distribution that maximizes $\log W$ subject to the constraints (3.44) is such that

$$f_n = \frac{w_n}{N} \propto e^{-\beta \varepsilon_n} \ , \tag{3.49}$$

where β is a Lagrange multiplier determined by the total energy. When applied to a gas, the possible states of a molecule are cells in phase space. Therefore

$$\log W = -N \int dz_1 \, f(x,p) \log f(x,p) \ , \tag{3.50}$$

where $dz_1 = d^3x d^3p$ and the most probable distribution is the equilibrium distribution found earlier by Maxwell and generalized by Boltzmann.

In this approach probabilities are central. The role of dynamics is minimized but it is not eliminated. The Hamiltonian enters the discussion in two places. One is quite explicit: there is a conserved energy the value of which is imposed as a constraint. The second is much more subtle; we saw above that the probability of a macrostate could be taken to be proportional to the multiplicity W provided microstates are assigned equal probabilities, or equivalently, equal volumes in phase space are assigned equal a priori weights. As always equal probabilities must ultimately be justified in terms of some form of underlying symmetry. In this case, the symmetry follows from Liouville's theorem – under a Hamiltonian time evolution a region in phase space will move around and its shape will be distorted but its volume will be conserved: Hamiltonian time evolution preserves volumes in phase space. The nearly universal applicability of the 'equal a priori postulate' can be traced to the fact that the only thing that is needed is a Hamiltonian; any Hamiltonian would do.

It is very remarkable that although Boltzmann calculated the maximized value $\log W$ for an ideal gas and knew that it agreed with the thermodynamical entropy except for a scale factor, he never wrote the famous equation that bears his name

$$S = k \log W \ . \tag{3.51}$$

This equation, as well as Boltzmann's constant k, were both first introduced by Planck.

There is, however, a problem with eq.(3.50): it involves the distribution function $f(x, p)$ in the one-particle phase space and therefore it cannot take correlations into account. Indeed, eq.(3.50) gives the correct form of the entropy only for ideal gases of non-interacting particles. The expression that applies to systems of interacting particles is[2]

$$\log W = -\int dz_N \; f_N \log f_N \; , \tag{3.52}$$

where $f_N = f_N(x_1, p_1, \ldots, x_N, p_N)$ is the probability distribution in the N-particle phase space. This equation is usually associated with the name of Gibbs who, in his "Principles of Statistical Mechanics" (1902), developed Boltzmann's combinatorial arguments into a very powerful theory of ensembles. The conceptual gap between eq.(3.50) and (3.52) is enormous; it goes well beyond the issue of intermolecular interactions. The probability in Eq.(3.50) is the single-particle distribution, it can be interpreted as a "mechanical" property, namely, the relative number of molecules in each cell. The entropy Eq.(3.50) can be interpreted as a mechanical property of the individual system. In contrast, eq.(3.52) involves the N-particle distribution which is not a property of any single individual system but a property of an ensemble of replicas of the system. Gibbs was not very explicit about his interpretation of probability. He wrote

> "The states of the bodies which we handle are certainly not *known* to us exactly. What we *know* about a body can generally be described most accurately and most simply by saying that it is one taken at random from a great number (ensemble) of bodies which are completely described." [my italics, Gibbs 1902, p.163]

It is clear that for Gibbs probabilities represent a state of knowledge, that the ensemble is a purely imaginary construction, just a tool for handling incomplete information. On the other hand, it is also clear that Gibbs still struggles thinking of probabilities in terms of frequencies. If the only reliable notion of probability that is available requires an ensemble and no such thing is anywhere to be found then either one adopts an imaginary ensemble or one altogether gives up on probability, an option too extreme to contemplate.

This brings our story of entropy up to about 1900. In the next chapter we start a more deliberate and systematic study of the connection between entropy and information.

3.7 Some remarks

I end with a disclaimer: this chapter has historical overtones but it is not history. Lines of research such as the Boltzmann equation and the ergodic hypothesis

[2]For the moment we disregard the question of the distinguishability of the molecules. The so-called Gibbs paradox and the extra factor of $1/N!$ will be discussed in detail in chapter 4.

that were historically very important have been omitted because they represent paths that diverge from the central theme of this book, namely that the laws of physics can be understood as rules for handling information and uncertainty. Our goal is to discuss thermodynamics and statistical mechanics as the first historical example of such an *information* physics. At first I tried to write a 'history as it should have happened'. I wanted to trace the development of the concept of entropy from its origins with Carnot in a manner that reflects the logical rather than the actual evolution. But I found that this approach would not do; it trivializes the enormous achievements of the 19th century thinkers and it misrepresents the actual nature of research. Scientific research is not a tidy business.

I mentioned that this chapter was inspired by a beautiful article by E. T. Jaynes with the same title [Jaynes 1988]. I think Jaynes' article has great pedagogical value but I disagree with him on how well Gibbs understood the logical status of thermodynamics and statistical mechanics as examples of inferential and probabilistic thinking. My own assessment runs in quite the opposite direction: the reason why the conceptual foundations of thermodynamics and statistical mechanics have been so controversial throughout the 20th century is precisely because neither Gibbs nor Boltzmann, nor anyone else at the time, were particularly clear on the interpretation of probability. I think that we could hardly expect them to have done much better; they did not benefit from the writings of Keynes (1921), Ramsey (1931), de Finetti (1937), Jeffreys (1939), Cox (1946), Shannon (1948), Polya (1954) and, of course, Jaynes himself (1957). Indeed, whatever clarity Jaynes attributes to Gibbs, is not Gibbs'; it is the hard-won clarity that Jaynes attained through his own efforts and after absorbing much of the best the 20th century had to offer.

Chapter 4

Entropy II: Measuring Information

What is information? Our central goal is to gain insight into the nature of information, how one manipulates it, and the implications such insights have for physics. In chapter 2 we provided a first partial answer. We might not yet know precisely what information is but sometimes we can recognize it. For example, it is clear that experimental data contains information, that the correct way to process it involves Bayes's rule, and that this is very relevant to the empirical aspect of all science, namely, to data analysis. Bayes' rule is the machinery that processes the information contained in data to update from a prior to a posterior probability distribution. This suggests a possible generalization: "information" is whatever induces one to update from one state of belief to another. This is a notion that will be explored in detail later.

In this chapter we pursue a different point of view that has turned out to be extremely fruitful. We saw that the natural way to deal with uncertainty, that is, with lack of information, is to introduce the notion of degrees of belief, and that these measures of plausibility should be manipulated and calculated using the ordinary rules of the calculus of probabilities. This achievement is a considerable step forward but it is not sufficient. The problem is that what the rules of probability theory allow us to do is to assign probabilities to some "complex" propositions on the basis of the probabilities that have been previously assigned to other, perhaps more "elementary" propositions. The solution is to introduce a new inference tool designed specifically for assigning those elementary probabilities. The new tool is Shannon's measure of an "amount of information" and the associated method of reasoning is Jaynes' Method of Maximum Entropy, or MaxEnt. [Shannon 1948, Jaynes 1957b, 1983, 2003]

4.1 Shannon's information measure

Consider a set of mutually exclusive and exhaustive alternatives i, for example, the possible values of a variable, or the possible states of a system. The state of the system is unknown. Suppose that on the basis of the incomplete information I we have somehow assigned probabilities $p(i|I) = p_i$. In order to figure out which state within the set $\{i\}$ is the correct one we need more information. The question we address here is how much more. Note that we are not asking the more difficult question of which particular piece of information is missing, but merely the quantity of information that is missing. It seems reasonable that the amount of information that is missing in a sharply peaked distribution is smaller than the amount missing in a broad distribution, but how much smaller? Is it possible to quantify the notion of amount of information? Can one find a unique quantity S that is a function of the p_i's, that tends to be large for broad distributions and small for narrow ones?

Consider a discrete set of n mutually exclusive and exhaustive discrete states i, each with probability p_i. According to Shannon, the measure S of the amount of information that is missing when all we know is the distribution p_i must satisfy three axioms. It is quite remarkable that these three conditions are sufficiently constraining to determine the quantity S uniquely. The first two axioms are deceptively simple.

Axiom 1. S is a real continuous function of the probabilities p_i, $S[p] = S(p_1, \ldots p_n)$.

Remark: It is explicitly assumed that $S[p]$ depends only on the p_i and on nothing else. What we seek here is an *absolute* measure of the amount of missing information in p. If the objective were to update from a prior q to a posterior distribution p – a problem that will be later tackled in chapter 6 – then we would require a functional $S[p, q]$ depending on both q and p. Such $S[p, q]$ would at best be a *relative* measure: the information in p relative to the reference distribution q.

Axiom 2. If all the p_i's are equal, $p_i = 1/n$. Then $S = S(1/n, \ldots, 1/n) = F(n)$, where $F(n)$ is an increasing function of n.

Remark: This means that it takes less information to pinpoint one alternative among a few than one alternative among many and also that knowing the number n of available states is already a valuable piece of information. Notice that the uniform distribution $p_i = 1/n$ is singled out to play a very special role. Indeed, although no reference distribution has been explicitly mentioned, the uniform distribution will, in effect, provide the standard of complete ignorance.

The third axiom is a consistency requirement and is somewhat less intuitive. The entropy $S[p]$ measures the amount of additional information beyond the incomplete information I already codified in the p_i that will be needed to pinpoint the actual state of the system. Imagine that this missing information were to be obtained not all at once but in installments. The consistency requirement is that the particular manner in which we obtain this information should not matter. This idea can be expressed as follows.

Imagine the n states are divided into N groups labeled by $g = 1, \ldots, N$ as

shown in Fig 4.1.

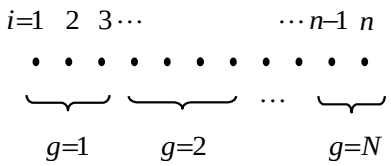


Figure 4.1: The n states are divided into N groups to formulate the grouping axiom.

The probability that the system is found in group g is

$$P_g = \sum_{i \in g} p_i \,. \tag{4.1}$$

Let $p_{i|g}$ denote the conditional probability that the system is in the state $i \in g$ given it is in group g,

$$p_{i|g} = \frac{p_i}{P_g} \quad \text{for} \quad i \in g. \tag{4.2}$$

Suppose we were to obtain the desired information in two steps, the first of which would allow us to single out one of the groups g while the second would allow us to decide on the actual i within the selected group g. The amount of information required in the first step is $S_G = S[P]$ where $P = \{P_g\}$ with $g = 1 \ldots N$. Now suppose we did get this information, and as a result we found, for example, that the system was in group g'. Then for the second step, to single out the state i within the group g', the amount of additional information needed would be $S_{g'} = S[p_{\cdot|g'}]$. But at the beginning of this process we do not yet know which of the gs is the correct one. The *expected amount of missing information* to take us from the gs to the actual is is $\sum_g P_g S_g$. The consistency requirement is that it should not matter whether we get the total missing information in one step, which completely determines i, or in two steps, the first of which has low resolution and only determines one of the groups, say g', while the second step

provides the fine tuning that determines i within g'. This gives us our third axiom:

Axiom 3. For all possible groupings $g = 1 \ldots N$ of the states $i = 1 \ldots n$ we must have

$$S[p] = S_G[P] + \sum_g P_g S_g[p_{\cdot|g}] \,. \tag{4.3}$$

This is called the "grouping" property.

Remark: Given axiom 3 it might seem more appropriate to interpret S as a measure of the *expected* rather than the *actual* amount of missing information, but if S is the expected value of something, it is not clear, at this point, what that something would be. We will return to this below.

The solution to Shannon's constraints is obtained in two steps. First assume that all states i are equally likely, $p_i = 1/n$. Also assume that the N groups g all have the same number of states, $m = n/N$, so that $P_g = 1/N$ and $p_{i|g} = p_i/P_g = 1/m$. Then by axiom 2,

$$S[p_i] = S\,(1/n, \ldots, 1/n) = F\,(n)\,, \tag{4.4}$$

$$S_G[P_g] = S\,(1/N, \ldots, 1/N) = F\,(N)\,, \tag{4.5}$$

and

$$S_g[p_{i|g}] = S(1/m, \ldots, 1/m) = F(m). \tag{4.6}$$

Then, axiom 3 gives

$$F\,(mN) = F\,(N) + F\,(m) \;\; . \tag{4.7}$$

This should be true for all integers N and m. It is easy to see that one solution of this equation is

$$F\,(m) = k\,\log\,m\;, \tag{4.8}$$

where k is any positive constant (just substitute), but it is also easy to see that eq.(4.7) has infinitely many other solutions. Indeed, since any integer m can be uniquely decomposed as a product of prime numbers, $m = \prod_r q_r^{\alpha_r}$, where α_i are integers and q_r are prime numbers, using eq.(4.7) we have

$$F\,(m) = \textstyle\sum_r \alpha_r F(q_r) \tag{4.9}$$

which means that eq.(4.7) can be satisfied by arbitrarily specifying $F(q_r)$ on the primes and then defining $F(m)$ for any other integer through eq.(4.9). A unique solution is obtained when we impose the additional requirement that $F(m)$ be monotonic increasing in m (axiom 2). The following argument is found in [Shannon Weaver 1949]; see also [Jaynes 2003]. Consider any two integers s and t both larger than 1. The ratio of their logarithms can be approximated arbitrarily closely by a rational number, i.e., we can find integers α and β (with β arbitrarily large) such that

$$\frac{\alpha}{\beta} \leq \frac{\log s}{\log t} < \frac{\alpha+1}{\beta} \quad \text{or} \quad t^\alpha \leq s^\beta < t^{\alpha+1}\;. \tag{4.10}$$

But F is monotonic increasing, therefore

$$F(t^\alpha) \leq F(s^\beta) < F(t^{\alpha+1}) \ , \tag{4.11}$$

and using eq.(4.7),

$$\alpha F(t) \leq \beta F(s) < (\alpha+1)F(t) \quad \text{or} \quad \frac{\alpha}{\beta} \leq \frac{F(s)}{F(t)} < \frac{\alpha+1}{\beta} \ . \tag{4.12}$$

Which means that the ratio $F(s)/F(t)$ can be approximated by the same rational number α/β. Indeed, comparing eqs.(4.10) and (4.12) we get

$$\left| \frac{F(s)}{F(t)} - \frac{\log s}{\log t} \right| \leq \frac{1}{\beta} \tag{4.13}$$

or,

$$\left| \frac{F(s)}{\log s} - \frac{F(t)}{\log t} \right| \leq \frac{F(t)}{\beta \log s} \tag{4.14}$$

We can make the right hand side arbitrarily small by choosing β sufficiently large, therefore $F(s)/\log s$ must be a constant, which proves (4.8) is the unique solution.

In the second step of our derivation we will still assume that all is are equally likely, so that $p_i = 1/n$ and $S[p] = F(n)$. But now we assume the groups g have different sizes, m_g, with $P_g = m_g/n$ and $p_{i|g} = 1/m_g$. Then axiom 3 becomes

$$F(n) = S_G[P] + \sum_g P_g \, F(m_g),$$

Therefore,

$$S_G[P] = F(n) - \sum_g P_g F(m_g) = \sum_g P_g \, [F(n) - F(m_g)] \ .$$

Substituting our previous expression for F we get

$$S_G[P] = \sum_g P_g \, k \, \log \frac{n}{m_g} = -k \sum_{i=1}^{N} P_g \, \log P_g \, .$$

Therefore Shannon's quantitative measure of the amount of missing information, the entropy of the probability distribution $p_1, \ldots, p_n$ is

$$S[p] = -k \sum_{i=1}^{n} p_i \, \log \, p_i \, . \tag{4.15}$$

Comments

Notice that for discrete probability distributions we have $p_i \leq 1$ and $\log p_i \leq 0$. Therefore $S \geq 0$ for $k > 0$. As long as we interpret S as the amount of uncertainty or of missing information it cannot be negative. We can also check that in cases where there is no uncertainty we get $S = 0$: if any state has probability one, all the other states have probability zero and every term in S vanishes.

The fact that entropy depends on the available information implies that there is no such thing as *the* entropy of a system. The same system may have many different entropies. Indeed, two different agents may reasonably assign different probability distributions p and p' so that $S[p] \neq S[p']$. But the non-uniqueness of entropy goes even further: the same agent may legitimately assign two entropies to the same system. This possibility is already shown in axiom 3 which makes explicit reference to two entropies $S[p]$ and $S_G[P]$ referring to two different descriptions of the same system. Colloquially, however, one does refer to *the* entropy of a system; in such cases the relevant information available about the system should be obvious from the context. For example, in thermodynamics what one means by *the* entropy is the particular entropy that one obtains when the only information available is specified by the known values of those few variables that specify the thermodynamic macrostate.

The choice of the constant k is purely a matter of convention. In thermodynamics the choice is Boltzmann's constant $k_B = 1.38 \times 10^{-16}$erg/K which reflects the historical choice of the Kelvin as the unit of temperature. A more convenient choice is $k = 1$ which makes temperature have energy units and entropy dimensionless. In communication theory and computer science, the conventional choice is $k = 1/\log_e 2 \approx 1.4427$, so that

$$S[p] = -\sum_{i=1}^{n} p_i \log_2 p_i \ . \tag{4.16}$$

The base of the logarithm is 2, and the entropy is said to measure information in units called 'bits'.

Now we turn to the question of interpretation. Earlier we mentioned that from axiom 3 it seems more appropriate to interpret S as a measure of the *expected* rather than the *actual* amount of missing information. If one adopts this interpretation, the actual amount of information that we gain when we find that i is the true alternative could be $\log 1/p_i$. But this is not quite satisfactory. Consider a variable that takes just two values, 0 and 1, with probabilities p and $1-p$ respectively. For very small p, $\log 1/p$ would be very large, while the information that communicates the true alternative is conveyed by a very short one bit message, namely "0". This shows that what $\log 1/p$ measures is not the *actual amount* of information but rather how unexpected or how surprising that piece of information might be. Accordingly, $\log 1/p_i$ is sometimes called the "surprise" of i.

Perhaps one could interpret $S[p]$ as the uncertainty implicit in p — we use the word 'uncertainty' as roughly synonymous to 'lack of information' so that

more information implies less uncertainty. But, as the following example shows, this does not always work either. I normally keep my keys in my pocket. My state of knowledge about the location of my keys is represented by a probability distribution that is sharply peaked at my pocket and reflects a small uncertainty. But suppose I check and I find that my pocket is empty. Then my keys could be virtually anywhere. My new state of knowledge is represented by a very broad distribution that reflects a high uncertainty. We have here a situation where the acquisition of more information has increased the uncertainty rather than decreased it. (This question is further discussed in section 4.6.)

The point of these remarks is not to suggest that there is something wrong with the mathematical derivation — there is not, eq.(4.15) does follow from the axioms — but to suggest caution when interpreting S. The notion of information is at this point still vague. Any attempt to find its measure will always be open to the objection that it is not clear what it is that is being measured. Is entropy the only way to measure uncertainty? Doesn't the variance also measure uncertainty? Both Shannon and Jaynes agreed that one should not place too much significance on the axiomatic derivation of eq.(4.15), that its use can be *fully* justified a posteriori by its formal properties, for example, by the various inequalities it satisfies. Thus, the standard practice is to define 'information' as a technical term using eq.(4.15) and proceed. Whether this meaning of information is in agreement with our colloquial meaning is quite another issue. However, this position can be questioned on the grounds that it is the axioms that confer meaning to the entropy; the disagreement is not about the actual equations, but about what they mean and, ultimately, about how they should be used. Other measures of uncertainty can be introduced and, indeed, they have been introduced by Renyi and by Tsallis, creating a whole industry of alternative theories [Renyi 1961, Tsallis 1988]. Whenever one can make an inference using Shannon's entropy, one can make other inferences using any one of Renyi's entropies. Which, among all those alternatives, should one choose?

The two-state case

To gain intuition about $S[p]$ consider the case of a variable that can take two values. The proverbial example is a biased coin — for example, a bent coin — for which the outcome 'heads' is assigned probability p and 'tails' probability $1-p$. The corresponding entropy, shown in figure 4.2 is

$$S(p) = -p \log p - (1-p) \log (1-p) \ , \tag{4.17}$$

where we chose $k = 1$. It is easy to check that $S \geq 0$ and that the maximum uncertainty, attained for $p = 1/2$, is $S_{\max} = \log 2$.

An important set of properties of the entropy follows from the concavity of the entropy which follows from the concavity of the logarithm. Suppose we can't decide whether the actual probability of heads is p_1 or p_2. We may decide to assign probability q to the first alternative and probability $1-q$ to the second.

Concave

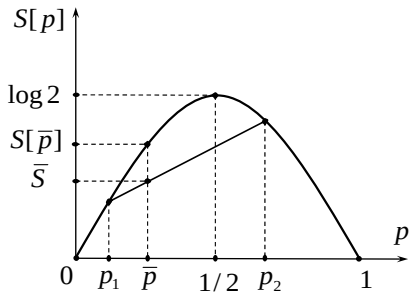


Figure 4.2: Showing the concavity of the entropy $S(\bar{p}) \geq \bar{S}$ for the case of two states.

The actual probability of heads then is the mixture $\bar{p} = qp_1 + (1-q)p_2$. The corresponding entropies satisfy the inequality

$$S(\bar{p}) \geq qS\left(p_1\right) + (1-q)\, S\left(p_2\right) = \bar{S}\ , \tag{4.18}$$

with equality in the extreme cases where $p_1 = p_2$, or $q = 0$, or $q = 1$. Eq.(4.18) says that however ignorant we might be when we invoke a probability distribution, an uncertainty about the probabilities themselves will introduce an even higher degree of ignorance.

4.2 Relative entropy

The following entropy-like quantity turns out to be useful

$$K[p,q] = +\sum_i p_i \,\log\,\frac{p_i}{q_i}\,. \tag{4.19}$$

Despite the positive sign K is sometimes read as the 'entropy of p relative to q,' and thus called "relative entropy." It is easy to see that in the special case when q_i is a uniform distribution then K is essentially equivalent to the Shannon entropy – they differ by a constant. Indeed, for $q_i = 1/n$, eq.(4.19) becomes

$$K[p,1/n] = \sum_i^n p_i\,\left(\log p_i + \log n\right) = \log n - S[p]\ . \tag{4.20}$$

The relative entropy is also known by many other names including cross entropy, information divergence, information for discrimination, and Kullback-Leibler distance [Kullback 1959]. The expression (4.19) has an old history. It was already used by Gibbs in his *Elementary Principles of Statistical Mechanics* [Gibbs 1902] and by Turing as the expected weight of evidence [Good 1983].

It is common to interpret $K[p,q]$ as the amount of information that is gained (thus the positive sign) when one thought the distribution that applies to a certain process is q and one learns that the distribution is actually p. Indeed, if the distribution q is the uniform distribution and reflects the minimum amount of information we can interpret $K[p,q]$ *as the amount of information in* p.

As we saw in section (2.10) the weight of evidence factor in favor of hypothesis θ_1 against θ_2 provided by data x is

$$w(\theta_1 : \theta_2) \stackrel{\text{def}}{=} \log \frac{p(x|\theta_1)}{p(x|\theta_2)} \ . \tag{4.21}$$

This quantity can be interpreted as the information gained from the observation of the data x. Indeed, this is precisely the way [Kullback 1959] defines the notion of 'information': the log-likelihood ratio is the information in the data x for discrimination in favor of θ_1 against θ_2. Accordingly, the relative entropy,

$$\int dx\, p(x|\theta_1) \log \frac{p(x|\theta_1)}{p(x|\theta_2)} = K(\theta_1, \theta_2) \ , \tag{4.22}$$

is interpreted as *the mean information per observation drawn from* $p(x|\theta_1)$ *in favor of* θ_1 *against* θ_2. The interpretation suffers from the same conceptual difficulties mentioned earlier concerning the Shannon entropy. In the next chapter we will see that the relative entropy turns out to be the fundamental quantity for inference – indeed, more fundamental, more general, and therefore, more useful than entropy itself – and that the interpretational difficulties that afflict the Shannon entropy can be avoided. (We will also redefine it with a negative sign, $S[p,q] \stackrel{\text{def}}{=} -K[p,q]$, so that it really is a *true* entropy.) In this chapter we just derive some properties and consider some applications.

An important property of the relative entropy is the Gibbs inequality,

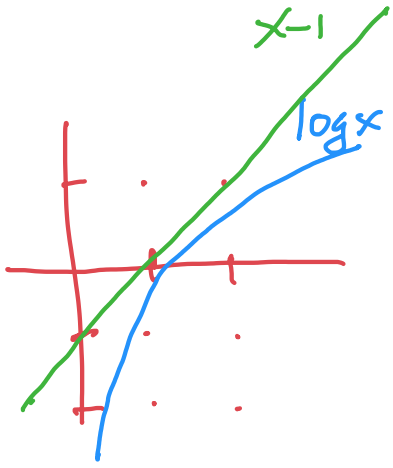
$$K[p,q] \geq 0 \ , \tag{4.23}$$

with equality if and only if $p_i = q_i$ for all i. The proof uses the concavity of the logarithm,

$$\log x \leq x - 1 \quad \text{or} \quad \log \frac{q_i}{p_i} \leq \frac{q_i}{p_i} - 1 \ , \tag{4.24}$$

which implies

$$\sum_i p_i \log \frac{q_i}{p_i} \leq \sum_i (q_i - p_i) = 0 \ . \tag{4.25}$$



The Gibbs inequality provides some justification to the common interpretation of $K[p,q]$ as a measure of the "distance" between the distributions p and q. Although useful, this language is not quite correct because $K[p,q] \neq K[q,p]$

while a true distance D is required to be symmetric, $D[p,q] = D[q,p]$. However, as we shall later see, if the two distributions are sufficiently close the relative entropy $K[p+\delta p, p]$ satisfies all the requirements of a metric. Indeed, it turns out that up to a constant factor, it is the only natural Riemannian metric on the manifold of probability distributions. It is known as the Fisher-Rao metric or, perhaps more appropriately, the information metric.

The two inequalities $S[p] \geq 0$ and $K[p,q] \geq 0$ together with eq.(4.20) imply

$$0 \leq S[p] \leq \log n \ , \tag{4.26}$$

which establishes the range of the entropy between the two extremes of complete certainty ($p_i = \delta_{ij}$ for some value j) and complete uncertainty (the uniform distribution) for a variable that takes n discrete values.

4.3 Joint entropy, additivity, and subadditivity

The entropy $S[p_x]$ reflects the uncertainty or lack of information about the variable x when our knowledge about it is codified in the probability distribution p_x. It is convenient to refer to $S[p_x]$ directly as the "entropy of the variable x" and write

$$S_x \stackrel{\text{def}}{=} S[p_x] = -\sum_x p_x \log p_x \ . \tag{4.27}$$

The virtue of this notation is its compactness but one must keep in mind the same symbol x is used to denote both a variable x and its values x_i. To be more explicit,

$$-\sum_x p_x \log p_x = -\sum_i p_x(x_i) \log p_x(x_i) \ . \tag{4.28}$$

The uncertainty or lack of information about two (or more) variables x and y is expressed by the joint distribution p_{xy} and the corresponding *joint* entropy is

$$S_{xy} = -\sum_{xy} p_{xy} \log p_{xy} \ . \tag{4.29}$$

When the variables x and y are independent, $p_{xy} = p_x p_y$, the joint entropy is *additive*

$$S_{xy} = -\sum_{xy} p_x p_y \log(p_x p_y) = S_x + S_y \ , \tag{4.30}$$

that is, the joint entropy of independent variables is the sum of the entropies of each variable. This *additivity* property also holds for the other measure of uncertainty we had introduced earlier, namely, the variance,

$$\mathrm{var}(x+y) = \mathrm{var}(x) + \mathrm{var}(y) \ . \tag{4.31}$$

In thermodynamics additivity is called *extensivity*: the entropy of an extended system is the sum of the entropies of its parts provided these parts are independent. The thermodynamic entropy can be extensive only when the interactions between various subsystems are sufficiently weak that correlations

between them can be neglected. Typically non-extensivity arises from correlations induced by short range surface effects (e.g., surface tension, wetting, capillarity) or by long-range Coulomb or gravitational forces (e.g., plasmas, black holes, etc.). Incidentally, the realization that extensivity is not a particularly fundamental property immediately suggests that it should not be given the very privileged role of a postulate in the formulation of thermodynamics.

When the two variables x and y are not independent the equality (4.30) can be generalized into an inequality. Consider the joint distribution $p_{xy} = p_x p_{y|x} = p_y p_{x|y}$. The relative entropy or Kullback "distance" of p_{xy} to the product distribution $p_x p_y$ that would represent uncorrelated variables is given by

$$\begin{aligned} K[p_{xy}, p_x p_y] &= \sum_{xy} p_{xy} \log \frac{p_{xy}}{p_x p_y} \\ &= -S_{xy} - \sum_{xy} p_{xy} \log p_x - \sum_{xy} p_{xy} \log p_y \\ &= -S_{xy} + S_x + S_y \ . \end{aligned} \tag{4.32}$$

Therefore, the Gibbs inequality, $K \geq 0$, leads to

$$S_{xy} \leq S_x + S_y \ , \tag{4.33}$$

with the equality holding when the two variables x and y are independent. This is called the *subadditivity* inequality. Its interpretation is clear: entropy increases when information about correlations is discarded.

4.4 Conditional entropy and mutual information

Consider again two variables x and y. We want to measure the amount of uncertainty about one variable x when we have some limited information about another variable y. This quantity, called the conditional entropy, and denoted $S_{x|y}$, is obtained by calculating the entropy of x as if the precise value of y were known and then taking the expectation over the possible values of y

$$S_{x|y} = \sum_y p_y S[p_{x|y}] = -\sum_y p_y \sum_x p_{x|y} \log p_{x|y} = -\sum_{x,y} p_{xy} \log p_{x|y} \ , \tag{4.34}$$

where p_{xy} is the joint distribution of x and y.

The conditional entropy is related to the entropy of x and to the joint entropy by the following "chain rule." Use the product rule for the joint distribution

$$\log p_{xy} = \log p_y + \log p_{x|y} \ , \tag{4.35}$$

and take the expectation over x and y to get

$$S_{xy} = S_y + S_{x|y} \ . \tag{4.36}$$

In words: the entropy of two variables is the entropy of one plus the conditional entropy of the other. Also, since S_y is positive we see that conditioning reduces entropy,

$$S_{xy} \geq S_{x|y} \ . \tag{4.37}$$

Another useful entropy-like quantity is the so-called "mutual information" of x and y, denoted M_{xy}, which "measures" how much information x and y have in common, or alternatively, how much information is lost when the correlations between x and y are discarded. This is given by the relative entropy between the joint distribution p_{xy} and the product distribution $p_x p_y$ that discards all information contained in the correlations. Using eq.(4.32),

$$\begin{aligned} M_{xy} \stackrel{\text{def}}{=} K[p_{xy}, p_x p_y] &= \sum_{xy} p_{xy} \log \frac{p_{xy}}{p_x p_y} \\ &= S_x + S_y - S_{xy} \geq 0 \ , \end{aligned} \tag{4.38}$$

where we used eq.(4.32). Note that M_{xy} is symmetrical in x and y. Using eq.(4.36) the mutual information is related to the conditional entropies by

$$M_{xy} = S_x - S_{x|y} = S_y - S_{y|x} \ . \tag{4.39}$$

An important application of mutual information to the problem of experimental design is given below in section 4.6.

4.5 Continuous distributions

Shannon's derivation of the expression for entropy, eq.(4.15), applies to probability distributions of discrete variables. The generalization to continuous variables is not quite straightforward.

The discussion will be carried out for a one-dimensional continuous variable; the generalization to more dimensions is trivial. The starting point is to note that the expression

$$-\int dx\, p(x) \log\, p(x) \tag{4.40}$$

is unsatisfactory. A change of variables $x \to y = y(x)$ changes the probability density $p(x)$ to $p'(y)$ but does not represent a loss or gain of information. Therefore, the actual probabilities do not change, $p(x)dx = p'(y)dy$, and neither should the entropy. However, one can check that (4.40) is not invariant,

$$\begin{aligned} \int dx\, p(x) \log\, p(x) &= \int dy\, p'(y) \log \left[p'(y) \left| \frac{dy}{dx} \right| \right] \\ &\neq \int dy\, p'(y) \log\, p'(y) \, . \end{aligned} \tag{4.41}$$

We approach the continuous case as a limit from the discrete case. Consider a continuous distribution $p(x)$ defined on an interval for $x_a \leq x \leq x_b$. Divide the interval into equal intervals $\Delta x = (x_b - x_a)/N$. For large N the distribution $p(x)$ can be approximated by a discrete distribution

$$p_n = p(x_n)\Delta x \ , \tag{4.42}$$

where $x_n = x_a + n\Delta x$ and n is an integer. The discrete entropy is

$$S_N = -\sum_{n=1}^{N} \Delta x\, p(x_n)\, \log\left[p(x_n)\Delta x\right] \quad , \tag{4.43}$$

and as $N \longrightarrow \infty$ we get

$$S_N \longrightarrow \log N - \int_{x_a}^{x_b} dx\, p(x)\, \log\left[\frac{p(x)}{1/\left(x_b - x_a\right)}\right] \tag{4.44}$$

which diverges. The divergence is what one would naturally expect: it takes a finite amount of information to identify one discrete alternative within a finite set, but it takes an infinite amount to single out one point in a continuum. The difference $S_N - \log N$ has a well defined limit and we are tempted to consider

$$-\int_{x_a}^{x_b} dx\, p(x)\, \log\left[\frac{p(x)}{1/\left(x_b - x_a\right)}\right] \tag{4.45}$$

as a candidate for the continuous entropy, until we realize that, except for an additive constant, it coincides with the unacceptable expression (4.40) and should be discarded for precisely the same reason: it is not invariant under changes of variables. Had we first changed variables to $y = y(x)$ and then discretized into N equal Δy intervals we would have obtained a different limit

$$-\int_{y_a}^{y_b} dy\, p'(y)\, \log\left[\frac{p'(y)}{1/\left(y_b - y_a\right)}\right] \quad . \tag{4.46}$$

The problem is that the limiting procedure depends on the particular choice of discretization; the limit depends on which particular set of intervals Δx or Δy we have arbitrarily decided to call equal. Another way to express the same idea is to note that the denominator $1/\left(x_b - x_a\right)$ in (4.45) represents a probability density that is uniform in the variable x, but not in y. Similarly, the density $1/\left(y_b - y_a\right)$ in (4.46) is uniform in y, but not in x.

Having identified the origin of the problem we can now suggest a solution. On the basis of our prior knowledge of the problem at hand we must decide on a privileged set of equal intervals, or alternatively, on one preferred probability distribution $\mu(x)$ we are willing to define as "uniform". Then, and only then, it makes sense to propose the following definition

$$S[p,\mu] \stackrel{\text{def}}{=} -\int_{x_a}^{x_b} dx\, p(x)\, \log\frac{p(x)}{\mu(x)} \quad . \tag{4.47}$$

It is easy to check that this is invariant,

$$\int_{x_a}^{x_b} dx\, p(x)\, \log\frac{p(x)}{\mu(x)} = \int_{y_a}^{y_b} dy\, p'(y)\, \log\frac{p'(y)}{\mu'(y)} \quad . \tag{4.48}$$

The following examples illustrate possible choices of the uniform $\mu(x)$:

1. When the variable x refers to position in "physical" Euclidean space, we can feel fairly comfortable with what we mean by equal volumes: use Cartesian coordinates and choose $\mu(x) =$ constant.

2. In a curved D-dimensional space with a known metric tensor g_{ij}, i.e., the distance between neighboring points with coordinates x^i and $x^i + dx^i$ is given by $d\ell^2 = g_{ij}dx^i dx^j$, the volume elements are given by $(\det g)^{1/2}\, d^D x$. (See the discussion in section 7.3.) The uniform distribution is that which assigns equal probabilities to equal volumes,

$$\mu(x)d^D x \propto (\det g)^{1/2}\, d^D x \ . \tag{4.49}$$

   Therefore we choose $\mu(x) \propto (\det g)^{1/2}$.

3. In classical statistical mechanics the Hamiltonian evolution in phase space is, according to Liouville's theorem, such that phase space volumes are conserved. This leads to a natural definition of equal intervals or equal volumes. The corresponding choice of uniform μ is called the postulate of "equal a priori probabilities." (See the discussion in section 5.2.)

Notice that the expression in eq.(4.47) is a relative entropy $-K[p, \mu]$. This is a hint for a theme that will be fully developed in chapter 6: relative entropy is the more fundamental quantity. Strictly, there is no Shannon entropy in the continuum – not only do we have to subtract an infinite constant and spoil its (already shaky) interpretation as an information measure, but we have to appeal to prior knowledge and introduce the measure μ. On the other hand there are no difficulties in obtaining the continuum limit from the discrete version of relative entropy. We can check that

$$K_N = \sum_{n=0}^{N} p_n \log \frac{p_n}{q_n} = \sum_{n=0}^{N} \Delta x\, p(x_n) \log \frac{p(x_n)\Delta x}{q(x_n)\Delta x} \tag{4.50}$$

has a well defined limit,

$$K[p, q] = \int_{x_a}^{x_b} dx\, p(x) \log \frac{p(x)}{q(x)} \ , \tag{4.51}$$

which is manifestly invariant under coordinate transformations.

4.6 Experimental design

A very useful and elegant application of the notion of mutual information is to the problem of experimental design. The usual problem of Bayesian data analysis is to make the best possible inferences about a certain variable θ on the basis of data obtained from a given experiment. The problem we now address concerns the decisions that must be made before the data is collected: Where should the detectors be placed? How many should there be? When

should the measurement be carried out? How to remain within the bounds of a budget? The goal is to choose the best possible experiment given a set of practical constraints. The idea is to compare the amounts of information available before and after the experiment. The difference is the amount of information provided by the experiment and this is the quantity that one must seek to maximize. The basic idea was proposed in [Lindley 1956]; a more modern application with references to the literature is [Loredo 2003].

The problem can be idealized as follows. We want to make inferences about a variable θ. Let $q(\theta)$ be the prior. We want to select the optimal experiment from within a family of experiments labeled by ε. The label ε can be discrete or continuous, one parameter or many, and each experiment ε is specified by its likelihood function $q_\varepsilon(x|\theta)$.[1]

The amount of information before the experiment is performed is given by

$$K_b = K[q,\mu] = \int d\theta\, q(\theta)\, \log \frac{q(\theta)}{\mu(\theta)}\ , \tag{4.52}$$

where $\mu(\theta)$ defines what we mean by the uniform distribution in the space of θs. If experiment ε were to be performed and data x were obtained the amount of information after the experiment would be

$$K_a(x) = K[q_\varepsilon,\mu] = \int d\theta\, q_\varepsilon(\theta|x)\, \log \frac{q_\varepsilon(\theta|x)}{\mu(\theta)}\ . \tag{4.53}$$

But the data x has not yet been collected; the expected amount of information to be obtained from experiment ε is

$$\langle K_a\rangle = \int dx d\theta\, q_\varepsilon(x) q_\varepsilon(\theta|x)\, \log \frac{q_\varepsilon(\theta|x)}{\mu(\theta)}\ , \tag{4.54}$$

where $q_\varepsilon(x)$ is the probability that data x is observed in experiment ε,

$$q_\varepsilon(x) = \int d\theta\, q_\varepsilon(x,\theta) = \int d\theta\, q(\theta) q_\varepsilon(x|\theta)\,. \tag{4.55}$$

Using Bayes theorem $\langle K_a\rangle$ can be written as

$$\langle K_a\rangle = \int dx d\theta\, q_\varepsilon(x,\theta)\, \log \frac{q_\varepsilon(x,\theta)}{q_\varepsilon(x) q(\theta)} + K_b\ . \tag{4.56}$$

Therefore, the expected information gained in experiment ε, which is $\langle K_a\rangle - K_b$, turns out to be

$$M_{x\theta}(\varepsilon) = \int dx d\theta\, q_\varepsilon(x,\theta)\, \log \frac{q_\varepsilon(x,\theta)}{q_\varepsilon(x) q(\theta)}\ , \tag{4.57}$$

which we recognize as the mutual information $M_{x\theta}$ of the data x from ε and the variable θ to be inferred, eq.(4.38). Clearly the best ε is that θ which maximizes

[1] The data x should include some label indicating the type of experiment that generated it, say x_ε. For simplicity of notation such a label is omitted.

$M_{x\theta}(\varepsilon)$ subject to whatever conditions (limited resources, etc.) apply to the situation at hand.

Incidentally, mutual information, eq.(4.38), satisfies the Gibbs inequality $M_{x\theta}(\varepsilon) \geq 0$. Therefore unless the data x and the variable θ are statistically independent (which represents a totally useless experiment in which information about one variable tells us absolutely nothing about the other) all experiments are to some extent informative, at least *on the average*. The qualification 'on the average' is important: individual data can lead to a negative information gain. Indeed, as we saw in the keys/pocket example discussed in section 4.1 a datum that turns out to be surprising can actually increase the uncertainty in θ.

An interesting special case is that of exploration experiments in which the goal is to find something [Loredo 2003]. The general background for this kind of problem is that observations have been made in the past leading to our current prior $q(\theta)$ and the problem is to decide where or when shall we make the next observation. The simplifying assumption is that we choose among experiments ε that differ only in that they are performed at different locations, in particular, the inevitable uncertainties introduced by noise are independent of ε; they are the same for all locations [Sebastiani Wynn 2000]. The goal is to identify the optimal location for the next observation. An example in astronomy could be as follows: the variable θ represents the location of a planet in the field of view of a telescope; the data x represents light intensity; and ε represents the time of observation and the orientation of the telescope.

The mutual information $M_{x\theta}(\varepsilon)$ can be written in terms of conditional entropy as in eq.(4.39). Explicitly,

$$\begin{aligned} M_{x\theta}(\varepsilon) &= \int dx d\theta\, q_\varepsilon(x,\theta)\, \log \frac{q_\varepsilon(x|\theta)}{q_\varepsilon(x)} \\ &= \int dx d\theta\, q_\varepsilon(x,\theta)\, \left[\log \frac{q_\varepsilon(x|\theta)}{\mu(x)} - \log \frac{q_\varepsilon(x)}{\mu(x)}\right] \\ &= \int d\theta\, q(\theta) \int dx\, q_\varepsilon(x|\theta)\, \log \frac{q_\varepsilon(x|\theta)}{\mu(x)} - \int dx\, q_\varepsilon(x)\, \log \frac{q_\varepsilon(x)}{\mu(x)}\ , \end{aligned}$$

where $\mu(x)$ defines what we mean by the uniform distribution in the space of xs. The assumption for these location experiments is that the noise is the same for all ε, that is, the entropy of the likelihood function $q_\varepsilon(x|\theta)$ is independent of ε. Therefore maximizing

$$M_{x\theta}(\varepsilon) = \text{const} - \int dx\, q_\varepsilon(x)\, \log \frac{q_\varepsilon(x)}{\mu(x)} = \text{const} + S_x(\varepsilon) \tag{4.58}$$

amounts to choosing the ε that maximizes the entropy of the data to be collected: we expect to learn the most by collecting data where we know the least.

4.7 Communication Theory

Here we give the briefest introduction to some basic notions of communication theory as originally developed by Shannon [Shannon 1948, Shannon Weaver 1949]. For a more comprehensive treatment see [Cover Thomas 1991].

Communication theory studies the problem of how a message that was selected at some point of origin can be reproduced at some later destination point. The complete communication system includes an *information source* that generates a message composed of, say, words in English, or pixels on a picture. A *transmitter* translates the message into an appropriate signal. For example, sound pressure is encoded into an electrical current, or letters into a sequence of zeros and ones. The signal is such that it can be transmitted over a *communication channel*, which could be electrical signals propagating in coaxial cables or radio waves through the atmosphere. Finally, a *receiver* reconstructs the signal back into a message to be interpreted by an agent at the destination point.

From the point of view of the engineer designing the communication system the challenge is that there is some limited information about the set of potential messages to be sent but it is not known which specific messages will be selected for transmission. The typical sort of questions one wishes to address concern the minimal physical requirements needed to communicate the messages that could potentially be generated by a particular information source. One wants to characterize the sources, measure the capacity of the communication channels, and learn how to control the degrading effects of noise. And after all this, it is somewhat ironic but nevertheless true that such "information theory" is completely unconcerned with whether any "information" is being communicated at all. As far as the engineer goes, whether the messages convey some meaning or not is completely irrelevant.

To illustrate the basic ideas consider the problem of data compression. A useful idealized model of an information source is a sequence of random variables $x_1, x_2, \ldots$ which take values from a finite alphabet of symbols. We will assume that the variables are independent and identically distributed. (Eliminating these limitations is both possible and important.) Suppose that we deal with a binary source in which the variables x_i, which are usually called 'bits', take the values zero or one with probabilities p or $1-p$ respectively. Shannon's idea was to classify the possible sequences $x_1, \ldots, x_N$ into *typical* and *atypical* according to whether they have high or low probability. For large N the expected number of zeros and ones is Np and $N(1-p)$ respectively. The probability of anyone of these *typical* sequences is

$$P(x_1, \ldots, x_N) \approx p^{Np}(1-p)^{N(1-p)} , \tag{4.59}$$

so that

$$-\log P(x_1, \ldots, x_N) \approx -N[p \log p - (1-p)\log(1-p)] = NS(p) \tag{4.60}$$

where $S(p)$ is the two-state entropy, eq.(4.17), the maximum value of which is $S_{\max} = \log 2$. Therefore, the probability of typical sequences is roughly

$$P(x_1, \ldots, x_N) \approx e^{-NS(p)} . \tag{4.61}$$

Since the total probability of typical sequences is less than one, we see that their number has to be less than about $e^{NS(p)}$ which for large N is considerably less than the total number of possible sequences, $2^N = e^{N \log 2}$. This fact is very significant. Transmitting an arbitrary sequence irrespective of whether it is typical or not requires a long message of N bits, but we do not have to waste resources in order to transmit all sequences. We only need to worry about the far fewer typical sequences because the atypical sequences are too rare. The number of typical sequences is about

$$e^{NS(p)} = 2^{NS(p)/\log 2} = 2^{NS(p)/S_{\max}} \tag{4.62}$$

and therefore we only need about $NS(p)/S_{\max}$ bits to identify each one of them. Thus, it must be possible to compress the original long but typical message into a much shorter one. The compression might imply some small probability of error because the actual message might conceivably turn out to be atypical but one can, if desired, avoid any such errors by using one additional bit to flag the sequence that follows as typical and short or as atypical and long. Actual schemes for implementing the data compression are discussed in [Cover Thomas 91].

Next we state these intuitive notions in a mathematically precise way.

Theorem: The Asymptotic Equipartition Property (AEP)
If $x_1, \ldots, x_N$ are independent variables with the same probability distribution $p(x)$, then

$$-\frac{1}{N} \log P(x_1, \ldots, x_N) \longrightarrow S[p] \quad \text{in probability.} \tag{4.63}$$

Proof: If the variables x_i are independent, so are functions of them such the logarithms of their probabilities, $\log p(x_i)$,

$$-\frac{1}{N} \log P(x_1, \ldots, x_N) = -\frac{1}{N} \sum_i^N \log p(x_i)\,, \tag{4.64}$$

and the law of large numbers (see section 2.7) gives

$$\lim_{N\to\infty} \text{Prob}\left[\left| -\frac{1}{N} \log P(x_1, \ldots, x_N) + \langle \log p(x) \rangle \right| \leq \varepsilon\right] = 1\,, \tag{4.65}$$

where

$$-\langle \log p(x) \rangle = S[p]\ . \tag{4.66}$$

This concludes the proof.

We can elaborate on the AEP idea further. The typical sequences are those for which eq.(4.61) or (4.63) is satisfied. To be precise let us define the typical set $A_{N,\varepsilon}$ as the set of sequences with probability $P(x_1, \ldots, x_N)$ such that

$$e^{-N[S(p)+\varepsilon]} \leq P(x_1, \ldots, x_N) \leq e^{-N[S(p)-\varepsilon]}\,. \tag{4.67}$$

Theorem of typical sequences:

(1) For N sufficiently large $\text{Prob}[A_{N,\varepsilon}] > 1-\varepsilon$.

(2) $|A_{N,\varepsilon}| \leq e^{N[S(p)+\varepsilon]}$ where $|A_{N,\varepsilon}|$ is the number of sequences in $A_{N,\varepsilon}$.

(3) For N sufficiently large $|A_{N,\varepsilon}| \geq (1-\varepsilon)e^{N[S(p)-\varepsilon]}$.

In words: the typical set has probability approaching certainty; typical sequences are nearly equally probable (thus the 'equipartition'); and there are about $e^{NS(p)}$ of them. To summarize:

The possible sequences are equally likely (well, most of them).

Proof: Eq.(4.65) states that for fixed ε, for any given δ there is an N_δ such that for all $N > N_\delta$, we have

$$\text{Prob}\left[\left|-\frac{1}{N}\log P(x_1,\ldots,x_N) + S[p]\right| \leq \varepsilon\right] \geq 1-\delta\,. \tag{4.68}$$

Thus, the probability that the sequence $(x_1,\ldots,x_N)$ is ε-typical tends to one, and therefore so must $\text{Prob}[A_{N,\varepsilon}]$. Setting $\delta = \varepsilon$ yields part **(1)**. To prove **(2)** write

$$\begin{aligned} 1 \geq \text{Prob}[A_{N,\varepsilon}] &= \sum_{(x_1,\ldots,x_N)\in A_{N,\varepsilon}} P(x_1,\ldots,x_N) \\ &\geq \sum_{(x_1,\ldots,x_N)\in A_{N,\varepsilon}} e^{-N[S(p)+\varepsilon]} = e^{-N[S(p)+\varepsilon]}\,|A_{N,\varepsilon}|\,. \end{aligned} \tag{4.69}$$

Finally, from part **(1)**,

$$\begin{aligned} 1-\varepsilon < \text{Prob}[A_{N,\varepsilon}] &= \sum_{(x_1,\ldots,x_N)\in A_{N,\varepsilon}} P(x_1,\ldots,x_N) \\ &\leq \sum_{(x_1,\ldots,x_N)\in A_{N,\varepsilon}} e^{-N[S(p)-\varepsilon]} = e^{-N[S(p)-\varepsilon]}\,|A_{N,\varepsilon}|\,, \end{aligned} \tag{4.70}$$

which proves **(3)**.

We can now quantify the extent to which messages generated by an information source of entropy $S[p]$ can be compressed. A scheme that produces compressed sequences that are longer than $NS(p)/S_{\max}$ bits is capable of distinguishing among all the typical sequences. The compressed sequences can be reliably decompressed into the original message. Conversely, schemes that yield compressed sequences of fewer than $NS(p)/S_{\max}$ bits cannot describe all typical sequences and are not reliable. This result is known as *Shannon's noiseless channel coding theorem*.

4.8 Assigning probabilities: MaxEnt

Probabilities are introduced to cope with uncertainty due to missing information. The notion that entropy $S[p]$ can be interpreted as a quantitative measure of the amount of missing information has one remarkable consequence: it provides us with a method to assign probabilities. Briefly the idea is simple: assign probabilities that do not reflect more knowledge than one actually has. More explicitly:

> *Among all possible probability distributions select that particular distribution that agrees with what we do know while reflecting least information about all else.*

The mathematical implementation of this idea involves entropy:

> *Since least information is expressed as maximum entropy, the selected distribution is that which maximizes entropy subject to whatever constraints are imposed by the available information.*

This method of reasoning is called the *Method of Maximum Entropy* and is often abbreviated as MaxEnt. Ultimately, the method of maximum entropy is based on an ethical principle of intellectual honesty that demands that one should not assume information one does not have. The idea is quite compelling but its justification relies heavily on interpreting entropy as a measure of missing information and therein lies its weakness: to what extent are we sure that entropy is the unique measure of information or of uncertainty?

As a simple illustration of MaxEnt in action consider a variable x about which absolutely nothing is known except that it can take n discrete values x_i with $i = 1 \ldots n$. The distribution that represents the state of maximum ignorance is that which maximizes the entropy $S = -\sum p \log p$ subject to the single constraint that the probabilities be normalized, $\sum p = 1$. Introducing a Lagrange multiplier α to handle the constraint, the variation $p_i \to p_i + \delta p_i$ gives

$$0 = \delta \left(S[p] - \alpha \sum_i p_i \right) = -\sum_i (\log\, p_i + 1 + \alpha)\, \delta p_i \ , \tag{4.71}$$

needs variational calculus

so that the selected distribution is

$$p_i = e^{-1-\alpha} \quad \text{or} \quad p_i = \frac{1}{n} \ , \tag{4.72}$$

where the multiplier α has been determined from the normalization constraint. We can check that the maximum value attained by the entropy,

$$S_{\max} = -\sum_i \frac{1}{n} \log \frac{1}{n} = \log n \, , \tag{4.73}$$

agrees with eq.(4.26).

Remark: The distribution of maximum ignorance turns out to be uniform. It coincides with what we would have obtained using Laplace's Principle of Insufficient Reason. It is sometimes asserted that MaxEnt provides a proof of Laplace's principle but such a claim is questionable. As we saw earlier, the privileged status of the uniform distribution was imposed through the Shannon's axioms from the very beginning.

4.9 Canonical distributions

Next we address a problem in which more information is available. The additional information is effectively a constraint that defines the family acceptable distributions. Although the constraints can take any form whatsoever in this section we develop the MaxEnt formalism for the special case of constraints that are linear in the probabilities. The most important applications are to situations of thermodynamic equilibrium where the relevant information is given in terms of the expected values of those few macroscopic variables such as energy, volume, and number of particles over which one has some experimental control. (In the next chapter we revisit this problem in detail.)

The goal is to select the distribution of maximum entropy from within the family of all distributions for which the expectations of some functions $f^k(x)$ labeled by superscripts $k = 1, 2, \ldots$ have known numerical values F^k,

$$\left\langle f^k \right\rangle = \sum_i p_i f_i^k = F^k \ , \tag{4.74}$$

where we set $f^k(x_i) = f_i^k$ for $i = 1 \ldots n$ to simplify the notation. To maximize $S[p]$ subject to (4.74) and normalization, $\sum p_i = 1$, introduce Lagrange multipliers α and λ_k,

$$\begin{aligned} 0 &= \delta \left(S[p] - \alpha \sum_i p_i - \lambda_k \left\langle f^k \right\rangle \right) \\ &= -\sum_i \left(\log\, p_i + 1 + \alpha + \lambda_k f_i^k \right) \delta p_i \ , \end{aligned} \tag{4.75}$$

where we adopt the Einstein summation convention that repeated upper and lower indices are summed over. The solution is the so-called 'canonical' distribution,

$$p_i = \exp -(\lambda_0 + \lambda_k f_i^k) \ , \tag{4.76}$$

where we have set $1 + \alpha = \lambda_0$. The normalization constraint determines λ_0,

$$e^{\lambda_0} = \sum_i \exp(-\lambda_k f_i^k) \stackrel{\text{def}}{=} Z\left(\lambda_1, \lambda_2, \ldots\right) \tag{4.77}$$

where we have introduced the partition function $Z(\lambda)$. The remaining multipliers λ_k are determined by eqs.(4.74): substituting eqs.(4.76) and (4.77) into eqs.(4.74) gives

$$-\frac{\partial \log Z}{\partial \lambda_k} = F^k \ . \tag{4.78}$$

check

This set of equations can in principle be inverted to give $\lambda_k = \lambda_k(F)$; in practice this is not usually necessary. Substituting eq.(4.76) into $S[p] = -\sum p_i \log p_i$ yields the maximized value of the entropy,

$$S_{\max} = \sum_i p_i(\lambda_0 + \lambda_k f_i^k) = \lambda_0 + \lambda_k F^k \ . \tag{4.79}$$

exponential family

Equations (4.76-4.78) are a generalized form of the "canonical" distributions first discovered by Maxwell, Boltzmann and Gibbs.

Strictly, the calculation above only shows that the entropy is stationary, $\delta S = 0$. To complete the argument we must show that (4.79) is indeed the absolute maximum rather than just a local extremum or a stationary point. Consider any other distribution q_i that satisfies precisely the same constraints, eqs.(4.74). According to the basic Gibbs inequality for the relative entropy of q and the canonical p,

$$K(q,p) = \sum_i q_i \log \frac{q_i}{p_i} \geq 0 \ , \tag{4.80}$$

or

$$S[q] \leq -\sum_i q_i \log p_i \ . \tag{4.81}$$

Substituting eq.(4.76) gives

$$S[q] \leq \sum_i q_i(\lambda_0 + \lambda_k f_i^k) = \lambda_0 + \lambda_k F^k \ . \tag{4.82}$$

Therefore

$$S[q] \leq S[p] = S_{\max} \ . \tag{4.83}$$

In words: within the family of all distributions q that satisfy the constraints (4.74) the distribution that achieves the maximum entropy is the canonical distribution p given in eq.(4.76).

Having found the maximum entropy distribution we can now develop the MaxEnt formalism along lines that closely parallel the formalism of statistical mechanics. Each distribution within the family of distributions of the form (4.76) can be thought of as a point in a continuous space – the manifold of canonical distributions. Each specific choice of expected values $(F^1, F^2, \ldots)$ determines a unique point within the space, and therefore the F^k play the role of coordinates. To each point $(F^1, F^2, \ldots)$ we can associate a number, the value of the maximized entropy. Therefore, $S_{\max}$ is a scalar field which we denote $S(F^1, F^2, \ldots) = S(F)$. In thermodynamics it is conventional to drop the suffix 'max' and to refer to $S(F)$ as *the entropy* of the system. This language can be misleading. We should constantly remind ourselves that $S(F)$ is just one out of many possible entropies that one could associate to the same physical system: $S(F)$ is that particular entropy that measures the amount of information that is missing for an agent whose knowledge consists of the numerical values of the Fs and nothing else. The multiplier

$$\lambda_0 = \log Z(\lambda_1, \lambda_2, \ldots) = \log Z(\lambda) \tag{4.84}$$

is sometimes called the "free energy" because it is closely related to the thermodynamic free energy,

$$S(F) = \log Z(\lambda) + \lambda_k F^k. \tag{4.85}$$

This shows that the quantities $S(F)$ and $\log Z(\lambda)$ are Legendre transforms of each other and therefore contain the same information. Just as the Fs are obtained from $\log Z(\lambda)$ from eq.(4.78), the λs can be obtained from $S(F)$

$$\frac{\partial S(F)}{\partial F^k} = \frac{\partial \log Z(\lambda)}{\partial \lambda_j}\frac{\partial \lambda_j}{\partial F^k} + \frac{\partial \lambda_j}{\partial F^k}F^j + \lambda_k \ . \tag{4.86}$$

Using eq.(4.78) we get

$$\frac{\partial S(F)}{\partial F^k} = \lambda_k \ , \tag{4.87}$$

which shows that the multipliers λ_k are the components of the gradient of the entropy $S(F)$ on the manifold of canonical distributions. Equivalently, $\delta S = \lambda_k \delta F^k$ is the change in entropy when the constraints are changed by δF^k (with the functions f^k held fixed).

A useful extension of the formalism is the following. Processes are common where the functions f^k can themselves be manipulated by controlling one or more "external" parameters v, $f_i^k = f^k(x_i, v)$. For example if f_i^k refers to the energy of the system when it is in state i, then the parameter v could represent the volume of the system or perhaps an externally applied magnetic field. Then a general change in the expected value F^k can be induced by changes in both f^k and λ_k,

$$\delta F^k = \delta \left\langle f^k \right\rangle = \sum_i \left(p_i \delta f_i^k + f_i^k \delta p_i \right) \ . \tag{4.88}$$

The first term on the right is

$$\left\langle \delta f^k \right\rangle = \sum_i p_i \frac{\partial f_i^k}{\partial v}\delta v = \left\langle \frac{\partial f^k}{\partial v} \right\rangle \delta v \ . \tag{4.89}$$

When F^k represents the internal energy then $\left\langle \delta f^k \right\rangle$ is a small energy transfer that can be controlled through an external parameter v. This suggests that $\left\langle \delta f^k \right\rangle$ represents a kind of "generalized work," δW^k, and the expectations $\left\langle \partial f^k / \partial v \right\rangle$ are analogues of pressure or susceptibility,

$$\delta W^k \stackrel{\text{def}}{=} \left\langle \delta f^k \right\rangle = \left\langle \frac{\partial f^k}{\partial v} \right\rangle \delta v \, . \tag{4.90}$$

The second term in eq.(4.88),

$$\delta Q^k \stackrel{\text{def}}{=} \sum_i f_i^k \delta p_i = \delta \left\langle f^k \right\rangle - \left\langle \delta f^k \right\rangle \tag{4.91}$$

is a kind of "generalized heat", and

$$\delta F^k = \delta W^k + \delta Q^k \tag{4.92}$$

is a "generalized first law." However, there is no implication that the quantity f^k is conserved (e.g., energy is a conserved quantity but magnetization is not).

The corresponding change in the entropy is obtained from eq.(4.85),

$$\begin{aligned}\delta S &= \delta \log Z(\lambda) + \delta(\lambda_k F^k) \\ &= -\frac{1}{Z}\sum_i \left[\delta\lambda_k f_i^k + \lambda_k \delta f_i^k\right] e^{-\lambda_k f_i^k} + \delta\lambda_k F^k + \lambda_k \delta F^k \\ &= \lambda_k \left(\delta \left\langle f^k \right\rangle - \left\langle \delta f^k \right\rangle\right), \end{aligned} \tag{4.93}$$

which, using eq.(4.91), gives

$$\delta S = \lambda_k \delta Q^k \ . \tag{4.94}$$

It is easy to see that this is equivalent to eq.(4.87) where the partial derivatives are derivatives at constant v. Thus the entropy remains constant in infinitesimal "adiabatic" processes — those with $\delta Q^k = 0$. From the point of view of information theory [see eq.(4.91)] this result is a triviality: the amount of information in a distribution cannot change when the probabilities do not change,

$$\delta p_i = 0 \Rightarrow \delta Q^k = 0 \Rightarrow \delta S = 0 \ . \tag{4.95}$$

4.10 On constraints and relevant information

MaxEnt is designed as a method to handle information in the form of constraints (while Bayes handles information in the form of data). The method is not at all restricted to constraints in the form of expected values (several examples will be given in later chapters) but this is a fairly common situation. To fix ideas consider a MaxEnt problem in which we maximize $S[p]$ subject to a constraint $\langle f \rangle = F$ to get a distribution $p(i|\lambda) \propto e^{-\lambda f_i}$. For example, the probability distribution that describes the state of thermodynamic equilibrium is obtained maximizing $S[p]$ subject to a constraint on the expected energy $\langle \varepsilon \rangle = E$ to yield the Boltzmann distribution $p(i|\beta) \propto e^{-\beta\varepsilon_i}$ where $\beta = 1/T$ is the inverse temperature. (See section 5.4.) The questions we address here are: How do we decide which is the right function f to choose? How do we decide the numerical value F? When can we expect the inferences to be reliable? The broader question "What is information?" shall be addressed in more detail in section 6.1.

When using the MaxEnt method to obtain, say, the canonical Boltzmann distribution it has been common to adopt the following language:

> We seek the probability distribution that codifies the information we actually have (*e.g.*, the expected energy) and is maximally unbiased (*i.e.* maximally ignorant or maximum entropy) about all the other information we do not possess.

This justification has stirred a considerable controversy that goes beyond the issue we discussed earlier of whether the Shannon entropy is the correct way

to measure information. Some of the objections that have been raised are the following:

(O1) The observed spectrum of black body radiation is whatever it is, independently of whatever information happens to be available to us.

(O2) In most realistic situations the expected value of the energy is not a quantity we happen to know. How, then, can we justify using it as a constraint?

(O3) Even when the expected values of some quantities happen to be known, there is no guarantee that the resulting inferences will be any good at all.

These objections deserve our consideration. They offer us an opportunity to attain a deeper understanding of entropic inference.

We can distinguish four epistemically different situations.

(A) **The ideal case**: We know that $\langle f \rangle = F$ and we know that it captures all the information that happens to be relevant to the problem at hand.

We have called case A the ideal situation because it reflects a situation in which the information that is necessary to reliably answer the questions that interest us is available. The requirements of both relevance and completeness are crucial. Note that a particular piece of evidence can be relevant and complete for some questions but not for others. For example, the expected energy $\langle \varepsilon \rangle = E$ is both relevant and complete for the question "Will system 1 be in thermal equilibrium with another system 2?" or alternatively, "What is the temperature of system 1?" But the same expected energy is far from relevant or complete for the vast majority of other possible questions such as, for example, "Where can we expect to find molecule #237 in this sample of ideal gas?"

Our goal here has been merely to describe the ideal epistemic situation one would like to achieve. We have not addressed the important question of how to assess whether a particular piece of evidence is relevant and complete for any specific issue at hand.

(B) **The important case**: We know that $\langle f \rangle$ captures all the information that happens to be relevant to the problem at hand but its actual numerical value F is not known.

This is the most common situation in physics. The answer to objection O2 starts from the observation that whether the value of the expected energy E is known or not, it is nevertheless still true that maximizing entropy subject to the energy constraint $\langle \varepsilon \rangle = E$ leads to the indisputably correct *family* of thermal equilibrium distributions (including, for example, the observed black-body spectral distribution). The justification behind imposing a constraint on the expected energy cannot be that the quantity E happens to be known — because of the brute fact that it is never actually known — but rather that it is the quantity that *should* be known. Even when the actual numerical value

is unknown, the epistemic situation described in case B is one in which we recognize the expected energy $\langle\varepsilon\rangle$ as the *relevant* information without which no successful predictions are possible. (In the next chapter we revisit this important question and provide the justification why it is the expected energy — and not some other conserved quantity such as $\langle\varepsilon^2\rangle$ — that is relevant to thermal equilibrium.)

Type B information is processed by allowing MaxEnt to proceed with the numerical value of $\langle\varepsilon\rangle = E$ handled as a free parameter. This leads us to the correct *family* of distributions $p(i|\beta) \propto e^{-\beta\varepsilon_i}$ containing the multiplier β as a free parameter. The actual value of the parameter β is at this point unknown. To determine it one needs additional information. The standard approach is to infer β either by a direct measurement using a thermometer, or infer it indirectly by Bayesian analysis from other empirical data.

(C) The predictive case: There is nothing special about the function f except that we happen to know its expected value, $\langle f\rangle = F$. In particular, we do not know whether information about $\langle f\rangle$ is complete or whether it is at all relevant to the problem at hand.

We do know something and this information, although limited, has some predictive value because it serves to constrain our attention to the subset of probability distributions that agree with it. Maximizing entropy subject to such a constraint will yield the best possible predictions but there is absolutely no guarantee that the predictions will be any good. Thus we see that, properly understood, objection O3 is not a flaw of the MaxEnt method; it is a legitimate warning that reasoning with incomplete information is a risky business.

(D) The extreme ignorance case: We know neither that $\langle f\rangle$ captures relevant information nor its numerical value F.

This is an epistemic situation that reflects complete ignorance. Case D applies to any arbitrary function f; it applies equally to all functions f. Since no specific f is singled out just maximize $S[p]$ subject to the normalization constraint. The result is as expected: extreme ignorance is described by a uniform distribution.

What distinguishes case C from D is that in C the value of F is actually known. This brute fact singles out a specific f and justifies using $\langle f\rangle = F$ as a constraint. What distinguishes D from B is that in B there is actual knowledge that singles out a specific f as being *relevant*. This justifies using $\langle f\rangle = F$ as a constraint. (How it comes to be that a particular f is singled out as relevant is an important question to be tackled on a case by case basis — a specific example is discussed in the next chapter.)

To summarize: between one extreme of ignorance (case D, we know neither which variables are relevant nor their expected values), and the other extreme of useful knowledge (case A, we know which variables are relevant and we also know their expected values), there are *intermediate states of knowledge* (cases B and C) — and these constitute the rule rather than the exception. Case B is the more common and important situation in which the relevant variables

have been correctly identified even though their actual expected values remain unknown. The situation described as case C is less common because information about expected values is not usually available. (What is usually available is information in the form of sample averages which is not in general quite the same thing — see the next section.)

Achieving the intermediate state of knowledge described as case B is the difficult problem presented by O2. Historically progress has been achieved in individual cases mostly by intuition and guesswork, that is, trial and error. Perhaps the seeds for a more systematic "theory of relevance" can already be seen in the statistical theories of model selection.

4.11 Avoiding pitfalls – I

The method of maximum entropy has been successful in many applications, but there are cases where it has failed or led to paradoxes and contradictions. Are these symptoms of irreparable flaws? I think not. What they are is valuable opportunities for learning. They teach us how to use the method and warn us about how not to use it; they allow us to explore its limitations; and what is perhaps most important is that they provide powerful hints for further development. Here I collect a few remarks about avoiding such pitfalls — a topic to which we shall later return (see section 8.4).

4.11.1 MaxEnt cannot fix flawed information

One point that must be made is that the issue of how the information was obtained in the first place should not be confused with the issue of how information is processed — which is the problem that MaxEnt is supposed to address. These are two separate issues.

The first issue is concerned with the prior judgements that are involved in assessing whether a particular piece of data or constraint or proposition is deemed worthy of acceptance as "information", that is, whether it is "true" or at least sufficiently reliable to provide the basis for the assignment of other probabilities. The particular process of how a particular piece of information was obtained — whether the data itself is uncertain — can serve to qualify and modify the information being processed. Once this first step has been completed and a sufficiently reliable information has been accepted one proceeds to tackle the second step of processing the newly available information.

MaxEnt only claims to address the second issue: once a constraint has been accepted as information, MaxEnt answers the question "What precise rule does one follow to assign probabilities?" Had the "information" turned out to be "false" our inferences about the world could be wildly misleading, but it is not the MaxEnt method that should be blamed for this failure. MaxEnt cannot fix flawed information nor should we expect it.

4.11.2 MaxEnt cannot supply missing information

It is not uncommon that we may find ourselves in situations where our intuition insists that our inferences are not right — this applies to inferences based on Bayes' rule just as much as MaxEnt or their (later) generalization into entropic inference (see chapter 6). The right way to proceed is to ask: How can we tell that something is wrong? The answer is that we must know something else about which we are not fully aware and that it is this something that clashes with our inferences. At the very least this is telling us that we had previous, although perhaps unrecognized, expectations and that we should reconsider them. The result of such analysis might indicate that those expectations were misguided — we have here an opportunity to educate our intuition and learn. Alternatively, the analysis might vindicate the earlier expectations and this is valuable too. It tells us that there existed additional prior information that happened to be relevant but, not having recognized it, we failed to take it into account. Either way, the right way to handle such situations is not to blame the method: first blame the user.

4.11.3 Sample averages are not expected values

Here is an example of a common temptation. A lucid analysis of the issues involved is given in [Uffink 1996]. Once we accept that certain constraints might refer to the expected values of certain variables, how do we decide their numerical magnitudes? The numerical values of expectations are seldom known and it is tempting to replace expected values by sample averages because it is the latter that are directly available from experiment. But the two are not the same: *Sample averages are experimental data. Expected values are not experimental data.*

For very large samples such a replacement can be justified by the law of large numbers — there is a high probability that sample averages will approximate the expected values. However, for small samples using one as an approximation for the other can lead to incorrect inferences. It is important to realize that these incorrect inferences do not represent an intrinsic flaw of the MaxEnt method; they are merely a warning of how the MaxEnt method should not be used.

Example – just data:

Here is a variation on the same theme. Suppose data $D = (x_1, x_2 \ldots x_n)$ has been collected. We might be tempted to maximize $S[p]$ subject to a constraint $\langle x \rangle = C_1$ where C_1 is unknown and then try to estimate C_1 from the data. The difficulty arises when we realize that if we know the data $(x_1, \ldots)$ then we also know their squares $(x_1^2, \ldots)$ and their cubes and also any arbitrary function of them $(f(x_1), \ldots)$. Which of these should we use as an expected value constraint? Or should we use all of them? The answer is that the MaxEnt method was not designed to tackle the kind of problem where the only information is data $D = (x_1, x_2 \ldots x_n)$. It is not that MaxEnt gives a wrong answer; it gives no answer at all because there is no constraint to impose; the MaxEnt engine cannot even get started.

Example – case B plus data:

One can imagine a different problem in order to see how MaxEnt could get some traction. Suppose, for example, that in addition to the data $D = (x_1, x_2 \ldots x_n)$ collected in n independent experiments we have additional information that singles out a specific function $f(x)$. Here we deal with an epistemic situation that was described as type B in the previous section: the expectation $\langle f \rangle$ captures relevant information. We proceed to maximize entropy imposing the constraint $\langle f \rangle = F$ with F treated as a free parameter. If the variable x can take k discrete values labeled by α we let $f(x_\alpha) = f_\alpha$ and the result is a canonical distribution

$$p(x_\alpha|\lambda) = \frac{e^{-\lambda f_\alpha}}{Z} \quad \text{where} \quad Z = \sum_{\alpha=1}^{k} e^{-\lambda f_\alpha} \tag{4.96}$$

with an unknown multiplier λ that can be estimated from the data D using Bayesian methods. If the n experiments are independent Bayes rule gives,

$$p(\lambda|D) = \frac{p(\lambda)}{p(D)} \prod_{j=1}^{n} \frac{e^{-\lambda f_j}}{Z} \ , \tag{4.97}$$

where $p(\lambda)$ is the prior. It is convenient to consider the logarithm of the posterior,

$$\begin{aligned} \log p(\lambda|D) &= \log p(\lambda) - \sum_{j=1}^{n} (\log Z + \lambda f_j) \\ &= \log p(\lambda) - n(\log Z + \lambda \bar{f}) \ , \end{aligned} \tag{4.98}$$

where $\bar{f}$ is the sample average,

$$\bar{f} = \frac{1}{n} \sum_{j=1}^{n} f_j \ . \tag{4.99}$$

The value of λ that maximizes the posterior $p(\lambda|D)$ is such that

$$\frac{\partial \log Z}{\partial \lambda} + \bar{f} = \frac{1}{n} \frac{\partial \log p(\lambda)}{\partial \lambda} \ . \tag{4.100}$$

As $n \to \infty$ the right hand side vanishes and we see that the optimal λ is such that

$$\langle f \rangle = -\frac{\partial \log Z}{\partial \lambda} = \bar{f} \tag{4.101}$$

This is to be expected: for large n the data overwhelms the prior and $\bar{f}$ tends to $\langle f \rangle$ (in probability). But the result eq.(4.100) also shows that when n is not so large then the prior can make a non-negligible contribution. In general one should not assume that $\langle f \rangle \approx \bar{f}$.

Let us emphasize that this analysis holds only when the selection of a privileged function $f(x)$ can be justified by additional knowledge about the physical nature of the problem. In the absence of such information we are back to the

previous example — just data — and we have no reason to prefer the distribution $e^{-\lambda f_j}$ over any other canonical distribution $e^{-\lambda g_j}$ for any arbitrary function $g(x)$.[2]

[2]Our conclusion differs from that reached in [Jaynes 1978, pp. 72-75] which did not include the effect of the prior $p(\lambda)$.

Chapter 5

Statistical Mechanics

Among the various theories that make up what we call physics, thermodynamics holds a very special place because it provided the first example of a fundamental theory that could be interpreted as a procedure for processing relevant information. Our goal in this chapter is to provide an explicit discussion of statistical mechanics as an example of entropic inference.

The challenge in constructing the models that we call theoretical physics lies in identifying the subject matter (the microstates) and the information (the constraints, the macrostates) that happens to be relevant to the problem at hand. First we consider the microstates and provide some necessary background on the dynamical evolution of probability distributions — Liouville's theorem — and use it to derive the so-called "postulate" of Equal a Priori Probabilities. Next, we show that for situations of thermal equilibrium the relevant information is encapsulated into a constraint on the expected value of the energy. Depending on the specific problem one can also include additional constraints on other conserved quantities such as number of particles or volume. Once the foundation has been established we can proceed to explore some consequences. We show how several central topics such as the second law of thermodynamics, irreversibility, reproducibility, and the Gibbs paradox can be considerably clarified when viewed from the information/inference perspective.

5.1 Liouville's theorem

Perhaps the most *relevant*, and therefore, most *important* piece of information that has to be incorporated into any inference about physical systems is that their time evolution is constrained by equations of motion. Whether these equations — those of Newton, Maxwell, Yang and Mills, or Einstein — can themselves be derived as examples of inference are questions which will not concern us at this point. (Later, in chapter 9 we revisit this question and show that quantum mechanics and its classical limit can also be derived as theories of inference.)

To be specific, in this chapter we will limit ourselves to discussing classical systems such as fluids. In this case there is an additional crucial piece of relevant information: these systems are composed of molecules. For simplicity we will assume that the molecules have no internal structure, that they are described by their positions and momenta, and that they behave according to classical mechanics.

The import of these remarks is that the proper description of the *microstate* of a fluid of N particles in a volume V is in terms of a "vector" in the N-particle phase space, $z = (\vec{x}_1, \vec{p}_1, \ldots \vec{x}_N, \vec{p}_N)$. The time evolution is given by Hamilton's equations,

$$\frac{d\vec{x}_i}{dt} = \frac{\partial H}{\partial \vec{p}_i} \quad \text{and} \quad \frac{d\vec{p}_i}{dt} = -\frac{\partial H}{\partial \vec{x}_i}\,, \tag{5.1}$$

where H is the Hamiltonian,

$$H = \sum_{i=1}^{N} \frac{p_i^2}{2m} + U(\vec{x}_1, \ldots \vec{x}_N, V)\ . \tag{5.2}$$

What makes phase space so convenient for the formulation of mechanics is that Hamilton's equations are first order in time. This means that through any given point $z(t_0)$, which can be thought as the initial condition, there is passes just one trajectory $z(t)$ and therefore trajectories can never intersect each other.

In a fluid the actual positions and momenta of the molecules are unknown and thus the *macrostate* of the fluid is described by a probability density in phase space, $f(z,t)$. When the system evolves continuously according to Hamilton's equations there is no information loss and the probability flow satisfies a local conservation equation,

$$\frac{\partial}{\partial t} f(z,t) = -\nabla_z \cdot J(z,t)\ , \tag{5.3}$$

where the probability current $J(z,t)$ is a vector with $6N$ components given by

$$J(z,t) = f(z,t)\dot{z} = \left(f(z,t)\frac{d\vec{x}_i}{dt},\, f(z,t)\frac{d\vec{p}_i}{dt} \right)\ . \tag{5.4}$$

Evaluating the divergence explicitly using (5.1) gives

$$\begin{aligned} \frac{\partial f}{\partial t} &= -\sum_{i=1}^{N} \left[\frac{\partial}{\partial \vec{x}_i} \cdot \left(f(z,t)\frac{d\vec{x}_i}{dt} \right) + \frac{\partial}{\partial \vec{p}_i} \cdot \left(f(z,t)\frac{d\vec{p}_i}{dt} \right) \right] \\ &= -\sum_{i=1}^{N} \left(\frac{\partial f}{\partial \vec{x}_i} \cdot \frac{\partial H}{\partial \vec{p}_i} - \frac{\partial f}{\partial \vec{p}_i} \cdot \frac{\partial H}{\partial \vec{x}_i} \right)\ . \end{aligned} \tag{5.5}$$

Thus the time derivative of $f(z,t)$ at a fixed point z is given by the Poisson bracket with the Hamiltonian H,

$$\frac{\partial f}{\partial t} = \{H, f\} \stackrel{\text{def}}{=} \sum_{i=1}^{N} \left(\frac{\partial H}{\partial \vec{x}_i} \cdot \frac{\partial f}{\partial \vec{p}_i} - \frac{\partial H}{\partial \vec{p}_i} \cdot \frac{\partial f}{\partial \vec{x}_i} \right)\ . \tag{5.6}$$

This is called the Liouville equation.

Two important corollaries are the following. Instead of focusing on the change in $f(z,t)$ at a fixed point z as in eq.(5.6) we can study the change in $f\left(z(t),t\right)$ at a point $z(t)$ as it is being carried along by the flow. This defines the so-called "convective" time derivative,

$$\frac{d}{dt}f\left(z(t),t\right)=\frac{\partial}{\partial t}f(z,t)+\sum_{i=1}^{N}\left(\frac{\partial f}{\partial\vec{x}_i}\cdot\frac{d\vec{x}_i}{dt}+\frac{\partial f}{\partial\vec{p}_i}\cdot\frac{d\vec{p}_i}{dt}\right)\ . \tag{5.7}$$

Using Hamilton's equations shows that the second term is $-\{H,f\}$ and cancels the first, therefore

$$\frac{d}{dt}f\left(z(t),t\right)=0\ , \tag{5.8}$$

which means that f is constant along a flow line. Explicitly,

$$f\left(z(t),t\right)=f\left(z(t'),t'\right)\ . \tag{5.9}$$

Next consider a small volume element $\Delta z(t)$ the boundaries of which are carried along by the fluid flow. Since trajectories cannot cross each other (because Hamilton's equations are first order in time) they cannot cross the boundary of the evolving volume $\Delta z(t)$ and therefore the total probability within $\Delta z(t)$ is conserved,

$$\frac{d}{dt}\,\mathrm{Prob}[\Delta z(t)]=\frac{d}{dt}[\Delta z(t)f\left(z(t),t\right)]=0\ . \tag{5.10}$$

But f itself is constant, eq.(5.8), therefore

$$\frac{d}{dt}\Delta z(t)=0\ , \tag{5.11}$$

which means that the shape of a region of phase space may get deformed by time evolution but its volume remains invariant. This result is usually known as Liouville's theorem.

5.2 Derivation of Equal a Priori Probabilities

Earlier, in section 4.5, we pointed out that a proper definition of entropy in a continuum, eq.(4.47), requires that one specify a privileged background measure $\mu(z)$,

$$S[f,\mu]=-\int dz\, f(z)\log\frac{f(z)}{\mu(z)}\ , \tag{5.12}$$

where $dz=d^{3N}xd^{3N}p$. The choice of $\mu(z)$ is important: it determines what we mean by a uniform or maximally ignorant distribution.

It is customary to set $\mu(z)$ equal to a constant which we might as well choose to be $\mu(z)=1$. This amounts to *postulating* that equal volumes of phase space are assigned the same a priori probabilities. Ever since the introduction of Boltzmann's ergodic hypothesis there have been many failed attempts to derive

it from purely dynamical considerations. It is easy to imagine alternatives that could appear to be just as plausible. One could, for example, divide phase space in slices of constant energy and assign equal probabilities to equal energy intervals. In this section we want to *derive* $\mu(z)$ by proving the following theorem
Theorem on Equal a Priori Probabilities: *Since* Hamiltonian dynamics involves no loss of information, *if the entropy* $S[f,\mu]$ *is to be interpreted as the measure of amount of information*, then $\mu(z)$ must be a uniform measure over phase space.
Proof: The main non-dynamical hypothesis is that entropy measures information. The *information* entropy of the time-evolved distribution $f(z,t)$ is

$$S(t) = -\int dz\, f(z,t) \log \frac{f(z,t)}{\mu(z)} \ . \tag{5.13}$$

The first input from Hamiltonian dynamics is that information is not lost and therefore we must require that $S(t)$ be constant,

$$\frac{d}{dt} S(t) = 0 \ . \tag{5.14}$$

Therefore,

$$\frac{d}{dt} S(t) = -\int dz \left[\frac{\partial f(z,t)}{\partial t} \log \frac{f(z,t)}{\mu(z)} + \frac{\partial f(z,t)}{\partial t} \right] \ . \tag{5.15}$$

The second term vanishes,

$$\int dz\, \frac{\partial f(z,t)}{\partial t} = \frac{d}{dt} \int dz\, f(z,t) = 0 \ . \tag{5.16}$$

A second input from Hamiltonian dynamics is that probabilities are not merely conserved, they are locally conserved, which is expressed by eqs.(5.3) and (5.4). The first term of eq.(5.15) can be rewritten,

$$\frac{d}{dt} S(t) = \int dz\, \nabla_z \cdot J(z,t) \log \frac{f(z,t)}{\mu(z)} \ , \tag{5.17}$$

so that integration by parts (the surface term vanishes) gives

$$\begin{aligned} \frac{d}{dt} S(t) &= -\int dz\, f(z,t) \dot{z} \cdot \nabla_z \log \frac{f(z,t)}{\mu(z)} \\ &= \int dz\, \left[-\dot{z} \cdot \nabla_z f(z,t) + f(z,t) \dot{z} \cdot \nabla_z \log \mu(z) \right] \ . \end{aligned} \tag{5.18}$$

Hamiltonian dynamics enters here once again: the first term vanishes by Liouville's equation (5.6),

$$-\int dz\, \dot{z} \cdot \nabla_z f(z,t) = \int dz\, \{H, f(z,t)\} = \int dz\, \frac{\partial f(z,t)}{\partial t} = 0 \ , \tag{5.19}$$

and therefore, imposing (5.14),

$$\frac{d}{dt}S(t) = \int dz\, f(z,t)\dot{z}\cdot\nabla_z \log\mu(z) = 0\ . \tag{5.20}$$

This integral must vanish for any arbitrary choice of the distribution $f(z,t)$, therefore

$$\dot{z}\cdot\nabla_z \log\mu(z) = 0\ . \tag{5.21}$$

Furthermore, we have considerable freedom about the particular Hamiltonian operating on the system. We could choose to change the volume in any arbitrarily prescribed way by pushing on a piston to change the volume, or we could choose to vary an external magnetic field. Either way we can change $H(t)$ and therefore $\dot{z}$ at will. The time derivative dS/dt must still vanish irrespective of the particular choice of the vector $\dot{z}$. We conclude that

$$\nabla_z \log\mu(z) = 0 \quad \text{or} \quad \mu(z) = \text{const}\,. \tag{5.22}$$

To summarize: the requirement that information is not lost in Hamiltonian dynamics implies that the measure of information must be a constant of the motion,

$$\frac{d}{dt}S(t) = 0\ , \tag{5.23}$$

and this singles out the Gibbs entropy,

$$S(t) = -\int dz\, f(z,t)\log f(z,t)\ , \tag{5.24}$$

(in $6N$-dimensional configuration space) as the correct *information* entropy.

It is sometimes asserted that (5.23) implies that the Gibbs entropy cannot be identified with the *thermodynamic* entropy because this would be in contradiction to the second law. As we shall see below, this is not true; in fact, it is quite the opposite.

Remark: In section 4.1 we pointed out that the interpretation of entropy $S[p,\mu]$ as a measure of information has its shortcomings. This could potentially undermine our whole program of deriving statistical mechanics as an example of entropic inference. Fortunately, as we shall see later in chapter 6 the framework of entropic inference can be considerably strengthened by removing any reference to questionable information measures. In this approach entropy $S[p,\mu]$ requires no interpretation; it is a tool designed for updating from a prior μ to a posterior p distribution. More explicitly the entropy $S[p,\mu]$ is introduced to rank distributions candidate p according to "preference" relative to a prior μ in accordance to certain "reasonable" design specifications. Recasting statistical mechanics into this entropic inference framework is straightforward. For example, the requirement that Hamiltonian time evolution does not affect the ranking of distributions — that is, if $f_1(z,t)$ is preferred over $f_2(z,t)$ at time t then the corresponding $f_1(z,t')$ is preferred over $f_2(z,t')$ at any other time t' — is expressed through eq.(5.14) so the proof of the Equal a Priori Theorem proceeds exactly as above.

5.3 The relevant constraints

Thermodynamics is mostly concerned with situations of thermal equilibrium. What is the relevant information needed to make inferences in these special cases? A problem here is that the notion of relevance is relative — a particular piece of information might be relevant to one specific question and irrelevant to another. So in addition to the explicit assumption of equilibrium we will also need to make a somewhat more vague assumption that our general interest is in those questions that are the typical concern of thermodynamics, namely, questions involving equilibrium macrostates and the processes that take us from one to another.

The first condition we must impose on $f\,(z,t)$ to describe equilibrium is that it be independent of time. Thus we require that $\{H,f\}=0$ and f must be a function of conserved quantities such as energy, momentum, angular momentum, or number of particles. But we do not want f to be merely stationary, as say, for a rotating fluid, we want it to be truly static. We want f to be invariant under time reversal. For these problems it turns out that it is not necessary to impose that the total momentum and total angular momentum vanish; these constraints will turn out to be satisfied automatically. To simplify the situation even more we will only consider problems where the number of particles is held fixed. Processes where particles are exchanged as in the equilibrium between a liquid and its vapor, or where particles are created and destroyed as in chemical reactions, constitute an important but straightforward extension of the theory.

It thus appears that it is sufficient to impose that f be some function of the energy. According to the formalism developed in section 4.9 and the remarks in 4.10 this is easily accomplished: the constraints codifying the information that could be relevant to problems of thermal equilibrium should be the expected values of functions $\phi(\varepsilon)$ of the energy. For example, $\langle\phi(\varepsilon)\rangle$ could include various moments, $\langle\varepsilon\rangle$, $\langle\varepsilon^2\rangle$,... or perhaps more complicated functions. The remaining question is which functions $\phi(\varepsilon)$ and how many of them.

To answer this question we look at thermal equilibrium from the point of view leading to what is known as the *microcanonical formalism*. Let us enlarge our description to include the system of interest A and its environment, that is, the thermal bath B with which it is in equilibrium. The advantage of this broader view is that the composite system $C=A+B$ can be assumed to be isolated and *we know that its energy ε_c is some fixed constant*. This is highly relevant information: when the value of ε_c is known, not only do we know $\langle\varepsilon_c\rangle=\varepsilon_c$ but we know the expected values $\langle\phi(\varepsilon_c)\rangle=\phi(\varepsilon_c)$ for absolutely all functions $\phi(\varepsilon_c)$. In other words, in this case we have succeeded in identifying the relevant information and we are finally ready to assign probabilities using the MaxEnt method. (When the value of ε_c is not known we are in that state of "intermediate" knowledge described as case (B) in section 4.10.)

To simplify the notation it is convenient to divide phase space into discrete cells of equal a priori probability so we can use the discrete Shannon entropy. By the equal a priori theorem the cells are of equal phase volume. For system A let the (discretized) microstate z_a have energy ε_a. For the thermal bath B a

much less detailed description is sufficient. Let the number of bath microstates with energy ε_b be $\Omega_B(\varepsilon_b)$. Our relevant information includes the fact that A and B interact very weakly, just barely enough to attain equilibrium, and thus the known total energy ε_c constrains the allowed microstates of $A+B$ to the subset that satisfies

$$\varepsilon_a + \varepsilon_b = \varepsilon_c \ . \tag{5.25}$$

The total number of such microstates is

$$\Omega(\varepsilon_c) = \textstyle\sum_a \Omega_B(\varepsilon_c - \varepsilon_a) \ . \tag{5.26}$$

We are in a situation where we know absolutely nothing beyond the fact that the composite system C can be in any one of its $\Omega(\varepsilon_c)$ allowed microstates. This is precisely the problem tackled in section 4.8: the maximum entropy distribution is uniform, eq.(4.72), the probability of any microstate of C is $1/\Omega(\varepsilon_c)$, and the entropy is $S_c = k \log \Omega(\varepsilon_c)$. More importantly, the probability that system A is in the particular microstate a with energy ε_a when it is in thermal equilibrium with the bath B is

$$p_a = \frac{\Omega_B(\varepsilon_c - \varepsilon_a)}{\Omega(\varepsilon_c)} \ . \tag{5.27}$$

This is the result we sought; now we need to interpret it. There is one final piece of relevant information we can use: the thermal bath B is usually much larger than system A, $\varepsilon_c \gg \varepsilon_a$, then it is convenient to rewrite p_a as

$$p_a \propto \exp \log \Omega_B(\varepsilon_c - \varepsilon_a) \ . \tag{5.28}$$

and Taylor expand

$$\log \Omega_B(\varepsilon_c - \varepsilon_a) = \log \Omega_B(\varepsilon_c) - \beta\varepsilon_a + \dots \ , \tag{5.29}$$

where the inverse temperature $\beta = 1/kT$ of the bath has been introduced according to the standard thermodynamic definition,

$$\left.\frac{\partial \log \Omega_B}{\partial \varepsilon_b}\right|_{\varepsilon_c} \stackrel{\text{def}}{=} \beta \ . \tag{5.30}$$

and we conclude that the distribution that codifies the relevant information about equilibrium is

$$p_a = \frac{1}{Z} \exp(-\beta\varepsilon_a) \ , \tag{5.31}$$

which has the canonical form of eq.(4.76). (Being independent of a the factor $\Omega_B(\varepsilon_c)/\Omega(\varepsilon_c)$ has been absorbed into the normalization Z.)

Our goal in this section was to identify the relevant variables. Here is the answer: the relevant information about thermal equilibrium can be summarized by the expected value of the energy $\langle\varepsilon\rangle$ because someone who just knows $\langle\varepsilon\rangle$ and is maximally ignorant about everything else is led to assign probabilities according to eq.(4.76) which coincides with (5.31).

But our analysis has also disclosed an important limitation. Eq.(5.27) shows that in general the distribution for a system in equilibrium with a bath depends in a complicated way on the properties of the bath. The information in $\langle\varepsilon\rangle$ is adequate only when the system and the bath interact weakly enough that the energy of the composite system C can be neatly partitioned into the energies of A and of B, eq.(5.25), and the bath is so much larger than the system that its effects can be represented by a single parameter, the temperature T.

Conversely, if these conditions are not met, then more information is needed. When the system-bath interactions are not sufficiently weak eq.(5.25) will not be valid and additional information concerning the correlations between A and B will be required. On the other hand if the system-bath interactions are too weak then within the time scales of interest the system A will reach only a partial thermal equilibrium with those few degrees of freedom in its very immediate vicinity. It is effectively surrounded by a thermal bath of finite size and the information contained in the single parameter β or the expected value $\langle\varepsilon\rangle$ will not suffice. This situation is briefly discussed in section 5.5.

So what's the big deal?

We have identified all the ingredients required to derive (see next section) the canonical formalism of statistical mechanics as an example of entropic inference. We saw that the identification of $\langle\varepsilon\rangle$ as relevant information relied on the micro-canonical formalism in an essential way. Does this mean that the information theory approach was ultimately unnecessary? That MaxEnt adds nothing to our understanding of statistical mechanics? Absolutely not.

Alternative derivations of statistical mechanics all rely on invoking the right cocktail of *ad hoc* hypothesis such as an ergodic assumption or a postulate for equal a priori probabilities. This is not too bad; all theories, MaxEnt included, require assumptions. Where MaxEnt can claim an unprecedented success is that the assumptions it does invoke are not at all *ad hoc*; they are precisely the type of assumptions one would naturally expect of any theory of inference — a specification of the subject matter (the microstates plus their underlying measure) plus an identification of the relevant constraints. Ultimately the justification of any formal system must be pragmatic: does the entropic model successfully predict, explain and unify? As we shall see in the next sections the answer is: yes.

5.4 The canonical formalism

We consider a system in thermal equilibrium. The energy of the (conveniently discretized) microstate z_a is $\varepsilon_a = \varepsilon_a(V)$ where V represents a parameter over which we have experimental control. For example, in fluids V is the volume of the system. We assume further that the expected value of the energy is known, $\langle\varepsilon\rangle = E$.

Maximizing the (discretized) Gibbs entropy,

$$S[p] = -k \sum_a p_a \log p_a \quad \text{where} \quad p_a = f(z_a)\Delta z \ , \tag{5.32}$$

subject to constraints on normalization and energy $\langle \varepsilon \rangle = E$ yields, eq.(4.76),

$$p_a = \frac{1}{Z} e^{-\beta \varepsilon_a} \tag{5.33}$$

where the Lagrange multiplier β is determined from

$$-\frac{\partial \log Z}{\partial \beta} = E \quad \text{and} \quad Z(\beta, V) = \textstyle\sum_a e^{-\beta \varepsilon_a} \ . \tag{5.34}$$

The maximized value of the Gibbs entropy is, eq.(4.79),

$$S(E, V) = k \log Z + k\beta E \ . \tag{5.35}$$

Differentiating with respect to E we obtain the analogue of eq.(4.87),

$$\left(\frac{\partial S}{\partial E} \right)_V = k \frac{\partial \log Z}{\partial \beta} \frac{\partial \beta}{\partial E} + k \frac{\partial \beta}{\partial E} E + k\beta \ = k\beta \ , \tag{5.36}$$

where eq.(5.34) has been used to cancel the first two terms.

The connection between the formalism above and thermodynamics hinges on a suitable identification of internal energy, work and heat. The first step is the crucial one: we adopt Boltzmann's assumption, eq.(3.37), and identify $\langle \varepsilon \rangle = E$ with the thermodynamical internal energy. Next we consider a small change in the internal energy,

$$\delta E = \delta \textstyle\sum_a p_a \varepsilon_a = \sum_a p_a \delta \varepsilon_a + \sum_a \varepsilon_a \delta p_a \, . \tag{5.37}$$

Since $\varepsilon_a = \varepsilon_a(V)$ the first term $\langle \delta \varepsilon \rangle$ on the right can be physically induced by pushing or pulling on a piston to change the volume,

$$\langle \delta \varepsilon \rangle = \textstyle\sum_a p_a \dfrac{\partial \varepsilon_a}{\partial V} \delta V = \left\langle \dfrac{\partial \varepsilon}{\partial V} \right\rangle \delta V \ . \tag{5.38}$$

Thus, it is reasonable to identify $\langle \delta \varepsilon \rangle$ with mechanical work,

$$\langle \delta \varepsilon \rangle = \delta W = -P \delta V \ , \tag{5.39}$$

where P is the pressure,

$$P = -\left\langle \frac{\partial \varepsilon}{\partial V} \right\rangle \ . \tag{5.40}$$

Having identified the work δW the second term in eq.(5.37) must therefore represent heat,

$$\delta Q = \delta E - \delta W = \delta \left\langle \varepsilon \right\rangle - \left\langle \delta \varepsilon \right\rangle . \tag{5.41}$$

The corresponding change in entropy is obtained from eq.(5.35),

$$\begin{aligned}\frac{1}{k}\delta S &= \delta \log Z + \delta(\beta E) \\ &= -\frac{1}{Z}\sum\nolimits_a e^{-\beta\varepsilon_a}\left(\varepsilon_a\delta\beta + \beta\delta\varepsilon_a\right) + E\delta\beta + \beta\delta E \\ &= \beta(\delta E - \langle\delta\varepsilon\rangle) \ , \end{aligned} \tag{5.42}$$

therefore,

$$\delta S = k\beta\delta Q \ . \tag{5.43}$$

In thermodynamics temperature is defined by

$$\left(\frac{\partial S}{\partial E}\right)_V \stackrel{\text{def}}{=} \frac{1}{T} \quad \text{so that} \quad \delta S = \frac{\delta Q}{T} \ . \tag{5.44}$$

which suggests the identification

$$k\beta = \frac{1}{T} \quad \text{or} \quad \beta = \frac{1}{kT} \ . \tag{5.45}$$

Therefore the maximized information entropy, $S(E,V)$, corresponds to the thermodynamic entropy originally introduced by Clausius and the Lagrange multiplier β corresponds to the inverse temperature.

Thus, the framework of entropic inference provides a natural *explanation* for both the temperature and the thermodynamic entropy. These are precisely the kind of theoretical concepts that must inevitably appear in all theories of inference.[1]

Substituting into eq.(5.41), yields the *fundamental thermodynamic identity*,

$$\delta E = T\delta S - P\delta V \ . \tag{5.46}$$

Incidentally, this identity shows that the "natural" variables for energy are S and V, that is, $E = E(S,V)$. Similarly, writing

$$\delta S = \frac{1}{T}\delta E + \frac{P}{T}\delta V \tag{5.47}$$

confirms that $S = S(E,V)$.

Equation (5.46) is useful either for processes at constant V so that $\delta E = \delta Q$, or for processes at constant S for which $\delta E = \delta W$. But except for these latter adiabatic processes ($\delta Q = 0$) the entropy is not a quantity that can be directly controlled in the laboratory. For processes that occur at constant temperature

[1] It might not be a bad idea to stop for a moment and let this marvelous notion sink in: temperature, that which we identify with hot things being hot and cold things being cold is, in the end, nothing but a Lagrange multiplier. It turns out that temperature is in some common cases also a measure of mean kinetic energy per molecule. This conception is useful but limited; it fails to capture the full significance of the concept of temperature.

it is more convenient to introduce a new quantity, called the free energy, that is a function of T and V. The free energy is given by a Legendre transform,

$$F(T,V) = E - TS \ , \tag{5.48}$$

so that

$$\delta F = -S\delta T - P\delta V \ . \tag{5.49}$$

For processes at constant T we have $\delta F = \delta W$ which justifies the name 'free' energy. Eq.(5.35) then leads to

$$F = -kT \log Z(T,V) \quad \text{or} \quad Z = e^{-\beta F} \ . \tag{5.50}$$

Several useful thermodynamic relations can be easily obtained from eqs.(5.46), (5.47), and (5.49). For example, the identities

$$\left(\frac{\partial F}{\partial T}\right)_V = -S \quad \text{and} \quad \left(\frac{\partial F}{\partial V}\right)_V = -P \, , \tag{5.51}$$

can be read directly from eq.(5.49).

5.5 Equilibrium with a heat bath of finite size

In section 5.3 we saw that the canonical Boltzmann-Gibbs distribution applies to situations where the system is in thermal equilibrium with an environment that is much larger than itself. But this latter condition can be violated. For example, when we deal with very fast phenomena or in situations where the system-environment interactions are very weak then over the time scales of interest the system will reach a partial equilibrium but only with those few degrees of freedom in its immediate vicinity. In such cases the effective environment has a finite size and the information contained in the single parameter β will not suffice. We will not pursue the subject beyond giving the briefest hints about how to approach the subject.

One might account for such finite size effects by keeping additional terms in the expansion (5.29),

$$\log \Omega_B(\varepsilon_c - \varepsilon_a) = \log \Omega_B(\varepsilon_c) - \beta\varepsilon_a - \frac{1}{2}\gamma\varepsilon_a^2 \ldots \, , \tag{5.52}$$

leading to corrections to the Boltzmann distribution,

$$p_a = \frac{1}{Z} \exp(-\beta\varepsilon_a - \frac{1}{2}\gamma\varepsilon_a^2 \ldots) \ . \tag{5.53}$$

An alternative path is to provide a more detailed model of the bath [Plastino 1994]. As before, we consider a system A that is weakly coupled to a heat bath B that has a finite size. The microstates of A and B are labelled a and b and have energies ε_a and ε_b respectively. The composite system $C = A + B$ can be assumed to be isolated and have a constant energy $\varepsilon_c = \varepsilon_a + \varepsilon_b$ (or more

precisely C has energy in some arbitrarily narrow interval about ε_c). To model the bath B we assume that the number of microstates of B with energy less than ε is $W(\varepsilon) = C\varepsilon^\alpha$, where the exponent α is some constant that depends on the size of the bath. Such a model is quite realistic; for example, $\alpha = N$ when the bath consists of N harmonic oscillators; and $\alpha = 3N/2$ when it is an ideal gas of N molecules.

Then the number of microstates of B in a narrow energy range $\delta\varepsilon$ is

$$\Omega_B(\varepsilon) = W(\varepsilon + \delta\varepsilon) - W(\varepsilon) = \alpha C \varepsilon^{\alpha-1}\delta\varepsilon \ , \tag{5.54}$$

and the probability that A is in a microstate a of energy ε_a is given by eq.(5.27),

$$p_a \propto \Omega_B(\varepsilon_c - \varepsilon_a) \propto (1 - \frac{\varepsilon_a}{\varepsilon_c})^{\alpha-1} \ , \tag{5.55}$$

so that

$$p_a = \frac{1}{Z}(1 - \frac{\varepsilon_a}{\varepsilon_c})^{\alpha-1} \quad \text{with} \quad Z = \textstyle\sum_a (1 - \frac{\varepsilon_a}{\varepsilon_c})^{\alpha-1} \, . \tag{5.56}$$

When the bath is sufficiently large $\varepsilon_a/\varepsilon_c \to 0$ and $\alpha \to \infty$ one recovers the Boltzmann distribution with appropriate corrections as in eq.(5.53). Indeed, using

$$\log(1+x) = x - \frac{1}{2}x^2 + \dots \tag{5.57}$$

we expand

$$(1 - \frac{\varepsilon_a}{\varepsilon_c})^{\alpha-1} = \exp(\alpha - 1)(-\frac{\varepsilon_a}{\varepsilon_c} - \frac{1}{2}\frac{\varepsilon_a^2}{\varepsilon_c^2} + \dots) \ , \tag{5.58}$$

to get eq.(5.53) with

$$\beta = \frac{\alpha - 1}{\varepsilon_c} \quad \text{and} \quad \gamma = \frac{\alpha - 1}{\varepsilon_c^2} \ . \tag{5.59}$$

We will not pursue the subject any further except to comment that distributions of this type have been proposed by C. Tsallis on the basis of a very different logic [Tsallis 1988].

Non-extensive thermodynamics

The idea proposed by Tsallis is to generalize the Boltzmann-Gibbs canonical formalism by adopting a different "non-extensive entropy",

$$T_\eta(p_1, \dots, p_n) = \frac{1 - \sum_i p_i^\eta}{\eta - 1} \ ,$$

that depends on a parameter η. Equivalent versions of such "entropies" have been proposed as alternative measures of information by several other authors; see, for example [Renyi 1961], [Aczel 1975], [Amari 1985].

One important feature is that the standard Shannon entropy is recovered in the limit $\eta \to 0$. Indeed, let $\eta = 1 + \delta$ and use

$$p_i^\delta = e^{\delta \log p_i} = 1 + \delta \log p_i + \dots \ . \tag{5.60}$$

As $\delta \to 0$ we get

$$\begin{aligned} T_{1+\delta} &= \frac{1}{\delta}(1 - \textstyle\sum_i p_i^{1+\delta}) \\ &= \frac{1}{\delta}[1 - \textstyle\sum_i p_i(1 + \delta \log p_i)] = -\textstyle\sum_i p_i \log p_i \ . \end{aligned} \tag{5.61}$$

The distribution that maximizes the Tsallis entropy subject to the usual normalization and energy constraints,

$$\textstyle\sum_i p_i = 1 \quad \text{and} \quad \sum_i \varepsilon_i p_i = E \, ,$$

is

$$p_i = \frac{1}{Z_\eta}[1 - \lambda\varepsilon_i]^{1/(\eta-1)} \, , \tag{5.62}$$

where Z_η is a normalization constant and the constant λ is a ratio of Lagrange multipliers. This distribution is precisely of the form (5.56) with $\lambda = 1/\varepsilon_c$ and $\eta = 1 + (\alpha - 1)^{-1}$.

Our conclusion is that Tsallis distributions make perfect sense within standard statistical mechanics. In order to justify them it is not necessary to introduce an alternative thermodynamics through new ad hoc entropies; it is merely necessary to recognize that sometimes a partial thermal equilibrium is reached with heat baths that are not extremely large. What changes is the relevant information on the basis of which we draw inferences and not the inference method. An added avantage is that the free and undetermined parameter η can, within the standard formalism advocated here, be calculated in terms of the size of the bath.

5.6 The Second Law of Thermodynamics

We saw that in 1865 Clausius summarized the two laws of thermodynamics into "The energy of the universe is constant. The entropy of the universe tends to a maximum." We can be a bit more explicit about the Second Law: In an adiabatic non-quasi-static process that starts and ends in equilibrium the total entropy increases; if the process is adiabatic and quasi-static the total entropy remains constant. The Second Law was formulated in a somewhat stronger form by Gibbs (1878): For irreversible processes not only does the entropy tend to increase, but it does increase to the maximum value allowed by the constraints imposed on the system.

We are now ready to prove the Second Law following [Jaynes 1965]. Jaynes' proof is mathematically very simple but it is also conceptually subtle. It may be useful to recall some of our previous results. The entropy mentioned in the Second Law is the thermodynamic entropy of Clausius S_C. It is defined only for equilibrium states,

$$S_C(B) - S_C(A) = \int_A^B \frac{dQ}{T} \ , \tag{5.63}$$

where the integral is along a reversible path of intermediate equilibrium states. But as we saw in section 5.4, in thermal equilibrium the *maximized* Gibbs entropy $S_G^{\rm can}$ — that is, the entropy computed from the canonical distribution — satisfies the same relation, eq.(5.43),

$$\delta S_G^{\rm can} = \frac{\delta Q}{T} \Rightarrow S_G^{\rm can}(B) - S_G^{\rm can}(A) = \int_A^B \frac{dQ}{T} \ , \tag{5.64}$$

which means that S_C and $S_G^{\rm can}$ differ only by an additive constant. Adjusting the constant so that $S_G^{\rm can}$ matches S_C for one equilibrium state they will match for all equilibrium states. Therefore, if at any time t the system is in thermal equilibrium and its relevant macrovariables agree with expected values, say $X(t)$, calculated using the canonical distribution then,

$$S_C(t) = S_G^{\rm can}(t) \ . \tag{5.65}$$

The system, which is assumed to be thermally insulated from its environment, is allowed (or forced) to evolve according to a certain Hamiltonian, $H(t)$. The evolution could, for example, be the free expansion of a gas into vacuum, or it could be given by the time-dependent Hamiltonian that describes some externally prescribed influence, say, a moving piston or an imposed field. Eventually a new equilibrium is reached at some later time t'. Such a process is adiabatic; no heat was exchanged with the environment. Under these circumstances the initial canonical distribution $f_{\rm can}(t)$, e.g. eq.(4.76) or (5.33), evolves according to Liouville's equation, eq.(5.6),

$$f_{\rm can}(t) \stackrel{H(t)}{\longrightarrow} f(t') \ , \tag{5.66}$$

and, according to eq.(5.23), the corresponding Gibbs entropy remains constant,

$$S_G^{\rm can}(t) = S_G(t') \ . \tag{5.67}$$

Since the Gibbs entropy remains constant it is sometimes argued that this contradicts the Second Law but note that the time-evolved $S_G(t')$ is not the thermodynamic entropy because $f(t')$ is not necessarily of the canonical form, eq.(4.76).

From the new distribution $f(t')$ we can, however, compute the new expected values $X(t')$ that apply to the state of equilibrium at t'. Of all distributions agreeing with the new values $X(t')$ the canonical distribution $f_{\rm can}(t')$ is that which has maximum Gibbs entropy, $S_G^{\rm can}(t')$. Therefore

$$S_G(t') \leq S_G^{\rm can}(t') \ . \tag{5.68}$$

But $S_G^{\rm can}(t')$ coincides with the thermodynamic entropy of the new equilibrium state,

$$S_G^{\rm can}(t') = S_C(t') \ . \tag{5.69}$$

Collecting all these results, eqs.(5.65)-(5.69), we conclude that the thermodynamic entropy has increased,

$$S_C(t) \leq S_C(t') \ . \tag{5.70}$$

This is the Second Law. The equality applies when the time evolution is quasi-static so that throughout the process the distribution is always canonical; in particular, $f(t') = f_{\rm can}(t')$. The argument above can be generalized considerably by allowing heat exchanges or by introducing uncertainties into the actual Hamiltonian dynamics.

To summarize, the chain of steps is

$$S_C(t) \underset{(1)}{=} S_G^{\rm can}(t) \underset{(2)}{=} S_G(t') \underset{(3)}{\leq} S_G^{\rm can}(t') \underset{(4)}{=} S_C(t') \ . \tag{5.71}$$

Steps (1) and (4) hinge on identifying the maximized Gibbs entropy with the thermodynamic entropy — which works provided we have correctly identified the relevant macrovariables for the particular problem at hand. Step (2) follows from the constancy of the Gibbs entropy under Hamiltonian evolution — this is the least controversial step. Of course, if we did not have complete knowledge about the exact Hamiltonian $H(t)$ acting on the system an inequality would have been introduced already at this point. The crucial inequality, however, is introduced in step (3) where *information is discarded*. The distribution $f(t')$ contains information about the macrovariables $X(t')$ at the final time t', but since the Hamiltonian is known, it also contains information about the whole previous history of $f(t)$ back to the initial time t and including the initial values $X(t)$. In contrast, a description in terms of the distribution $f_{\rm can}(t')$ contains information about the macrovariables $X(t')$ at time t' *and nothing else*. In a thermodynamic description all memory of the history of the system is lost.

The Second Law refers to thermodynamic entropies only. These entropies measure the amount of information available to someone with only macroscopic means to observe and manipulate the system. *The irreversibility implicit in the Second Law arises from this restriction to thermodynamic descriptions.*

Thus, the Second law is not a Law of Nature; it is not even a law within those imperfect and idealized models — such as classical mechanics — that attempt to describe Nature itself. The Second Law is a law but its connection to Nature is more indirect. It is a law within those models that attempt — not to describe Nature itself — but our limited and inadequate information about Nature.

It is important to emphasize what has just been proved: in an adiabatic process from one state of equilibrium to another the *thermodynamic* entropy increases. This is the Second Law. Many questions remain unanswered: We have assumed that the system tends towards and finally reaches an equilibrium; how do we know that this happens? What are the relaxation times, transport coefficients, etc.? There are all sorts of aspects of non-equilibrium irreversible processes that remain to be explained but this does not detract from what Jaynes' explanation did in fact accomplish, namely, it explained the Second Law, no more and, most emphatically, no less.

5.7 The thermodynamic limit

If the Second Law "has only statistical certainty" (Maxwell, 1871) and any violation "seems to be reduced to improbability" (Gibbs, 1878) how can thermodynamic predictions attain so much certainty? Part of the answer hinges on restricting the kind of questions we are willing to ask to those concerning the few macroscopic variables over which we have some control. Most other questions are not "interesting" and thus they are never asked. For example, suppose we are given a gas in equilibrium within a cubic box, and the question is where will we find particle #23. The answer is that we expect the particle to be at the center of the box but with a very large standard deviation — the particle can be anywhere in the box. Such an answer is not particularly impressive. On the other hand, if we ask for the energy of the gas at temperature T, or how it changes as the volume is changed by δV, then the answers are truly impressive.

Consider a system in thermal equilibrium in a macrostate described by a canonical distribution $f(z)$ assigned on the basis of constraints on the values of certain macrovariables X. For simplicity we will assume X is a single variable, the energy, $X = \langle\varepsilon\rangle = E$. The microstates z can be divided into typical and atypical microstates. The typical microstates are all contained within a "high probability" region $\mathcal{R}_\delta$ to be defined below that has total probability $1-\delta$, where δ is a small positive number, and within which $f(z)$ is greater than some lower bound. The "phase" volume of the typical region is

$$\mathrm{Vol}(\mathcal{R}_\delta) = \int_{\mathcal{R}_\delta} dz = W_\delta \ . \tag{5.72}$$

Our goal is to establish that the thermodynamic entropy and the volume of the region $\mathcal{R}_\delta$ are related through Boltzmann's equation,

$$S_C \approx k \log W_\delta \ . \tag{5.73}$$

The surprising feature is that the result is essentially independent of δ. The following theorems which are adaptations of the Asymptotic Equipartition Property (section 4.7) state this result in a mathematically precise way.

The Asymptotic Equipartition Theorem: Let $f(z)$ be the canonical distribution and $kS = S_G = S_C$ the corresponding entropy,

$$f(z) = \frac{e^{-\beta\varepsilon(z)}}{Z} \quad \text{and} \quad S = \beta E + \log Z \ . \tag{5.74}$$

Then as $N \to \infty$,

$$-\frac{1}{N}\log f(z) \longrightarrow \frac{S}{N} \quad \text{in probability,} \tag{5.75}$$

provided that the system is such that the energy fluctuations $\Delta\varepsilon$ increase slower than N, that is, $\lim_{N\to\infty} \Delta\varepsilon/N = 0$. ($\Delta$ denotes the standard deviation.)

The theorem roughly means that *the probabilities of the accessible microstates are essentially equal.* The microstates z for which $(-\log f(z))/N$ differs substantially from S/N have either too low probability and are deemed "inaccessible"

or they might individually have a high probability but are too few to contribute significantly.

Remark: The word 'essentially' is tricky because $f(z)$ may differ from e^{-S} by a huge *multiplicative* factor — perhaps several billion — but $\log f(z)$ still differs from $-S$ by an unimportant amount that grows less rapidly than N.

Remark: The left hand side of (5.75) is a quantity associated to a microstate z while the right side contains the entropy S. This may mislead us into thinking that the entropy S can be associated to microstate rather than macrostates. The crucial point is that the limit in (5.75) is valid 'in probability' only.

Proof: Apply the Tchebyshev inequality, eq.(2.93),

$$P\left(|x - \langle x\rangle| \geq \delta\right) \leq \left(\frac{\Delta x}{\delta}\right)^2 \ , \tag{5.76}$$

to the variable

$$x = \frac{-1}{N} \log f(z) \ . \tag{5.77}$$

Its expected value is the entropy per particle,

$$\begin{aligned} \langle x\rangle &= \frac{-1}{N} \langle \log f\rangle \\ &= \frac{S}{N} = \frac{1}{N}\left(\beta E + \log Z\right) \ . \end{aligned} \tag{5.78}$$

To calculate the variance,

$$(\Delta x)^2 = \frac{1}{N^2}\left[\left\langle (\log f)^2\right\rangle - \langle \log f\rangle^2\right] \ , \tag{5.79}$$

use

$$\begin{aligned} \left\langle (\log f)^2\right\rangle &= \left\langle (\beta\varepsilon + \log Z)^2\right\rangle \\ &= \beta^2 \left\langle \varepsilon^2\right\rangle + 2\beta \langle \varepsilon\rangle \log Z + (\log Z)^2 \ , \end{aligned} \tag{5.80}$$

so that

$$(\Delta x)^2 = \frac{\beta^2}{N^2}\left(\left\langle \varepsilon^2\right\rangle - \langle \varepsilon\rangle^2\right) = \left(\frac{\beta \Delta\varepsilon}{N}\right)^2 \ . \tag{5.81}$$

Collecting these results gives

$$\text{Prob}\left[\left|-\frac{1}{N}\log f(z) - \frac{S}{N}\right| \geq \delta\right] \leq \left(\frac{\beta}{\delta}\right)^2 \left(\frac{\Delta\varepsilon}{N}\right)^2 \ . \tag{5.82}$$

For systems such that the relative energy fluctuations $\Delta\varepsilon/N$ tend to 0 as $N \to \infty$ the limit on the right is zero,

$$\lim_{N\to\infty} \text{Prob}\left[\left|-\frac{1}{N}\log f(z) - \frac{S}{N}\right| \geq \delta\right] = 0 \ , \tag{5.83}$$

which concludes the proof.

Remark: Note that the theorem applies only to those systems with interparticle interactions such that the energy fluctuations $\Delta\varepsilon$ are sufficiently well behaved. For example, it is not uncommon that $\Delta\varepsilon/E \propto N^{-1/2}$ and that the energy is an extensive quantity, $E/N \to$ const. Then

$$\frac{\Delta\varepsilon}{N} = \frac{\Delta\varepsilon}{E}\frac{E}{N} \propto \frac{1}{N^{1/2}} \to 0 \ . \tag{5.84}$$

Typically this requires that as N and V tend to infinity with N/V constant, the spatial correlations fall sufficiently fast that distant particles are uncorrelated. Under these circumstances both energy and entropy are extensive quantities.

The following theorem elaborates on these ideas further. To be precise let us redefine the typical region $\mathcal{R}_\delta$ as the set of microstates with probability $f(z)$ such that

$$e^{-S-N\delta} \le f(z) \le e^{-S+N\delta} \ , \tag{5.85}$$

or, using eq.(5.74),

$$\frac{1}{Z}e^{-\beta E-N\delta} \le f(z) \le \frac{1}{Z}e^{-\beta E+N\delta} \ . \tag{5.86}$$

This last expression shows that typical microstates are those for which the energy per particle $\varepsilon(z)/N$ lies within a narrow interval $2\delta kT$ about its expected value E/N.

Remark: Even though some states z (namely those with energy $\varepsilon(z) < E$) can individually be more probable than the typical states it turns out (see below) that they are too few and their volume is negligible compared to W_δ.

Theorem of typical microstates: For N sufficiently large

(1) $\text{Prob}[\mathcal{R}_\delta] > 1-\delta$

(2) $\text{Vol}(\mathcal{R}_\delta) = W_\delta \le e^{S+N\delta}$.

(3) $W_\delta \ge (1-\delta)e^{S-N\delta}$.

(4) $\lim_{N\to\infty}(\log W_\delta - S)/N = 0$.

In words:

> *The typical region has probability close to one; typical microstates are almost equally probable; the phase volume they occupy is about e^S, that is, $S = k\log W$.*

For large N the entropy is a measure of the logarithm of the phase volume of typical states**,**

$$S = \log W_\delta \pm N\delta \ , \tag{5.87}$$

where $\log W_\delta = N \times O(1)$ while $\delta \ll 1$ and it does not much matter what we precisely we mean by typical (i.e., whether we choose for $\delta = 10^{-6}$ or 10^{-12}). Incidentally, note that it is the (maximized) Gibbs entropy that satisfies the

Boltzmann formula $S_G = S_C = k \log W$ (where the irrelevant subscript δ has been dropped).

Proof: Eq.(5.83) states that for fixed δ, for any given η there is an N_η such that for all $N > N_\eta$, we have

$$\text{Prob}\left[\left|-\frac{1}{N}\log f(z) - \frac{S}{N}\right| \leq \delta\right] \geq 1 - \eta\,. \tag{5.88}$$

Thus, the probability that a microstate z drawn from the distribution $f(z)$ is δ-typical tends to one, and therefore so must Prob$[\mathcal{R}_\delta]$. Setting $\eta = \delta$ yields part **(1)**. This also shows that the total probability of the set of states with

$$\varepsilon(z) < E \quad \text{or} \quad f(z) > e^{-S+N\delta} = \frac{1}{Z}e^{-\beta E + N\delta} \tag{5.89}$$

is negligible — states that individually are more probable than typical occupy a negligible volume. To prove **(2)** write

$$\begin{aligned} 1 \geq \text{Prob}[\mathcal{R}_\delta] &= \textstyle\int_{\mathcal{R}_\delta} dz\, f(z) \\ &\geq e^{-S-N\delta} \textstyle\int_{\mathcal{R}_\delta} dz \; = e^{-S-N\delta} W_\delta\,. \end{aligned} \tag{5.90}$$

Similarly, to prove **(3)** use **(1)**,

$$\begin{aligned} 1 - \delta < \text{Prob}[\mathcal{R}_\delta] &= \textstyle\int_{\mathcal{R}_\delta} dz\, f(z) \\ &\leq e^{-S+N\delta} \textstyle\int_{\mathcal{R}_\delta} dz = e^{-S+N\delta} W_\delta\,, \end{aligned} \tag{5.91}$$

Finally, from (2) and (3),

$$(1-\delta)e^{S-N\delta} \leq W_\delta \leq e^{S+N\delta}\ , \tag{5.92}$$

which is the same as

$$\frac{S}{N} - \delta + \frac{\log(1-\delta)}{N} \leq \frac{\log W_\delta}{N} \leq \frac{S}{N} + \delta\ , \tag{5.93}$$

and proves **(4).**

Remark: The theorems above can be generalized to situations involving several macrovariables X^k in addition to the energy. In this case, the expected value of $\log f(z)$ is

$$\langle -\log f \rangle = S = \lambda_k \left\langle X^k \right\rangle + \log Z\ , \tag{5.94}$$

and its variance is

$$(\Delta \log f)^2 = \lambda_k \lambda_m \left(\left\langle X^k X^m \right\rangle - \left\langle X^k \right\rangle \left\langle X^m \right\rangle \right)\ . \tag{5.95}$$

5.8 Interpretation of the Second Law: Reproducibility

First a summary of the previous sections: We saw that within the typical region $\mathcal{R}(t)$ fluctuations of the $X(t)$ are negligible — all microstates are characterized by the same values of X, eq.(5.86) — and that given $X(t)$ the typical region has probability one — it includes essentially all possible initial states compatible with the values $X(t)$. Having been prepared in equilibrium at time t the system is then subjected to an adiabatic process and it eventually attains a new equilibrium at time t'. The Hamiltonian evolution deforms the initial region $\mathcal{R}(t)$ into a new region $\mathcal{R}(t')$ with exactly the same volume $W(t) = W(t')$; the macrovariables evolve from their initial values $X(t)$ to new values $X(t')$. Now suppose that for the new equilibrium we adopt a thermodynamic description: the preparation history is forgotten, and all we know are the new values $X(t')$. The new typical region $\mathcal{R}'(t')$ has a volume $W'(t') > W(t)$ and it includes all microstates compatible with the information $X(t')$.

The volume $W(t) = e^{S_C(t)/k}$ of the typical region can be interpreted in two ways. On one hand it is a measure of our ignorance as to the true microstate when all we know are the macrovariables $X(t)$. On the other hand, the volume $W(t)$ is also a measure of the extent that we can experimentally control the actual microstate of the system when the $X(t)$ are the only parameters we can manipulate.

After these preliminaries we come to the crux of the argument: With the limited experimental means at our disposal we can guarantee that the initial microstate will be somewhere within $W(t)$ and therefore that in due course of time it will evolve to be within $W(t')$. (See fig.6-1.) In order for the process $X(t) \to X(t')$ to be experimentally reproducible it must be that all the microstates in $W(t')$ will also evolve to be within $W'(t')$ which means that $W(t) = W(t') \leq W'(t')$. Conversely, if it were true that $W(t) > W'(t')$ we would sometimes observe that an initial microstate within $W(t)$ would evolve into a final microstate lying outside $W'(t')$ that is, sometimes we would observe that $X(t)$ does not evolve to $X(t')$. Thus, when $W(t) > W'(t')$ the experiment is definitely not reproducible.

A new element has been introduced into the discussion of the Second Law: *reproducibility* [Jaynes 1965]. Thus, we can express the Second Law in the somewhat tautological form:

> *In a reproducible adiabatic process from one state of equilibrium to another the thermodynamic entropy cannot decrease.*

We can address this question from a different angle: How do we know that the chosen constraints X are the relevant macrovariables that provide an *adequate* thermodynamic description? In fact, what do we mean by an *adequate* description? Let us rephrase these questions differently: Could there exist additional unknown physical constraints Y that significantly restrict the microstates

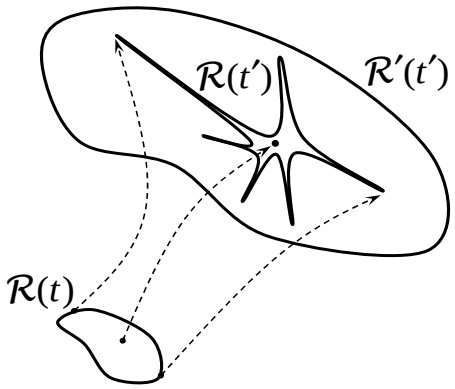


Figure 5.1: A system prepared by specifying $X(t)$ at the initial time t lies somewhere within $\mathcal{R}(t)$. It evolves to the region $\mathcal{R}(t')$ characterized by values $X(t')$. The experiment is reproducible because all states within the larger region $\mathcal{R}'(t')$ are characterized by the same $X(t')$.

compatible with the initial macrostate and which therefore provide an even better description? The answer is that such variables can, of course, exist but that including them in the description need not necessarily lead to improved predictions. If the process $X(t) \to X(t')$ is reproducible when no particular care has been taken to control the values of Y we can expect that to the extent that we are only interested in the Xs; keeping track of the Ys will not yield a better description. *Reproducibility is the pragmatic criterion whereby we can decide whether a particular thermodynamic description is adequate or not.*

5.9 Remarks on irreversibility

A considerable source of confusion on the question of reversibility is that the same word 'reversible' is used with several different meanings [Uffink 2001]:
(a) *Mechanical or microscopic reversibility* refers to the possibility of reversing the velocities of every particle. Such reversals would allow a completely isolated

system not just to retrace its steps from the final macrostate to the initial macrostate but it would also allow it to retrace its detailed microstate trajectory as well.

(b) *Carnot or macroscopic reversibility* refers to the possibility of retracing the history of macrostates of a system in the opposite direction. The required amount of control over the system can be achieved by forcing the system along a prescribed path of intermediate macroscopic equilibrium states that are infinitesimally close to each other. Such a reversible process is appropriately called *quasi-static*. There is no implication that the trajectories of the individual particles will be retraced.

(c) *Thermodynamic reversibility* refers to the possibility of starting from a final macrostate and completely recovering the initial macrostate without any other external changes. There is no need to retrace the intermediate macrostates in reverse order. In fact, rather than 'reversibility' it may be more descriptive to refer to '*recoverability*'. Typically a state is irrecoverable when there is friction, decay, or corruption of some kind.

Notice that when one talks about the "irreversibility" of the Second Law and about the "reversibility" of mechanics there is no inconsistency or contradiction: the word 'reversibility' is being used with two entirely different meanings.

Classical thermodynamics assumes that isolated systems approach and eventually attain a state of equilibrium. By definition the state of equilibrium is such that, once attained, it will not spontaneously change in the future. On the other hand, it is understood that changes might have happened in the past. Classical thermodynamics introduces a time asymmetry: it treats the past and the future differently.

The situation with statistical mechanics, however, is different. Once equilibrium has been attained fluctuations are possible. In fact, if we are willing to wait long enough we can be certain that large fluctuations will necessarily happen in the future just as they might have happened in the past. The situation is quite symmetric. The interesting asymmetry arises when we realize that for an improbable state — a large fluctuation — to happen spontaneously in the future we may have to wait an extremely long time while we are perfectly willing to entertain the possibility that a similarly improbable state was observed in the very recent past. This might seem paradoxical because the formalism of statistical mechanics does not introduce any time asymmetry. The solution to the puzzle is that the improbable state in the recent past did not in all probability happen spontaneously but was brought about by some external intervention. The system might, for example, have been deliberately prepared in some unusual state by applying appropriate constraints which were subsequently removed — we do this all the time.

Thus, the time asymmetry is not introduced by the laws of mechanics; it is introduced through the information we accept as relevant for this kind of situation: we know that deliberate manipulations have happened in the past and that they will not happen in the future.

5.10 Entropies, descriptions and the Gibbs paradox

Under the generic title of "Gibbs Paradox" one usually considers a number of related questions in both phenomenological thermodynamics and in statistical mechanics: (1) The entropy change when two distinct gases are mixed happens to be independent of the nature of the gases. Is this in conflict with the idea that in the limit as the two gases become identical the entropy change should vanish? (2) Should the thermodynamic entropy of Clausius be an extensive quantity or not? (3) Should two microstates that differ only in the exchange of identical particles be counted as two or just one microstate?

The conventional wisdom asserts that the resolution of the paradox rests on quantum mechanics but this analysis is unsatisfactory; at best it is incomplete. While it is true that the exchange of identical quantum particles does not lead to a new microstate this approach ignores the case of classical, and even non-identical particles. For example, nanoparticles in a colloidal suspension or macromolecules in solution are both classical and non-identical. Several authors (e.g., [Grad 1961, Jaynes 1992]) have recognized that quantum theory has no bearing on the matter; indeed, as remarked in section 3.5, this was already clear to Gibbs.

Our purpose here is to discuss the Gibbs paradox from the point of view of information theory. The discussion follows [Tseng Caticha 2001]. Our conclusion will be that the paradox is resolved once it is realized that there is no such thing as *the* entropy of a system, that there are *many* entropies. The choice of entropy is a choice between a description that treats particles as being distinguishable and a description that treats them as indistinguishable; which of these alternatives is more convenient depends on the resolution of the particular experiment being performed.

The "grouping" property of entropy, eq.(4.3),

$$S[p] = S_G[P] + \textstyle\sum_g P_g S_g[p_{\cdot|g}]$$

plays an important role in our discussion. It establishes a relation between several different descriptions and refers to three different entropies. One can describe the system with high resolution as being in a microstate i (with probability p_i), or alternatively, with lower resolution as being in one of the groups g (with probability P_g). Since the description in terms of the groups g is less detailed we might refer to them as 'mesostates'. A thermodynamic description, on the other hand, corresponds to an even lower resolution that merely specifies the equilibrium macrostate. For simplicity, we will define the macrostate with a single variable, the energy. Including additional variables is easy and does not modify the gist of the argument.

The standard connection between the thermodynamic description in terms of macrostates and the description in terms of microstates is established in section 5.4. If the energy of microstate a is ε_a, to the macrostate of energy $E = \langle\varepsilon\rangle$ we

associate that canonical distribution (5.33)

$$p_a = \frac{e^{-\beta\varepsilon_a}}{Z_H}\,, \tag{5.96}$$

where the partition function Z_H and the Lagrange multiplier β are determined from eqs.(5.34),

$$Z_H = \sum_i e^{-\beta\varepsilon_i} \quad \text{and} \quad \frac{\partial \log Z_H}{\partial \beta} = -E\ . \tag{5.97}$$

The corresponding entropy, eq.(5.35) is (setting $k = 1$)

$$S_H = \beta E + \log Z_H\,, \tag{5.98}$$

measures the amount of information required to specify the microstate when all we know is the value E.

Identical particles

Before we compute and interpret the probability distribution over mesostates and its corresponding entropy we must be more specific about which mesostates we are talking about. Consider a system of N classical particles that are exactly identical. The interesting question is whether these identical particles are also "distinguishable". By this we mean the following: we look at two particles now and we label them. We look at the particles later. Somebody might have switched them. Can we tell which particle is which? The answer is: it depends. Whether we can distinguish identical particles or not depends on whether we were able and willing to follow their trajectories.

A slightly different version of the same question concerns an N-particle system in a certain state. Some particles are permuted. Does this give us a different state? As discussed earlier the answer to this question requires a careful specification of what we mean by a state.

Since by a *microstate* we mean a point in the N-particle phase space, then a permutation does indeed lead to a new microstate. On the other hand, our concern with particle exchanges suggests that it is useful to introduce the notion of a *mesostate* defined as the group of those $N!$ microstates that are obtained by particle permutations. With this definition it is clear that a permutation of the identical particles does not lead to a new mesostate.

Now we can return to discussing the connection between the thermodynamic macrostate description and the description in terms of mesostates using, as before, the method of Maximum Entropy. Since the particles are (sufficiently) identical, all those $N!$ microstates i within the same mesostate g have the same energy, which we will denote by E_g (*i.e.*, $E_i = E_g$ for all $i \in g$). To the macrostate of energy $\bar{E} = \langle E \rangle$ we associate the canonical distribution,

$$P_g = \frac{e^{-\beta E_g}}{Z_L}\,, \tag{5.99}$$

where

$$Z_L = \sum_g e^{-\beta E_g} \quad \text{and} \quad \frac{\partial \log Z_L}{\partial \beta} = -\bar{E}\,. \tag{5.100}$$

The corresponding entropy, eq.(5.35) is (setting $k = 1$)

$$S_L = \beta \bar{E} + \log Z_L\,, \tag{5.101}$$

measures the amount of information required to specify the mesostate when all we know is $\bar{E}$.

Two different entropies S_H and S_L have been assigned to the same macrostate $\bar{E}$; they measure the different amounts of additional information required to specify the state of the system to a high resolution (the microstate) or to a low resolution (the mesostate).

The relation between Z_H and Z_L is obtained from

$$Z_H = \sum_i e^{-\beta E_i} = N! \sum_g e^{-\beta E_g} = N! Z_L \quad \text{or} \quad Z_L = \frac{Z_H}{N!}\,. \tag{5.102}$$

The relation between S_H and S_L is obtained from the "grouping" property, eq.(4.3), with $S = S_H$ and $S_G = S_L$, and $p_{i|g} = 1/N!$. The result is

$$S_L = S_H - \log N!\,. \tag{5.103}$$

Incidentally, note that

$$S_H = -\textstyle\sum_a p_a \log p_a = -\sum_g P_g \log P_g/N!\ . \tag{5.104}$$

Equations (5.102) and (5.103) both exhibit the Gibbs $N!$ "corrections." Our analysis shows (1) that the justification of the $N!$ factor is not to be found in quantum mechanics, and (2) that the $N!$ does not correct anything. The $N!$ is not a fudge factor that fixes a wrong (possibly nonextensive) entropy S_H into a correct (possibly extensive) entropy S_L. Both entropies S_H and S_L are correct. They differ because they measure different things: one measures the information to specify the microstate, the other measures the information to specify the mesostate.

An important goal of statistical mechanics is to provide a justification, an explanation of thermodynamics. Thus, we still need to ask which of the two statistical entropies, S_H or S_L, should be identified with the thermodynamic entropy of Clausius S_T. Inspection of eqs.(5.102) and (5.103) shows that, as long as one is not concerned with experiments that involve changes in the number of particles, the same thermodynamics will follow whether we set $S_H = S_T$ or $S_L = S_T$.

But, of course, experiments involving changes in N are very important (for example, in the equilibrium between different phases, or in chemical reactions). Since in the usual thermodynamic experiments we only care that some number of particles has been exchanged, and we do not care which were the actual particles exchanged, we expect that the correct identification is $S_L = S_T$. Indeed,

the quantity that regulates the equilibrium under exchanges of particles is the chemical potential defined by

$$\mu = -kT \left(\frac{\partial S_T}{\partial N} \right)_{E,V,\ldots} \tag{5.105}$$

The two identifications $S_H = S_T$ or $S_L = S_T$, lead to two different chemical potentials, related by

$$\mu_L = \mu_H - NkT\,. \tag{5.106}$$

It is easy to verify that, under the usual circumstances where surface effects can be neglected relative to the bulk, μ_L has the correct functional dependence on N: it is intensive and can be identified with the thermodynamic μ. On the other hand, μ_H is not an intensive quantity and cannot therefore be identified with μ.

Non-identical particles

We saw that classical identical particles can be treated, depending on the resolution of the experiment, as being distinguishable or indistinguishable. Here we go further and point out that even non-identical particles can be treated as indistinguishable. Our goal is to state explicitly in precisely what sense it is up to the observer to decide whether particles are distinguishable or not.

We defined a mesostate as a subset of $N!$ microstates that are obtained as permutations of each other. With this definition it is clear that a permutation of particles does not lead to a new mesostate even if the exchanged particles are not identical. This is an important extension because, unlike quantum particles, classical particles cannot be expected to be exactly identical down to every minute detail. In fact in many cases the particles can be grossly different – examples might be colloidal suspensions or solutions of organic macromolecules. A high resolution device, for example an electron microscope, would reveal that no two colloidal particles or two macromolecules are exactly alike. And yet, for the purpose of modelling most of our macroscopic observations it is not necessary to take account of the myriad ways in which two particles can differ.

Consider a system of N particles. We can perform rather crude macroscopic experiments the results of which can be summarized with a simple phenomenological thermodynamics where N is one of the relevant variables that define the macrostate. Our goal is to construct a statistical foundation that will explain this macroscopic model, reduce it, so to speak, to "first principles." The particles might ultimately be non-identical, but the crude phenomenology is not sensitive to their differences and can be explained by postulating mesostates g and microstates i with energies $E_i \approx E_g$, for all $i \in g$, as if the particles were identical. As in the previous section this statistical model gives

$$Z_L = \frac{Z_H}{N!} \quad \text{with} \quad Z_H = \sum_i e^{-\beta E_i}\,, \tag{5.107}$$

and the connection to the thermodynamics is established by postulating

$$S_T = S_L = S_H - \log N! \,. \tag{5.108}$$

Next we consider what happens when more sophisticated experiments are performed. The examples traditionally offered in discussions of this sort refer to the new experiments that could be made possible by the discovery of membranes that are permeable to some of the N particles but not to the others. Other, perhaps historically more realistic examples, are afforded by the availability of new experimental data, for example, more precise measurements of a heat capacity as a function of temperature, or perhaps measurements in a range of temperatures that had previously been inaccessible.

Suppose the new phenomenology can be modelled by postulating the existence of two kinds of particles. (Experiments that are even more sophisticated might allow us to detect three or more kinds, perhaps even a continuum of different particles.) What we previously thought were N identical particles we will now think as being N_a particles of type a and N_b particles of type b. The new description is in terms of macrostates defined by N_a and N_b as the relevant variables.

To construct a statistical explanation of the new phenomenology from 'first principles' we need to revise our notion of mesostate. Each new mesostate will be a group of microstates which will include all those microstates obtained by permuting the a particles among themselves, and by permuting the b particles among themselves, but will not include those microstates obtained by permuting a particles with b particles. The new mesostates, which we will label $\hat{g}$ and to which we will assign energy $\varepsilon_{\hat{g}}$, will be composed of $N_a!N_b!$ microstates $\hat{\imath}$, each with a well defined energy $E_{\hat{\imath}} = E_{\hat{g}}$, for all $\hat{\imath} \in \hat{g}$. The new statistical model gives

$$\hat{Z}_L = \frac{\hat{Z}_H}{N_a!N_b!} \quad \text{with} \quad \hat{Z}_H = \sum_{\hat{\imath}} e^{-\beta E_{\hat{\imath}}} \,, \tag{5.109}$$

and the connection to the new phenomenology is established by postulating

$$\hat{S}_T = \hat{S}_L = \hat{S}_H - \log N_a!N_b! \,. \tag{5.110}$$

In discussions of this topic it is not unusual to find comments to the effect that in the limit as particles a and b become identical one expects that the entropy of the system with two kinds of particles tends to the entropy of a system with just one kind of particle. The fact that this expectation is not met is one manifestation of the Gibbs paradox.

From the information theory point of view the paradox does not arise because there is no such thing as *the entropy of the system*, there are several entropies. It is true that as $a \to b$ we will have $\hat{Z}_H \to Z_H$, and accordingly $\hat{S}_H \to S_H$, but there is no reason to expect a similar relation between $\hat{S}_L$ and S_L because these two entropies refer to mesostates $\hat{g}$ and g that remain different even as a and b became identical. In this limit the mesostates $\hat{g}$, which are useful for

descriptions that treat particles a and b as indistinguishable among themselves but distinguishable from each other, lose their usefulness.

Conclusion

The Gibbs paradox in its various forms arises from the widespread misconception that entropy is a real physical quantity and that one is justified in talking about *the entropy* of the system. The thermodynamic entropy is not a property of the system. Entropy is a property of our description of the system, it is a property of the macrostate. More explicitly, it is a function of the macroscopic variables used to define the macrostate. To different macrostates reflecting different choices of variables there correspond different entropies for the very same system.

But this is not the complete story: entropy is not just a function of the macrostate. Entropies reflect a relation between two descriptions of the same system: in addition to the macrostate, we must also specify the set of microstates, or the set of mesostates, as the case might be. Then, having specified the macrostate, an entropy can be interpreted as the amount of additional information required to specify the microstate or mesostate. We have found the 'grouping' property very valuable precisely because it emphasizes the dependence of entropy on the choice of micro or mesostates.

Chapter 6

Entropy III: Updating Probabilities

Inductive inference is a framework for reasoning with incomplete information, for coping with uncertainty. The framework must include a means to represent a state of partial knowledge — this is handled through the introduction of probabilities — and it must allow us to change from one state of partial knowledge to another when new information becomes available. Indeed any inductive method that recognizes that a situation of incomplete information is in some way unfortunate — by which we mean that it constitutes a problem in need of a solution — would be severely deficient if it failed to address the question of how to proceed in the fortunate circumstance that some additional information has become available. *The theory of probability, if it is to useful at all, cannot be separate from a theory for updating probabilities.*

The challenge is to develop updating methods that are both systematic, objective and practical. In Chapter 2 we saw that Bayes' rule is the natural way to update when the information is available in the form of data and of a likelihood function. We also saw that Bayes' rule could not be derived just from the requirements of consistency implicit in the sum and product rules of probability theory. An additional principle of parsimony — the Principle of Minimal Updating (PMU) — was necessary: *Whatever was learned in the past is valuable and should not be disregarded; beliefs should be revised but only to the extent required by the new data.* A few interesting questions were just barely hinted at: How do we update when the information is not in the form of data? If the information is not data, what else could it possibly be? Indeed what, after all, is information?

Then in Chapter 4 we saw that the method of maximum entropy, MaxEnt, allowed one to deal with information in the form of constraints on the allowed probability distributions. So here we have a partial answer to one of our questions: in addition to data information can also take the form of constraints. However, MaxEnt was not designed as a method for updating; it is a method

for assigning probabilities on the basis of the constraint information, but it does not allow us to take into account the information contained in generic prior distributions.

Thus, Bayes' rule allows for the information contained in arbitrary priors and in data, but not in arbitrary constraints,[1] while on the other hand, MaxEnt can handle arbitrary constraints but not arbitrary priors. In this chapter we bring those two methods together: by generalizing the PMU we show how the MaxEnt method can be extended beyond its original scope, as a rule to assign probabilities, to a full-fledged method for inductive inference, that is, a method for updating from arbitrary priors given information in the form of arbitrary constraints. It should not be too surprising that the extended Maximum Entropy method — which we will henceforth abbreviate as ME, and also refer to as 'entropic inference' or 'entropic updating' — includes both MaxEnt and Bayes' rule as special cases.

Historically the ME method is a direct descendant of MaxEnt. As we saw in chapter 4 in the MaxEnt framework entropy is interpreted through the Shannon axioms as a measure of the amount of information that is missing in a probability distribution. We discussed some limitations of this approach. The Shannon axioms refer to probabilities of discrete variables; for continuous variables the entropy is not defined. But a more serious objection was raised: even if we grant that the Shannon axioms do lead to a reasonable expression for the entropy, to what extent do we believe the axioms themselves? Shannon's third axiom, the grouping property, is indeed sort of reasonable, but is it necessary? Is entropy the only consistent measure of uncertainty or of information? What is wrong with, say, the standard deviation? Indeed, there exist examples in which the Shannon entropy does not seem to reflect one's intuitive notion of information [Uffink 1995]. Other entropies, justified by a different choice of axioms, can be introduced (for example, [Renyi 1961] and [Tsallis 1988]); which one should one adopt?

From our point of view the real limitation is that neither Shannon nor Jaynes were concerned with the problem of updating. Shannon was analyzing the capacity of communication channels and characterizing the potential diversity of messages generated by information sources (section 4.7). His entropy makes no reference to prior distributions. On the other hand, as we already mentioned, Jaynes conceived MaxEnt as a method to assign probabilities on the basis of constraint information and a fixed underlying measure, not an arbitrary prior. He never meant to update from one probability distribution to another.

Considerations such as these motivated several attempts to develop ME directly as a method for updating probabilities without invoking questionable measures of uncertainty. Prominent among them are [Shore and Johnson 1980, Skilling 1988-90, Csiszar 1991]. The important contribution by Shore and Johnson was the realization that one could axiomatize the updating method itself rather than the information measure. Their axioms are justified on the basis of

[1] Bayes' rule can handle constraints when they are expressed in the form of data that can be plugged into a likelihood function but not all constraints are of this kind.

a fundamental principle of consistency — if a problem can be solved in more than one way the results should agree — but the axioms themselves and other assumptions they make have raised some objections [Karbelkar 1986, Uffink 1995]). Despite such criticism Shore and Johnson's pioneering papers have had an enormous influence: they identified the correct goal to be achieved.

The main goal of this chapter is to design a framework for updating — the method of entropic inference. The concept of relative entropy is introduced as a tool for reasoning — it is designed to perform a certain function defined through certain *design criteria* or *specifications*. There is no implication that the method is "true", or that it succeeds because it achieves some special contact with reality. Instead the claim is that the method succeeds in the sense that it works as designed — and that this is satisfactory because it leads to empirically adequate models. The presentation below is based on [Caticha 2003, Caticha Giffin 2006, Caticha 2007].

As we argued earlier when developing the theory of degrees of belief, our general approach differs from the way in which many physical theories have been developed in the past. The more traditional approach consists of first setting up the mathematical formalism and then seeking an acceptable interpretation. The drawback of this procedure is that questions can always be raised about the uniqueness of the proposed interpretation, and about the criteria that makes it acceptable or not.

In contrast, here we proceed in the opposite order: we first decide what we are talking about, what goal we want to achieve, and only then we construct a suitable mathematical formalism designed with that specific goal in mind. The advantage is that issues of meaning and interpretation are resolved from the start. The preeminent example of this approach is Cox's algebra of probable inference (see chapter 2) which clarified the meaning and use of the notion of probability: after Cox it is no longer possible to raise doubts about the legitimacy of adopting the degree of rational belief interpretation. Similarly, the concept of entropy is introduced as a tool for reasoning without recourse to notions of heat, multiplicity of states, disorder, uncertainty, or even in terms of an amount of information. In this approach *entropy needs no interpretation*. We do not need to know what 'entropy' means; we only need to know how to use it. Incidentally, this may help explain why previous searches failed to find a uniquely correct and unobjectionably precise meaning for the concept of entropy — there is none to be found.

Since the PMU is the driving force behind both Bayesian and ME updating it is worthwhile to investigate the precise relation between the two. We show that Bayes' rule can be derived as a special case of the ME method. This important result was first obtained by Williams (see [Williams 80][Diaconis 82]) long before the use of relative entropy as a tool for inference had been properly justified — that is, without appealing to questionable measures of information. Accordingly Williams' achievement did not receive the widespread appreciation it deserved. The virtue of the derivation presented here [Caticha Giffin 2006], which hinges on translating information in the form of data into a constraint that can be processed using ME, is that it is particularly clear. It throws light

on Bayes' rule and demonstrates its complete compatibility with ME updating. Thus, within the ME framework maximum entropy and Bayesian methods are unified into a single consistent theory of inference. One advantage of this insight is that it allows a number of generalizations of Bayes' rule (see section 2.9.2). Another is that it has implications for physics: it provides an important missing piece for the big puzzle of quantum mechanics (see chapter 10).

There is another function that the ME method must perform in order to fully qualify as a method of inductive inference: once we have decided that the distribution of maximum entropy is to be preferred over all others the following question arises immediately: the maximum of the entropy functional is never infinitely sharp, are we really confident that distributions that lie very close to the maximum are totally ruled out? We must find a quantitative way to assess the extent to which distributions with lower entropy are ruled out. This matter will be later addressed in chapter 8.

6.1 What is information?

The term 'information' is used with a wide variety of different meanings [Cover Thomas 1991, Jaynes 2003, Caticha 2007, Golan 2008, Floridi 2011, Adriaans 2012]. There is the Shannon notion of information, a technical term that, as we have seen, is meant to measure an amount of information and is quite divorced from semantics. There is also an algorithmic notion of information, which captures the notion of complexity and originates in the work of Solomonov, Kolmogorov and Chaitin, and has been developed as an alternative approach to induction, learning, artificial intelligence, and as a general theory of knowledge — it has been suggested that data compression is one of the principles that governs human cognition. Despite the obvious relevance to our subject, the algorithmic approach will not be pursued here. Instead we develop an epistemic notion of information that is somewhat closer to the everyday colloquial use of the term — roughly, information is what we seek when we ask a question.

It is not unusual to hear that systems "carry" or "contain" information and that "information is physical". This mode of expression can perhaps be traced to the origins of information theory in Shannon's theory of communication. We say that we have received information when among the vast variety of messages that could conceivably have been generated by a distant source, we discover which particular message was actually sent. It is thus that the message "carries" information. The analogy with physics is immediate: the set of all possible states of a physical system can be likened to the set of all possible messages, and the actual state of the system corresponds to the message that was actually sent. Thus, the system "conveys" a message: the system "carries" information about its own state. Sometimes the message might be difficult to read, but it is there nonetheless.

This language — information is physical — useful as it has turned out to be, does not exhaust the meaning of the word 'information'. The goal of information theory, or better, communication theory, is to characterize the sources

of information, to measure the capacity of the communication channels, and to learn how to control the degrading effects of noise. It is somewhat ironic but nevertheless true that this "information" theory is unconcerned with the central Bayesian issue of how message affect the beliefs of a rational agent.[2]

A fully Bayesian information theory demands an explicit account of the relation between information and the beliefs of an ideally rational agent.

Implicit in the recognition that most of our beliefs are held on the basis of incomplete information is the idea that our beliefs would be better if only we had more information. Thus a theory of probability demands a theory for updating probabilities. The desire and need to update our beliefs is driven by the conviction that not all probability assignments are equally good. In fact, it is a presupposition of thought itself that some beliefs are better than others — otherwise why go through the trouble of thinking?

The concern with 'good' and 'better' bears on the issue of whether probabilities are subjective, objective, or somewhere in between. We argued earlier that what makes one probability assignment better than another is that it better reflects something "objective" about the world. What precisely it is that we are being objective about is a difficult philosophical conundrum which, fortunately, we do not need to address. Suffice it to say that the adoption of better beliefs has real consequences: they provide a better guidance about how to cope with the world, and in this pragmatic sense, they provide a better guide to the "truth".

Objectivity is desirable; objectivity is the goal. Probabilities are useful to the extent that they incorporate some degree of objectivity. What we seek are updating mechanisms that allow us to process information and incorporate its objective features into our beliefs. Bayes' rule behaves precisely in this way. We saw in section 2.9.3 that as more and more data are taken into account the original (possibly subjective) prior becomes less and less relevant, and all rational agents become more and more convinced of the *same* truth. This is crucial: were it not this way Bayesian reasoning would not be deemed acceptable.

To set the stage for the discussion below consider some examples. Suppose a new piece of information is acquired. This could take a variety of forms. The typical example in data analysis would be something like: The prior probability of a certain proposition might have been q and after analyzing some data we feel rationally justified in asserting that a better assignment would be p. More explicitly, propositions such as "the value of the variable X lies between $x - \varepsilon$ and $x + \varepsilon$" might initially have had probabilities that were broadly spread over the range of x and after a measurement is performed the new data might induce us to revise our beliefs to a distribution that favors values in a narrower more localized region.

The typical example in statistical mechanics would run something like this: Total ignorance about the state of a system is expressed by a prior distribution that assigns equal probabilities to all microstates. The information that the

[2]We mentioned earlier, and emphasize again here, that the qualifier 'rational' is crucial: we are interested in the reasoning of an idealized rational agent and not of real imperfect humans.

system happens to be in thermal equilibrium induces us to update such beliefs to a probability distribution satisfying the constraint that the expected energy takes on a specific value, $\langle \varepsilon \rangle = E$.

Here is another more generic example. Let's say we have received a message — but the carrier of information could equally well have been in the form of input from our senses or data from an experiment. If the message agrees with our prior beliefs we can safely ignore it. The message is boring; it carries no news; literally, it carries no information. The interesting situation arises when the message surprises us; it is not what we expected. A message that disagrees with our prior beliefs presents us with a problem that demands a decision. If the source of the message is not deemed reliable then the contents of the message can be safely ignored — it carries no information; it is no different from noise. On the other hand, if the source of the message is deemed reliable then we have an opportunity to improve our beliefs — we ought to update our beliefs to agree with the message. Choosing between these two options requires a rational decision, a judgement. The message (or the sensation, or the data) become information precisely at that moment when as a result of our evaluation we feel compelled to revise our beliefs.

We are now ready to address the question: What, after all, is 'information'? The main observation is that the result of being confronted with new information is to restrict our options as to what we are honestly and rationally allowed to believe. This, I propose, is the defining characteristic of information.

> *Information, in its most general form, is whatever affects and therefore constrains rational beliefs.*

Since our objective is to update from a prior distribution to a posterior when *new* information becomes available we can state that

> *New information is what forces a change of beliefs.*

New information is a set of constraints on the family of acceptable posterior distributions.

An important aspect of this notion is that for a rational agent, the identification of what constitutes information — as opposed to mere noise — already involves a judgement, an evaluation; it is a matter of facts and also a matter of values. Furthermore, once a certain proposition has been identified as information, the revision of beliefs acquires a moral component; it is no longer optional: it becomes a moral imperative.

Our definition captures an idea of information that is directly related to changing our minds: information is the driving force behind the process of learning. But note that although there is no need to talk about amounts of information, whether measured in units of bits or otherwise, our notion of information allows precise quantitative calculations. Indeed, constraints on the acceptable posteriors are precisely the kind of information the method of maximum entropy (to be developed below) is designed to handle.

The constraints that convey, or rather, that *are* information can take a wide variety of forms including, in addition to the examples above, anything that affects beliefs. For example, in Bayesian inference both the prior distribution and the likelihood function constitute valuable information — they are not something that can be measured but they certainly contribute to constrain our beliefs. And constraints need not be just in the form of expected values; they can specify the functional form of a distribution or be imposed through various geometrical relations. (See chapters 8, 9, and also [Caticha 2001, Caticha Cafaro 2007].)

Concerning the act of updating it may be worthwhile to point out an analogy with dynamics — the study of change. In Newtonian dynamics the state of motion of a system is described in terms of momentum — the "quantity" of motion — while the change from one state to another is explained in terms of an applied force. Similarly, in Bayesian inference a state of belief is described in terms of probabilities — a "quantity" of belief — and the change from one state to another is due to information. Just as a force is that which induces a change from one state of motion to another, so *information is that which induces a change from one state of belief to another*. Updating is a form of dynamics.

What about prejudices and superstitions? What about divine revelations? Do they constitute information? Perhaps they lie outside our restriction to beliefs of *ideally rational agents*, but to the extent that their effects are indistinguishable from those of other sorts of information, namely, they affect beliefs, they should qualify as information too. Whether the sources of such information are reliable or not is quite another matter. False information is information too. In fact, even ideally rational agents can be affected by false information because the evaluation that assures them that the data was competently collected or that the message originated from a reliable source involves an act of judgement that is not completely infallible. Strictly, all those judgements, which constitute the first step of the inference process, are themselves the end result of other inference processes that are not immune from uncertainty.

What about limitations in our computational power? Such practical limitations are unavoidable and they do influence our inferences so should they be considered information? No. Limited computational resources may affect the numerical approximation to the value of, say, an integral, but they do not affect the actual value of the integral. Similarly, limited computational resources may affect the approximate imperfect reasoning of real humans and real computers but they do not affect the reasoning of those ideal rational agents that are the subject of our present concerns.

6.2 The design of entropic inference

Once we have decided, as a result of the confrontation of new information with old beliefs, that our beliefs require revision the problem becomes one of deciding how precisely this ought to be done. First we identify some general features of the kind of belief revision that one might consider desirable, of the kind of

belief revision that one might count as rational. Then we design a method, a systematic procedure, that implements those features. To the extent that the method performs as desired we can claim success. The point is not that success derives from our method having achieved some intimate connection to the inner wheels of reality; success just means that the method seems to be working. Whatever criteria of rationality we choose, they are meant to be only provisional — they are not immune from further change and improvement.

Typically the new information will not affect our beliefs in just one proposition — in which case the updating would be trivial. Tensions immediately arise because the beliefs in various propositions are not independent; they are interconnected by demands of consistency. Therefore the new information also affects our beliefs in all those "neighboring" propositions that are directly linked to it, and these in turn affect their neighbors, and so on. The effect can potentially spread over the whole network of beliefs; it is the whole web of beliefs that must be revised.

The one obvious requirement is that the updated beliefs ought to agree with the newly acquired information. Unfortunately, this requirement, while necessary, is not sufficiently restrictive: we can update in many ways that preserve both internal consistency and consistency with the new information. Additional criteria are needed. What rules is it rational to choose?

6.2.1 General criteria

The rules are motivated by the same pragmatic design criteria that motivate the design of probability theory itself — universality, consistency, and practical utility. But this is admittedly too vague; we must be very specific about the precise way in which they are implemented.

Universality

The goal is to design a method for induction, for reasoning when not much is known. In order for the method to perform its function we must impose that it be of *universal* applicability. Consider the alternative: We could design methods that are problem-specific, and employ different induction methods for different problems. Such a framework, unfortunately, would fail us precisely when we need it most, namely, in those situations where the information available is so incomplete that we do not know which method to employ.

We can argue this point somewhat differently: It is quite conceivable that different situations could require different problem-specific induction methods. What we want to design here is a general-purpose method that captures what all the other problem-specific methods have in common.

Parsimony

To specify the updating we adopt a very conservative criterion that recognizes the value of information: what has been laboriously learned in the past is valu-

able and should not be disregarded unless rendered obsolete by new information. The only aspects of one's beliefs that should be updated are those for which new evidence has been supplied. Thus we adopt a

Principle of Minimal Updating: *Beliefs should be updated only to the extent required by the new information.*

This version of the principle generalizes the earlier version presented in section 2.9.2 which was restricted to information in the form of data.

The special case of updating in the absence of new information deserves special attention. It states that when there is no new information an ideally rational agent should not change its mind.[3] In fact, it is difficult to imagine any notion of rationality that would allow the possibility of changing one's mind for no apparent reason. This is important and it is worthwhile to consider it from a different angle. Degrees of belief, probabilities, are said to be subjective: two different individuals might not share the same beliefs and could conceivably assign probabilities differently. But subjectivity does not mean arbitrariness. It is not a blank check allowing the rational agent to change its mind for no good reason. Valuable prior information should not be discarded unless it is absolutely necessary.

Minimal updating offers yet another pragmatic advantage. As we shall see below, rather than identifying what features of a distribution are singled out for updating and then specifying the detailed nature of the update, we will adopt design criteria that stipulate what is not to be updated. The practical advantage of this approach is that it enhances objectivity — there are many ways to change something but only one way to keep it the same.

The analogy with mechanics can be pursued further: if updating is a form of dynamics, then minimal updating is the analogue of inertia. Rationality and objectivity demand a considerable amount of inertia.

Independence

The next general requirement turns out to be crucially important: without it the very possibility of scientific theories would not be possible. The point is that every scientific model, whatever the topic, if it is to be useful at all, must assume that all relevant variables have been taken into account and that whatever was left out — the rest of the universe — should not matter. To put it another way: in order to do science we must be able to understand parts of the universe without having to understand the universe as a whole. Granted, it is not necessary that the understanding be complete and exact; it must just be adequate for our purposes.

The assumption, then, is that it is possible to focus our attention on a suitably chosen system of interest and neglect the rest of the universe because they

[3] We refer to ideally rational agents who have fully processed all information acquired in the past. Humans do not normally behave this way; they often change their minds by processes that are not fully conscious.

are "sufficiently independent". Thus, in any form of science the notion of statistical independence must play a central and privileged role. This idea — that some things can be neglected, that not everything matters — is implemented by imposing a criterion that tells us how to handle independent systems. The requirement is quite natural: *When two systems are a priori believed to be independent and we receive information about one then it should not matter if the other is included in the analysis or not (and vice versa).* This amounts to requiring that independence be preserved unless information about correlations is explicitly introduced.

The independence requirement is rather subtle and one must be careful about its precise implementation. To demonstrate the robustness of the design we provide (in a later section) an alternative version that takes the form of a consistency constraint: *Whenever systems are known to be independent it should not matter whether the analysis treats them jointly or separately.*

Again we emphasize: none of these criteria are imposed by Nature. They are desirable for pragmatic reasons; they are imposed by design.

6.2.2 Entropy as a tool for updating probabilities

Consider a variable x the value of which is uncertain; x can be discrete or continuous, in one or in several dimensions. It could, for example, represent the possible microstates of a physical system, a point in phase space, or an appropriate set of quantum numbers. The uncertainty about x is described by a probability distribution $q(x)$. Our goal is to update from the prior distribution $q(x)$ to a posterior distribution $p(x)$ when new information — by which we mean a set of constraints — becomes available. The question is: which distribution among all those that are in principle acceptable — they all satisfy the constraints — should we select?

Our goal is to design a method that allows a systematic search for the preferred posterior distribution. The central idea, first proposed in [Skilling 1988],[4] is disarmingly simple: to select the posterior first rank all candidate distributions in increasing *order of preference* and then pick the distribution that ranks the highest. Irrespective of what it is that makes one distribution preferable over another (we will get to that soon enough) it is clear that any ranking according to preference must be transitive: if distribution p_1 is preferred over distribution p_2, and p_2 is preferred over p_3, then p_1 is preferred over p_3. Such transitive rankings are implemented by assigning to each $p(x)$ a real number $S[p]$, which is called the entropy of p, in such a way that if p_1 is preferred over p_2, then $S[p_1] > S[p_2]$. The selected distribution (one or possibly many, for there may be several equally preferred distributions) is that which maximizes the entropy functional.

The importance of this particular approach to ranking distributions cannot be overestimated: it implies that the updating method will take the form of a

[4] [Skilling 88] deals with the more general problem of ranking positive additive distributions which includes intensity as well as probability distributions.

variational principle — the method of Maximum Entropy (ME) — and that the latter will involve a certain functional — the entropy — that maps distributions to real numbers and that is designed to be maximized. These features are not imposed by Nature; they are all imposed by design. They are dictated by the function that the ME method is supposed to perform. (Thus, it makes no sense to seek a generalization in which entropy is a complex number or a vector; such a generalized entropy would just not perform the desired function.)

Next we specify the ranking scheme, that is, we choose a specific functional form for the entropy $S[p]$. Note that *the purpose of the method is to update from priors to posteriors* so the ranking scheme must depend on the particular prior q and therefore the entropy S must be a functional of both p and q. The entropy $S[p, q]$ describes a ranking of the distributions p *relative* to the given prior q. $S[p, q]$ is the entropy of p *relative* to q, and accordingly $S[p, q]$ is commonly called *relative entropy*. This is appropriate and sometimes we will follow this practice. However, since all entropies are relative, even when relative to a uniform distribution, the qualifier 'relative' is redundant and can be dropped. This is somewhat analogous to the situation with energy: it is implicitly understood that all energies are relative to some reference frame but there is no need to constantly refer to a 'relative energy' — it is just not convenient.

The functional $S[p, q]$ is designed by a process of elimination — one might call it a process of *eliminative induction*. First we state the desired design criteria; this is the crucial step that defines what makes one distribution preferable over another. Then we analyze how each criterion constrains the form of the entropy. As we shall see the design criteria adopted below are sufficiently constraining that there is a single entropy functional $S[p, q]$ that survives the process of elimination.

This approach has a number of virtues. First, to the extent that the design criteria are universally desirable, then the single surviving entropy functional will be of universal applicability too. Second, the reason why alternative entropy candidates are eliminated is quite explicit — at least one of the design criteria is violated. Thus, *the justification behind the single surviving entropy is not that it leads to demonstrably correct inferences, but rather, that all other candidate entropies demonstrably fail to perform as desired.*

6.2.3 Specific design criteria

Three criteria and their consequences for the functional form of the entropy are given below. Detailed proofs are deferred to the next section.

Locality

DC1 *Local information has local effects.*

Suppose the information to be processed does *not* refer to a particular subdomain $\mathcal{D}$ of the space $\mathcal{X}$ of xs. In the absence of any new information about $\mathcal{D}$ the PMU demands we do not change our minds about probabilities that are conditional on $\mathcal{D}$. Thus, we design the inference method so that $q(x|\mathcal{D})$, the prior

probability of x conditional on $x \in \mathcal{D}$, is not updated. The selected conditional posterior is

$$P(x|\mathcal{D}) = q(x|\mathcal{D}) \ . \tag{6.1}$$

(The notation will be as follows: we denote priors by q, candidate posteriors by lower case p, and the selected posterior by upper case P.)

We emphasize: the point is not that we make the unwarranted assumption that keeping $q(x|\mathcal{D})$ unchanged is guaranteed to lead to correct inferences. It need not; induction is risky. The point is, rather, that in the absence of any evidence to the contrary there is no reason to change our minds and the prior information takes priority.

The consequence of DC1 is that non-overlapping domains of x contribute additively to the entropy,

$$S[p,q] = \int dx\, F\left(p(x), q(x), x\right) \ , \tag{6.2}$$

where F is some unknown function — not a functional, just a regular function of three arguments. The proof is given in section 6.3.

Comment 1:

It is essential that DC1 refers to conditional probabilities. An example may help to see why: Consider a loaded die with faces $f = 1 \ldots 6$. A priori we have no reason to favor any face, therefore $q(f) = 1/6$. Then we are told that the die is loaded in favor of 2. The criterion DC1 tells nothing about how to update the $P(f)$s. If the die were *very* loaded in favor of 2, say, $P(2) = 0.9$ then it must be that $P(f) < 1/6$ for $f \neq 2$ and therefore all $P(f)$s must be updated. Let us continue with the example: suppose we are further told that the die is loaded so that $p(2) = 2p(5)$. The criterion DC1 is meant to capture the fact that information about faces 2 and 5 does not change our preferences among the remaining four faces $\mathcal{D} = \{1, 3, 4, 6\}$; the DC1 implies that $P(f|\mathcal{D}) = q(f|\mathcal{D}) = 1/4$; it says nothing about whether $P(f)$ for $f \in \mathcal{D}$ is less or more than $1/6$.[5]

Comment 2:

If the variable x is continuous the criterion DC1 requires that information that refers to points infinitely close but just outside the domain $\mathcal{D}$ will have no influence on probabilities conditional on $\mathcal{D}$. This may seem surprising as it may lead to updated probability distributions that are discontinuous. Is this a problem? No.

In certain situations (*e.g.*, physics) we might have explicit reasons to believe that conditions of continuity or differentiability should be imposed and this information might be given to us in a variety of ways. The crucial point, however — and this is a point that we keep and will keep reiterating — is that unless such information is in fact explicitly given we should not assume it. If the new information leads to discontinuities, so be it.

Comment 3:

[5] For $f \in \mathcal{D}$, if $p(2) < 2/9$ then $P(f) > 1/6$; if $p(2) > 2/9$ then $P(f) < 1/6$.

The locality criterion DC1 includes Bayesian conditionalization as a special case. Indeed, if the information is given through the constraint $p(\mathcal{D}) = 1$ — or more precisely $p(\tilde{\mathcal{D}}) = 0$ where $\tilde{\mathcal{D}}$ is the complement of $\mathcal{D}$ so that the information does not directly refer to $\mathcal{D}$ — then $P(x|\mathcal{D}) = q(x|\mathcal{D})$, which is known as Bayesian conditionalization. More explicitly, if θ is the variable to be inferred on the basis of information about a likelihood function $q(x|\theta)$ and observed data x', then the update from the prior q to the posterior P,

$$q(x,\theta) = q(x)q(\theta|x) \rightarrow P(x,\theta) = P(x)P(\theta|x) \tag{6.3}$$

consists of updating $q(x) \rightarrow P(x) = \delta(x - x')$ to agree with the new information and invoking the PMU so that $P(\theta|x') = q(\theta|x')$ remains unchanged. Therefore,

$$P(x,\theta) = \delta(x - x')q(\theta|x) \quad \text{and} \quad P(\theta) = q(\theta|x') \ , \tag{6.4}$$

which is Bayes' rule (see sections 2.9.2 and 6.6 below). Thus, *entropic inference is designed to include Bayesian inference as a special case.* Note however that imposing locality is not identical to imposing Bayesian conditionalization — locality is more general because it is not restricted to absolute certainties such as $p(\mathcal{D}) = 1$.

Coordinate invariance

DC2 *The system of coordinates carries no information.*

The points $x \in \mathcal{X}$ can be labeled using any of a variety of coordinate systems. In certain situations we might have explicit reasons to believe that a particular choice of coordinates should be preferred over others and this information might have been given to us in a variety of ways, but unless it was in fact given we should not assume it: the ranking of probability distributions should not depend on the coordinates used.

The consequence of DC2 is that $S[p,q]$ can be written in terms of coordinate invariants such as $dx\, m(x)$ and $p(x)/m(x)$, and $q(x)/m(x)$:

$$S[p,q] = \int dx\, m(x) \Phi\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}\right) \ . \tag{6.5}$$

The proof is given in section 6.3. Thus the single unknown function F which had three arguments has been replaced by two unknown functions: Φ which has two arguments, and the density $m(x)$.

To grasp the meaning of DC2 it may be useful to recall some facts about coordinate transformations. Consider a change from old coordinates x to new coordinates x' such that $x = \Gamma(x')$. The new volume element dx' includes the corresponding Jacobian,

$$dx = \gamma(x')dx' \quad \text{where} \quad \gamma(x') = \left|\frac{\partial x}{\partial x'}\right| . \tag{6.6}$$

Let $m(x)$ be any density; the transformed density $m'(x')$ is such that $m(x)dx = m'(x')dx'$. This is true, in particular, for probability densities such as $p(x)$ and $q(x)$, therefore

$$m(x) = \frac{m'(x')}{\gamma(x')} \ , \quad p(x) = \frac{p'(x')}{\gamma(x')} \quad \text{and} \quad q(x) = \frac{q'(x')}{\gamma(x')} \, . \tag{6.7}$$

The coordinate transformation gives

$$\begin{aligned} S[p,q] &= \int dx \, F\left(p(x), q(x), x\right) \\ &= \int \gamma(x')dx' \, F\left(\frac{p'(x')}{\gamma(x')}, \frac{q'(x')}{\gamma(x')}, \Gamma(x')\right) , \end{aligned} \tag{6.8}$$

which is a mere change of variables. The identity above is valid always, for all Γ and for all F; it imposes absolutely no constraints on $S[p,q]$. The real constraint arises from realizing that we could have *started* in the x' coordinate frame, in which case we would have ranked the distributions using the entropy

$$S[p',q'] = \int dx' \, F\left(p'(x'), q'(x'), x'\right) \ , \tag{6.9}$$

but this should have no effect on our conclusions. This is the nontrivial content of DC2. It is not that we can change variables, we can always do that; but rather that the two rankings, the one according to $S[p,q]$ and the other according to $S[p',q']$ must coincide. This requirement is satisfied if, for example, $S[p,q]$ and $S[p',q']$ turn out to be numerically equal, but this is not necessary.

Locality (again)

Next we determine the density $m(x)$ by invoking the locality criterion DC1 once again. A situation in which no new information is available is dealt by allowing the domain $\mathcal{D}$ to cover the whole space of xs, $\mathcal{D} = \mathcal{X}$ and DC1 requires that in the absence of any new information the prior conditional probabilities should not be updated, $P(x|\mathcal{X}) = q(x|\mathcal{X})$ or $P(x) = q(x)$. Thus, when there are no constraints the selected posterior distribution should coincide with the prior distribution, which is expressed as

DC1′ *When there is no new information there is no reason to change one's mind and one shouldn't.*

The consequence of DC1′ (a second use of locality) is that the arbitrariness in the density $m(x)$ is removed: up to normalization $m(x)$ must be the prior distribution $q(x)$, and therefore at this point we have succeeded in restricting the entropy to functionals of the form

$$S[p,q] = \int dx \, q(x) \Phi\left(\frac{p(x)}{q(x)}\right) \ . \tag{6.10}$$

Independence

DC3 *When two systems are a priori believed to be independent and we receive independent information about each then it should not matter if one is included in the analysis of the other or not (and vice versa).*

Consider a composite system, $x = (x_1, x_2) \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$. Assume that all prior evidence led us to believe the systems were independent. This belief is reflected in the prior distribution: if the individual system priors $q_1(x_1)$ and $q_2(x_2)$, then the prior for the whole system is $q_1(x_1)q_2(x_2)$. Further suppose that new information is acquired such that $q_1(x_1)$ would by itself be updated to $P_1(x_1)$ and that $q_2(x_2)$ would be itself be updated to $P_2(x_2)$. DC3 requires that $S[p, q]$ be such that the joint prior $q_1(x_1)q_2(x_2)$ updates to the product $P_1(x_1)P_2(x_2)$ so that inferences about one do not affect inferences about the other.

The consequence of DC3 is that the remaining unknown function Φ is determined to be $\Phi(z) = -z \log z$. Thus, probability distributions $p(x)$ should be ranked relative to the prior $q(x)$ according to the relative entropy,

$$S[p, q] = -\int dx\, p(x) \log \frac{p(x)}{q(x)}. \tag{6.11}$$

Comment:

We emphasize that the point is not that when we have no evidence for correlations we draw the firm conclusion that the systems must necessarily be independent. They could indeed have turned out to be correlated and then our inferences would be wrong. Induction involves risk. The point is rather that if the joint prior reflected independence and the new evidence is silent on the matter of correlations, then the prior takes precedence and there is no reason to change our minds. As before, a feature of the probability distribution — in this case, independence — will not be updated unless the evidence requires it.

Comment:

We also emphasize that *DC3 is not a consistency requirement.* The argument we deploy is *not* that both the prior *and* the new information tell us the systems are independent in which case consistency requires that it should not matter whether the systems are treated jointly or separately. DC3 refers to a situation where the new information does not say whether the systems are independent or not. Rather, the updating is being *designed* so that the independence reflected in the prior is maintained in the posterior by default.

6.2.4 The ME method

We can now summarize the overall conclusion:

The ME method: *We want to update from a prior distribution q to a posterior distribution when there is new information in the form of constraints $\mathcal{C}$ that specify a family $\{p\}$ of allowed posteriors. The posterior is selected*

through a ranking scheme that recognizes the value of prior information and the privileged role of independence. Within the family $\{p\}$ the preferred posterior P is that which maximizes the relative entropy $S[p,q]$ subject to the available constraints. No interpretation for $S[p,q]$ is given and none is needed.

This extends the method of maximum entropy beyond its original purpose as a rule to assign probabilities from a given underlying measure (MaxEnt) to a method for updating probabilities from any arbitrary prior (ME). Furthermore, the logic behind the updating procedure does not rely on any particular meaning assigned to the entropy, either in terms of information, or heat, or disorder. Entropy is merely a tool for inductive inference; we do not need to know what entropy means; we only need to know how to use it.

Comment: In chapter 8 we will refine the method further. There we will address the question of assessing the extent to which distributions close to the entropy maximum ought to be included in the analysis. Their contribution — which accounts for fluctuation phenomena — turns out to be particularly significant in situations where the entropy maximum is not particularly sharp.

The derivation above has singled out *a unique $S[p,q]$ to be used in inductive inference.* Other "entropies" (such as, *e.g.*, $S_\eta[p,q]$ in eq.(6.53) below] might turn out to be useful for other purposes — perhaps as measures of some kinds of information, or measures of discrimination or distinguishability among distributions, or of ecological diversity, or for some altogether different function — but they are unsatisfactory in that they do not perform the function stipulated by the design criteria DC1-3.

6.3 The proofs

In this section we establish the consequences of the three criteria leading to the final result eq.(6.11). The details of the proofs are important not just because they lead to our final conclusions, but also because the translation of the verbal statement of the criteria into precise mathematical form is a crucial part of unambiguously specifying what the criteria actually say.

DC1: Locality

Here we prove that criterion DC1 leads to the expression eq.(6.2) for $S[p,q]$. The requirement that probabilities be normalized is handled by imposing normalization as one among so many other constraints that one might wish to impose. To simplify the proof we initially consider the case of a discrete variable, p_i with $i = 1 \ldots n$, so that $S[p,q] = S(p_1 \ldots p_n, q_1 \ldots q_n)$. The generalization to a continuum is straightforward.

Suppose the space of states $\mathcal{X}$ is partitioned into two non-overlapping domains $\mathcal{D}$ and $\tilde{\mathcal{D}}$ with $\mathcal{D} \cup \tilde{\mathcal{D}} = \mathcal{X}$, and that the information to be processed is

in the form of a constraint that refers to the domain $\tilde{\mathcal{D}}$,

$$\sum_{j\in\tilde{\mathcal{D}}} a_j p_j = A \ . \tag{6.12}$$

DC1 states that the constraint on $\tilde{\mathcal{D}}$ does not have an influence on the *conditional* probabilities $p_{i|\mathcal{D}}$. It may however influence the probabilities p_i within $\mathcal{D}$ through an overall multiplicative factor. To deal with this complication consider then a special case where the overall probabilities of $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are constrained too,

$$\sum_{i\in\mathcal{D}} p_i = P_{\mathcal{D}} \quad \text{and} \quad \sum_{j\in\tilde{\mathcal{D}}} p_j = P_{\tilde{\mathcal{D}}} \ , \tag{6.13}$$

with $P_{\mathcal{D}} + P_{\tilde{\mathcal{D}}} = 1$. Under these special circumstances constraints on $\tilde{\mathcal{D}}$ will not influence p_is within $\mathcal{D}$, and vice versa.

To obtain the posterior maximize $S[p,q]$ subject to these three constraints,

$$0 = \Bigg[\delta S - \lambda\left(\sum_{i\in\mathcal{D}} p_i - P_{\mathcal{D}}\right) + \\ - \tilde{\lambda}\left(\sum_{j\in\tilde{\mathcal{D}}} p_i - P_{\tilde{\mathcal{D}}}\right) + \mu\left(\sum_{j\in\tilde{\mathcal{D}}} a_j p_j - A\right)\Bigg] \ ,$$

leading to

$$\frac{\partial S}{\partial p_i} = \lambda \quad \text{for} \quad i\in\mathcal{D} \ , \tag{6.14}$$

$$\frac{\partial S}{\partial p_j} = \tilde{\lambda} + \mu a_j \quad \text{for} \quad j\in\tilde{\mathcal{D}} \ . \tag{6.15}$$

Eqs.(6.12-6.15) are $n+3$ equations we must solve for the p_is and the three Lagrange multipliers. Since $S = S(p_1\ldots p_n, q_1\ldots q_n)$ its derivative

$$\frac{\partial S}{\partial p_i} = f_i(p_1\ldots p_n, q_1\ldots q_n) \tag{6.16}$$

could in principle also depend on all $2n$ variables. But this violates the locality criterion because any arbitrary change in a_j within $\tilde{\mathcal{D}}$ would influence the p_is within $\mathcal{D}$. The only way that probabilities within $\mathcal{D}$ can be shielded from arbitrary changes in the constraints pertaining to $\tilde{\mathcal{D}}$ is that the functions f_i with $i\in\mathcal{D}$ depend only on p_is while the functions f_j depend only on p_js. Furthermore, this must hold not just for one particular partition of $\mathcal{X}$ into domains $\mathcal{D}$ and $\tilde{\mathcal{D}}$, it must hold for all conceivable partitions. Therefore f_i can depend only on p_i and, at this point, on any of the qs,

$$\frac{\partial S}{\partial p_i} = f_i(p_i, q_1\ldots q_n) \ . \tag{6.17}$$

But the power of the locality criterion is not exhausted yet. The information to be incorporated into the posterior can enter not just through constraints but

also through the prior. Suppose that the local information about domain $\tilde{\mathcal{D}}$ is altered by changing the prior within $\tilde{\mathcal{D}}$. Let $q_j \to q_j + \delta q_j$ for $j \in \tilde{\mathcal{D}}$. Then (6.17) becomes

$$\frac{\partial S}{\partial p_i} = f_i(p_i, q_1 \ldots q_j + \delta q_j \ldots q_n) \tag{6.18}$$

which shows that p_i with $i \in \mathcal{D}$ will be influenced by information about $\tilde{\mathcal{D}}$ unless f_i with $i \in \mathcal{D}$ is independent of all the q_js for $j \in \tilde{\mathcal{D}}$. Again, this must hold for all partitions into $\mathcal{D}$ and $\tilde{\mathcal{D}}$, and therefore,

$$\frac{\partial S}{\partial p_i} = f_i(p_i, q_i) \quad \text{for all} \quad i \in \mathcal{X} \ . \tag{6.19}$$

Integrating, one obtains

$$S[p, q] = \sum_i F_i(p_i, q_i) + \text{constant} \, . \tag{6.20}$$

for some undetermined functions F_i. The corresponding expression for a continuous variable x is obtained replacing i by x, and the sum over i by an integral over x leading to eq.(6.2),

$$S[p, q] = \int dx \, F\left(p(x), q(x), x\right) \ . \tag{6.21}$$

Remark: One might wonder whether in taking the continuum limit there might be room for introducing first and higher derivatives of p and q so that the function F might include more arguments,

$$F \stackrel{?}{=} F(p, q, \frac{dp}{dx}, \frac{dq}{dx}, \ldots; x) \ . \tag{6.22}$$

The answer is no! As discussed in the previous section one must not allow the inference method to introduce assumptions about continuity or differentiability unless such conditions are explicitly introduced as information. In the absence of any information to the contrary the prior information takes precedence; if this leads to discontinuities we must accept them. On the other hand, we may find ourselves in situations where our intuition insists that the discontinuities should just not be there. The right way to handle such situations (see section 4.11) is not to blame the method but the user: perhaps there is additional information concerning continuity that is relevant but we did not recognize it and failed to take it into account.

DC2: Coordinate invariance

Next we prove eq.(6.5). It is convenient to introduce an unspecified function $m(x)$ which transforms as a density and rewrite the expression (6.2) in the form

$$S[p, q] = \int dx \, m(x) \frac{1}{m(x)} F\left(\frac{p(x)}{m(x)} m(x), \frac{q(x)}{m(x)} m(x), x\right) \tag{6.23}$$

$$= \int dx \, m(x) \Phi\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m(x), x\right) , \tag{6.24}$$

where the function Φ is defined by

$$\Phi(\alpha, \beta, m, x) \stackrel{\text{def}}{=} \frac{1}{m} F(\alpha m, \beta m, x). \tag{6.25}$$

Next, we consider a special situation where the new information are constraints which do not favor one coordinate system over another. For example consider the constraint

$$\int dx\, p(x) a(x) = A \tag{6.26}$$

where $a(x)$ is a scalar function, that is, it is invariant under coordinate transformations,

$$a(x) \to a'(x') = a(x). \tag{6.27}$$

The usual normalization condition $\int dx\, p(x) = 1$ is a simple example of a scalar constraint.

Maximizing $S[p, q]$ subject to the constraint,

$$\delta \left[S[p, q] + \lambda \left(\int dx\, p(x) a(x) - A \right) \right] = 0, \tag{6.28}$$

gives

$$\dot{\Phi} \left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m(x), x \right) = \lambda a(x) \ , \tag{6.29}$$

where the dot represents the derivative with respect to the first argument,

$$\dot{\Phi}\left(\alpha, \beta, m, x\right) \stackrel{\text{def}}{=} \frac{\partial \Phi\left(\alpha, \beta, m, x\right)}{\partial \alpha} \tag{6.30}$$

But we could have started using the primed coordinates,

$$\dot{\Phi} \left(\frac{p'(x')}{m'(x')}, \frac{q'(x')}{m'(x')}, m'(x'), x' \right) = \lambda' a'(x'), \tag{6.31}$$

or, using (6.7) and (6.27),

$$\dot{\Phi} \left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}, m(x)\gamma(x'), x' \right) = \lambda' a(x). \tag{6.32}$$

Dividing (6.32) by (6.29) we get

$$\frac{\dot{\Phi}\left(\alpha, \beta, m\gamma, x'\right)}{\dot{\Phi}\left(\alpha, \beta, m, x\right)} = \frac{\lambda'}{\lambda}. \tag{6.33}$$

This identity should hold for any transformation $x = \Gamma(x')$. On the right hand side the multipliers λ and λ' are just constants; the ratio λ'/λ might depend on the transformation Γ but it does not depend on x. Consider the special case of a transformation Γ that has unit determinant everywhere, $\gamma = 1$, and differs from the identity transformation only within some arbitrary region $\mathcal{D}$. Since for

x outside this region $\mathcal{D}$ we have $x = x'$, the left hand side of eq.(6.33) equals 1. Thus, for this particular Γ the ratio is $\lambda'/\lambda = 1$; but $\lambda'/\lambda =$ constant, so $\lambda'/\lambda = 1$ holds within $\mathcal{D}$ as well. Therefore, for x within $\mathcal{D}$,

$$\dot{\Phi}\left(\alpha, \beta, m, x'\right) = \dot{\Phi}\left(\alpha, \beta, m, x\right). \tag{6.34}$$

Since the choice of $\mathcal{D}$ is arbitrary we conclude is that the function $\dot{\Phi}$ cannot depend on its fourth argument, $\dot{\Phi} = \dot{\Phi}\left(\alpha, \beta, m\right)$.

Having eliminated the fourth argument, let us go back to eq.(6.33),

$$\frac{\dot{\Phi}\left(\alpha, \beta, m\gamma\right)}{\dot{\Phi}\left(\alpha, \beta, m\right)} = \frac{\lambda'}{\lambda}\ , \tag{6.35}$$

and consider a different transformation Γ, one with unit determinant $\gamma = 1$ only outside the region $\mathcal{D}$. Therefore the constant ratio λ'/λ is again equal to 1, so that

$$\dot{\Phi}\left(\alpha, \beta, m\gamma\right) = \dot{\Phi}\left(\alpha, \beta, m\right). \tag{6.36}$$

But within $\mathcal{D}$ the transformation Γ is quite arbitrary, it could have any arbitrary Jacobian $\gamma \neq 1$. Therefore the function $\dot{\Phi}$ cannot depend on its third argument either, and therefore $\dot{\Phi} = \dot{\Phi}(\alpha, \beta)$. Integrating with respect to α gives $\Phi = \Phi(\alpha, \beta) +$ constant. The additive constant, which could depend on β, has no effect on the maximization and can be dropped. This completes the proof of eq.(6.5).

DC1′: Locality again

The locality criterion implies that when there are no constraints the selected posterior distribution should coincide with the prior distribution. This provides us with an interpretation of the as yet unspecified density $m(x)$. The argument is simple: maximize $S[p, q]$ in (6.5) subject to the single requirement of normalization,

$$\delta\left[S[p, q] + \lambda\left(\int dx\, p(x) - 1\right)\right] = 0, \tag{6.37}$$

to get

$$\dot{\Phi}\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}\right) = \lambda. \tag{6.38}$$

Since λ is a constant, the left hand side must be independent of x for arbitrary choices of the prior $q(x)$. This could, for example, be accomplished if the function $\dot{\Phi}(\alpha, \beta)$ were itself a constant, independent of its arguments α and β. But this gives

$$\Phi(\alpha, \beta) = c_1\alpha + c_2 \tag{6.39}$$

where c_1 and c_2 are constants and leads to the unacceptable form $S[p, q] \propto \int dx\, p(x) +$ constant.

If the independence on x cannot be eliminated by an appropriate choice of $\dot{\Phi}$, we must secure it by a choice of $m(x)$. Eq.(6.38) is an equation for

$p(x)$. In the absence of new information the selected posterior distribution must coincide with the prior, $P(x) = q(x)$. The obvious way to secure that (6.38) be independent of x is to choose $m(x) \propto q(x)$. Therefore $m(x)$ must, except for an overall normalization, be chosen to coincide with the prior distribution.

DC3: independence

If $x = (x_1, x_2) \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, and the individual system priors $q_1(x_1)$ and $q_2(x_2)$ are separately updated to $P_1(x_1)$ and $P_2(x_2)$ respectively, then we determine the unknown function Φ so that the joint prior for the combined system $q_1(x_1)q_2(x_2)$ is updated to $P_1(x_1)P_2(x_2)$.

We need only consider a special case of extremely constraining information: for system 1 we want to maximize $S_1[p_1, q_1]$ subject to the constraint that $p_1(x_1)$ is $P_1(x_1)$, the result being, naturally, $p_1(x_1) = P_1(x_1)$. A similar result holds for system 2. When the systems are treated jointly, however, the inference is not nearly as trivial. We want to maximize the entropy of the joint system,

$$S[p, q] = \int dx_1 dx_2\, q_1(x_1)q_2(x_2)\Phi\left(\frac{p(x_1, x_2)}{q_1(x_1)q_2(x_2)}\right) , \tag{6.40}$$

subject to the following constraints on the joint distribution $p(x_1, x_2)$:

$$\int dx_2\, p(x_1, x_2) = P_1(x_1) \qquad \text{and} \qquad \int dx_1\, p(x_1, x_2) = P_2(x_2) . \tag{6.41}$$

Notice that here we have not written just two constraints. We actually have one constraint for each value of x_1 and of x_2; this is an infinity of constraints, each of which must be multiplied by its own Lagrange multiplier, $\lambda_1(x_1)$ or $\lambda_2(x_2)$. Then,

$$\delta\left[S - \int dx_1 \lambda_1(x_1)\left(\int dx_2\, p(x_1, x_2) - p_1(x_1)\right) - \{1 \leftrightarrow 2\}\right] = 0, \tag{6.42}$$

where $\{1 \leftrightarrow 2\}$ indicates a third term, similar to the second, with 1 and 2 interchanged. The independent variations $\delta p(x_1, x_2)$ yield

$$\Phi'\left(\frac{p(x_1, x_2)}{q_1(x_1)q_2(x_2)}\right) = \lambda_1(x_1) + \lambda_2(x_2). \tag{6.43}$$

(The prime indicates a derivative with respect to the argument.)

Next we impose that the selected posterior be the product $P(x_1, x_2) = P_1(x_1)P_2(x_2)$ and find the Φ that delivers the desired posterior. Then,

$$\Phi'(y) = \lambda_1(x_1) + \lambda_2(x_2), \quad \text{where} \quad y = \frac{P_1(x_1)P_2(x_2)}{q_1(x_1)q_2(x_2)}. \tag{6.44}$$

Differentiating with respect to x_1 and to x_2, yields

$$y\Phi'''(y) + \Phi''(y) = 0 , \tag{6.45}$$

which can easily be integrated three times to give

$$\Phi(y) = ay \log y + by + c. \tag{6.46}$$

The additive constant c may be dropped: its contribution to the entropy would appear in a term that does not depend on the probabilities and would have no effect on the ranking scheme. At this point the entropy takes the form

$$S[p,q] = \int dx \left(ap(x) \log \frac{p(x)}{q(x)} + bp(x) \right). \tag{6.47}$$

This $S[p]$ will be maximized subject to constraints which will always include normalization. Since this is implemented by adding a term $\lambda \int dx\, p(x)$, the b constant can always be absorbed into the undetermined multiplier λ. Thus, the term $bp(x)$ has no effect on the selected distribution and can be dropped.

Finally, a is just an overall multiplicative constant, it also does not affect the overall ranking except in the trivial sense that inverting the sign of a will transform the maximization problem to a minimization problem or vice versa. We can therefore set $a = -1$ so that maximum S corresponds to maximum preference which gives us eq.(6.11) and concludes our derivation.

6.4 An alternative independence criterion: consistency

The robustness of the entropic inference framework can be illustrated by exploring an alternative version of the independence criterion. DC3 referred to new information that was silent on the matter of correlations among systems. Now we consider an alternative where the new information is not silent; it explicitly states that there are no correlations. Instead of DC3 we require the following consistency requirement:[6]

DC3′ *When systems are* ***known*** *to be independent it should not matter whether the inference procedure treats them separately or jointly.*

The consequence of DC3′ is that the remaining unknown function Φ is determined to be $\Phi(z) = -z \log z$. Thus, probability distributions $p(x)$ should be ranked relative to the prior $q(x)$ according to the relative entropy $S[p,q]$ given by eq.(6.11).

Criterion DC3′ is perhaps subtler than it might appear at first sight. Two points must be made clear. The first point concerns how the information about independence is to be handled as a constraint. Consider two (or more) systems that we know are independent. This means that both the prior and the posterior

[6]Since DC3 and its alternative DC3′ lead to the same entropic inference framework this section may be skipped on a first reading. The use of DC3′ involves a considerably more involved derivation but has the advantage of illuminating how it is that the alternative η-entropies, eq.(6.53), are ruled out.

are products. If the priors for the individual systems are $q_1(x_1)$ and $q_2(x_2)$, then the prior for the combined system is the product

$$q(x_1, x_2) = q_1(x_1)q_2(x_2) \ , \tag{6.48}$$

while the joint posterior is constrained within the family

$$p(x_1, x_2) = p_1(x_1)p_2(x_2) \ . \tag{6.49}$$

Further suppose that new information is acquired, say constraints $\mathcal{C}_1$ such that $q_1(x_1)$ is updated to $P_1(x_1)$, and constraints $\mathcal{C}_2$ such that $q_2(x_2)$ is updated to $P_2(x_2)$. The constraints $\mathcal{C}_1$ could, for example, include normalization, and in addition they could also involve the known expected value of a function $f_1(x_1)$,

$$\int dx_1 f_1(x_1)p_1(x_1) = \int dx_1 dx_2 \, f_1(x_1)p(x_1, x_2) = F_1 \ . \tag{6.50}$$

Criterion DC3′ is implemented as follows: First we treat the two individual systems separately. For system 1 the result of maximizing

$$S[p_1, q_1] = \int dx_1 \, q_1(x_1)\Phi\left(\frac{p_1(x_1)}{q_1(x_1)}\right) \ , \tag{6.51}$$

subject to constraints $\mathcal{C}_1$ on $p_1(x_1)$ is to select the posterior $P_1(x_1)$. Similarly, for system 2 maximizing $S[p_2, q_2]$ subject to constraints $\mathcal{C}_2$ on $p_2(x_2)$ is to select the posterior $P_2(x_2)$.

Next the two systems are treated jointly. Since we are concerned with a situation where we have the information that the systems are independent, we are *required* to search for the posterior within the restricted family of joint distributions that take the form of the product (6.49); this is an *additional* constraint $\mathcal{C}_3$ over and above the original $\mathcal{C}_1$ and $\mathcal{C}_2$. The new constraint $\mathcal{C}_3 = \{ p = p_1p_2\}$ is easily implemented by direct substitution. Instead of maximizing the joint entropy, $S[p, q_1q_2]$, we maximize

$$S[p_1p_2, q_1q_2] = \int dx_1 dx_2 \, q_1(x_1)q_2(x_2)\Phi\left(\frac{p_1(x_1)p_2(x_2)}{q_1(x_1)q_2(x_2)}\right), \tag{6.52}$$

under independent variations δp_1 and δp_2 subject to the same constraints $\mathcal{C}_1$ and $\mathcal{C}_2$. The function Φ is then designed — or at least constrained — so that that the selected posterior be $P_1(x_1)P_2(x_2)$.

The second point is that the criterion DC3′ is universal; it applies to *all instances* of systems that happen to be independent. DC3′ applies to situations where the independent systems are identical, and also when they are not; it applies when we deal with just two systems — as in the previous paragraph — and it also applies when we deal with many, whether just a few or a very large number.

Remark: Imposing DC3′ in its full generality, which includes the limit of a large number of independent systems, leads to an inductive framework that is consistent with the laws of large numbers.[7]

[7]Conversely, imposing DC3′ for just two identical independent systems is bad design: it leads to inductive frameworks that are inconsistent with laws of large numbers.

The lengthy proof leading to (6.11) given below involves three steps. First we show that applying DC3′ to independent systems that happen to be identical restricts the entropy functional to a member of the family of entropies

$$S_\eta[p,q] = \frac{1}{\eta(\eta+1)}\left(1 - \int dx\, p^{\eta+1} q^{-\eta}\right) \tag{6.53}$$

labeled by a single real parameter η.

It is easy to see that there are no singularities for $\eta = 0$ or -1; the limits $\eta \to 0$ and $\eta \to -1$ are well behaved. For example, to take $\eta \to 0$ use

$$y^\eta = \exp(\eta \log y) \approx 1 + \eta \log y \ , \tag{6.54}$$

which leads to the usual logarithmic entropy, $S_0[p,q] = S[p,q]$ given in eq.(6.11). Similarly, for $\eta \to -1$ we get $S_{-1}[p,q] = S[q,p]$.

In the second step DC3′ is applied to two independent systems that are not identical and could in principle be described by different parameters η_1 and η_2. The consistency demanded by DC3′ implies that the two parameters must be equal, $\eta_1 = \eta_2$, and since this must hold for all pairs of independent systems consistency demands that η must be a universal constant. In the third and final step the value of this constant — which turns out to be $\eta = 0$ — is determined by demanding that DC3′ apply to N identical systems where N is very large.

Step 1: Consistency for identical independent systems

In this subsection we show that applying DC3′ to systems that happen to be identical restricts the entropy functional to a member of the one-parameter family of η-entropies $S_\eta[p,q]$ parametrized by η. For $\eta = 0$ one obtains the standard logarithmic entropy, eq.(6.11),

$$S_0[p,q] = -\int dx\, p(x) \log \frac{p(x)}{q(x)} \ . \tag{6.55}$$

For $\eta = -1$ one obtains

$$S_{-1}[p,q] = \int dx\, q(x) \log \frac{p(x)}{q(x)} \ , \tag{6.56}$$

which coincides with $S_0[q,p]$ with the arguments switched. Finally, for a generic value of $\eta \neq -1, 0$ the result is

$$S_\eta[p,q] = -\int dx\, p(x) \left(\frac{p(x)}{q(x)}\right)^\eta \ . \tag{6.57}$$

It is worthwhile to recall that the objective of this whole exercise is to rank probability distributions according to preference and therefore different entropies that induce the same ranking scheme are effectively equivalent. This is very convenient as it allows considerable simplifications by an appropriate

choice of additive and multiplicative constants. Taking advantage of this freedom we can, for example, combine the three expressions (6.55), (6.56), and (6.57) into the single expression

$$S_\eta[p,q] = \frac{1}{\eta(\eta+1)}\left(1 - \int dx\, p^{\eta+1} q^{-\eta}\right) \ , \tag{6.58}$$

that we met earlier in eq.(6.53).

The proof below is fairly lengthy and may be skipped on a first reading. It follows the treatment in [Caticha Giffin 06] and is based upon and extends a previous proof by Karbelkar who showed that belonging to the family of η-entropies is a sufficient condition to satisfy the consistency criterion for *identical* systems. He conjectured but did not prove that this was perhaps also a necessary condition [Karbelkar 86]. Although necessity was not essential to his argument it is crucial for ours. We show below that for identical systems there are no acceptable entropies outside the S_η family.

First we treat the two systems separately. For system 1 we maximize the entropy $S[p_1, q_1]$ subject to normalization and the constraint $\mathcal{C}_1$ in eq.(6.50). Introduce Lagrange multipliers α_1 and λ_1,

$$\delta\left[S[p_1,q_1] - \lambda_1\left(\int dx_1 f_1 p_1 - F_1\right) - \alpha_1\left(\int dx_1 p_1 - 1\right)\right] = 0, \tag{6.59}$$

which gives

$$\Phi'\left(\frac{p_1(x_1)}{q_1(x_1)}\right) = \lambda_1 f_1(x_1) + \alpha_1 \ , \tag{6.60}$$

where the prime indicates a derivative with respect to the argument, $\Phi'(y) = d\Phi(y)/dy$. For system 2 we need only consider the extreme situation where the constraints $\mathcal{C}_2$ determine the posterior completely: $p_2(x_2) = P_2(x_2)$.

Next we treat the two systems jointly. The constraints $\mathcal{C}_2$ are easily implemented by direct substitution and thus, we maximize the entropy $S[p_1P_2, q_1q_2]$ by varying p_1 subject to normalization and the constraint $\mathcal{C}_1$ in eq.(6.50). Introduce Lagrange multipliers α and λ,

$$\delta\left[S[p_1P_2,q_1q_2] - \lambda\left(\int dx_1 f_1 p_1 - F_1\right) - \alpha\left(\int dx_1\, p_1 - 1\right)\right] = 0, \tag{6.61}$$

which gives

$$\int dx_2\, P_2 \Phi'\left(\frac{p_1P_2}{q_1q_2}\right) = \lambda[P_2,q_2] f_1(x_1) + \alpha[P_2,q_2] \ , \tag{6.62}$$

where the multipliers λ and α are independent of x_1 but could in principle be functionals of P_2 and q_2.

The consistency condition that constrains the form of Φ is that if the solution to eq.(6.60) is $P_1(x_1)$ then the solution to eq.(6.62) must also be $P_1(x_1)$, and this must be true irrespective of the choice of $P_2(x_2)$. Let us then consider a small change $P_2 \to P_2 + \delta P_2$ that preserves the normalization of P_2. First introduce a Lagrange multiplier α_2 and rewrite eq.(6.62) as

$$\int dx_2\, P_2 \Phi'\left(\frac{P_1P_2}{q_1q_2}\right) - \alpha_2\left[\int dx_2 P_2 - 1\right] = \lambda[P_2,q_2] f_1(x_1) + \alpha[P_2,q_2] \ , \tag{6.63}$$

where we have replaced p_1 by the known solution P_1 and thereby effectively transformed eqs.(6.60) and (6.62) into an equation for Φ. The δP_2 variation gives,

$$\Phi'\left(\frac{P_1P_2}{q_1q_2}\right)+\frac{P_1P_2}{q_1q_2}\Phi''\left(\frac{P_1P_2}{q_1q_2}\right)=\frac{\delta\lambda}{\delta P_2}f_1(x_1)+\frac{\delta\alpha}{\delta P_2}+\alpha_2\ . \tag{6.64}$$

Next use eq.(6.60) to eliminate $f_1(x_1)$,

$$\Phi'\left(\frac{P_1P_2}{q_1q_2}\right)+\frac{P_1P_2}{q_1q_2}\Phi''\left(\frac{P_1P_2}{q_1q_2}\right)=A[P_2,q_2]\Phi'\left(\frac{P_1}{q_1}\right)+B[P_2,q_2]\ , \tag{6.65}$$

where

$$A[P_2,q_2]=\frac{1}{\lambda_1}\frac{\delta\lambda}{\delta P_2}\quad\text{and}\quad B[P_2,q_2]=-\frac{\delta\lambda}{\delta P_2}\frac{\alpha_1}{\lambda_1}+\frac{\delta\alpha}{\delta P_2}+\alpha_2\ , \tag{6.66}$$

are at this point unknown functionals of P_2 and q_2. Differentiating eq.(6.65) with respect to x_1 the B term drops out and we get

$$A[P_2,q_2]=\left[\frac{d}{dx_1}\Phi'\left(\frac{P_1}{q_1}\right)\right]^{-1}\frac{d}{dx_1}\left[\Phi'\left(\frac{P_1P_2}{q_1q_2}\right)+\frac{P_1P_2}{q_1q_2}\Phi''\left(\frac{P_1P_2}{q_1q_2}\right)\right]\ , \tag{6.67}$$

which shows that A is not a functional of P_2 and q_2 but a mere function of P_2/q_2. Substituting back into eq.(6.65) we see that the same is true for B. Therefore eq.(6.65) can be written as

$$\Phi'\left(y_1y_2\right)+y_1y_2\Phi''\left(y_1y_2\right)=A(y_2)\Phi'\left(y_1\right)+B(y_2)\ , \tag{6.68}$$

where $y_1=P_1/q_1$, $y_2=P_2/q_2$, and $A(y_2)$, $B(y_2)$ are unknown functions of y_2.

Now we specialize to identical systems. Then we can exchange the labels $1\leftrightarrow 2$, and we get

$$A(y_2)\Phi'\left(y_1\right)+B(y_2)=A(y_1)\Phi'\left(y_2\right)+B(y_1)\ . \tag{6.69}$$

To find the unknown functions A and B differentiate with respect to y_2,

$$A'(y_2)\Phi'\left(y_1\right)+B'(y_2)=A(y_1)\Phi''\left(y_2\right) \tag{6.70}$$

and then with respect to y_1 to get

$$\frac{A'(y_1)}{\Phi''\left(y_1\right)}=\frac{A'(y_2)}{\Phi''\left(y_2\right)}=a=\text{const}\ . \tag{6.71}$$

Integrate to get

$$A(y_1)=a\Phi'\left(y_1\right)+b\ , \tag{6.72}$$

then substitute back into eq.(6.70) and integrate again to get

$$B'(y_2)=b\Phi''\left(y_2\right)\quad\text{and}\quad B(y_2)=b\Phi'\left(y_2\right)+c\ , \tag{6.73}$$

where b and c are constants. We can check directly that $A(y)$ and $B(y)$ are indeed solutions of eq.(6.69). Substituting into eq.(6.68) gives

$$\Phi'(y_1y_2) + y_1y_2\Phi''(y_1y_2) = a\Phi'(y_1)\,\Phi'(y_2) + b\,[\Phi'(y_1) + \Phi'(y_2)] + c\ . \quad (6.74)$$

This is a peculiar differential equation. We can think of it as one differential equation for $\Phi'(y_1)$ for each given constant value of y_2 but there is a complication in that the various (constant) coefficients $\Phi'(y_2)$ are themselves unknown. To solve for Φ choose a fixed value of y_2, say $y_2 = 1$,

$$y\Phi''(y) - \eta\Phi'(y) - \kappa = 0\ , \quad (6.75)$$

where $\eta = a\Phi'(1) + b - 1$ and $\kappa = b\Phi'(1) + c$. To eliminate the constant κ differentiate with respect to y,

$$y\Phi''' + (1-\eta)\,\Phi'' = 0\ , \quad (6.76)$$

which is a linear homogeneous equation and is easy to integrate.

For generic values of $\eta \neq -1, 0$ the solution is

$$\Phi''(y) \propto y^{\eta-1} \Rightarrow \Phi'(y) = \alpha y^\eta + \beta\ . \quad (6.77)$$

The constants α and β are chosen so that this is a solution of eq.(6.74) for all values of y_2 (and not just for $y_2 = 1$). Substituting into eq.(6.74) and equating the coefficients of various powers of y_1y_2, y_1, and y_2 gives three conditions on the two constants α and β,

$$\alpha(1+\eta) = a\alpha^2, \quad 0 = a\alpha\beta + b\alpha, \quad \beta = a\beta^2 + 2b\beta + c\ . \quad (6.78)$$

The nontrivial ($\alpha \neq 0$) solutions are $\alpha = (1+\eta)/a$ and $\beta = -b/a$, while the third equation gives $c = b(1-b)/4a$. We conclude that for generic values of η the solution of eq.(6.74) is

$$\Phi(y) = \frac{1}{a}y^{\eta+1} - \frac{b}{a}y + C\ , \quad (6.79)$$

where C is a new constant. Substituting into eq.(6.10) yields

$$S_\eta[p,q] = \frac{1}{a}\int dx\, p(x)\left(\frac{p(x)}{q(x)}\right)^\eta - \frac{b}{a}\int dx\, p(x) + C\int dx\, q(x)\ . \quad (6.80)$$

This complicated expression can be simplified considerably by exploiting the freedom to choose additive and multiplicative constants. We can drop the last two terms and choose $a = -1$ so that the preferred distribution is that which maximizes entropy. This reproduces eq.(6.57).

For $\eta = 0$ we return to eq.(6.76) and integrate twice to get

$$\Phi(y) = a'y\log y + b'y + c'\ , \quad (6.81)$$

for some new constants a', b', and c'. Substituting into eq.(6.10) yields

$$S_0[p,q] = a' \int dx\, p(x) \log \frac{p(x)}{q(x)} + b' \int dx\, p(x) + c' \int dx\, q(x) \ . \tag{6.82}$$

Again, choosing $a' = -1$ and dropping the last two terms does not affect the ranking scheme. This yields the standard expression for relative entropy, eq.(6.55).

Finally, for $\eta = -1$ integrating eq.(6.76) twice gives

$$\Phi(y) = a'' \log y + b'' y + c'' \ , \tag{6.83}$$

for some new constants a'', b'', and c''. Substituting into eq.(6.10) yields

$$S_0[p,q] = a'' \int dx\, q(x) \log \frac{p(x)}{q(x)} + b'' \int dx\, p(x) + c'' \int dx\, q(x) \ . \tag{6.84}$$

Again, choosing $a'' = 1$ and dropping the last two terms yields eq.(6.56). This completes our derivation.

Step 2: Consistency for non-identical systems

Let us summarize our results so far. The goal is to update probabilities by ranking the distributions according to an entropy S that is of general applicability. The allowed functional forms of the entropy S have been constrained down to a member of the one-dimensional family S_η. One might be tempted to conclude that there is no S of universal applicability; that inferences about different systems could to be carried out with different η-entropies. But we have not yet exhausted the full power of the consistency DC3. Consistency is universally desirable; there is no reason why it should be limited to identical systems.

To proceed further we ask: What is η? Is it a property of the individual carrying out the inference or of the system under investigation? The former is unacceptable; we insist that the updating must be objective in that different individuals with the same prior and with the same constraints must make the same inference. Therefore the "inference parameter" η can only be a property of the system.

Consider two different systems characterized by η_1 and η_2. Let us further suppose that these systems are known to be independent (perhaps system #1 lives here on Earth while system #2 lives in a distant galaxy) so that they fall under the jurisdiction of DC3′. Separate inferences about systems #1 and #2 are carried out with $S_{\eta_1}[p_1, q_1]$ and $S_{\eta_2}[p_2, q_2]$ respectively. For the combined system we are also required to use an η-entropy, say $S_\eta[p_1 p_2, q_1 q_2]$. Consistency is possible only if we impose $\eta_1 = \eta_2$ from the start.

But this is not all: consider a third system #3 that also lives here on Earth. We do not know whether system #3 is independent from system #1 or not but we can confidently assert that it will certainly be independent of the system #2 living in the distant galaxy. The argument of the previous paragraph leads

us to conclude that $\eta_3 = \eta_2$, and therefore that $\eta_3 = \eta_1$ even when systems #1 and #3 are not known to be independent! We conclude that *all systems must be characterized by the same parameter* η whether they are independent or not because we can always find a common reference system that is sufficiently distant to be independent of any two of them. The inference parameter η is a universal constant, the value of which is at this point still unknown.

The power of a consistency argument resides in its universal applicability: if an entropy $S[p,q]$ exists then it must be one chosen from among the $S_\eta[p,q]$. The remaining problem is to determine this universal constant η. Here we give one argument; in the next subsection we give another one.

One possibility is to regard η as a quantity to be determined experimentally. Are there systems for which inferences based on a known value of η have repeatedly led to success? The answer is yes; they are quite common.

As we discussed in Chapter 5 statistical mechanics and thus thermodynamics are theories of inference based on the value $\eta = 0$. The relevant entropy, which is the Boltzmann-Gibbs-Shannon entropy, can be interpreted as the special case of the ME when one updates from a uniform prior. It is an experimental fact *without any known exceptions* that inferences about *all* physical, chemical and biological systems that are in thermal equilibrium or close to it can be carried out by assuming that $\eta = 0$. Let us emphasize that this is not an obscure and rare example of purely academic interest; these systems comprise essentially all of natural science. (Included is every instance where it is useful to introduce a notion of temperature.)

In conclusion: consistency for non-identical systems requires that η be a universal constant and there is abundant experimental evidence for its value being $\eta = 0$. Other η-entropies may turn out to be useful for other purposes but *the logarithmic entropy* $S[p,q]$ *in eq.(6.11) provides the only consistent ranking criterion for updating probabilities that can claim general applicability.*

Step 3: Consistency with the law of large numbers

Here we offer a second argument, also based on a broader application of DC3′, that the value of the universal constant η must be $\eta = 0$.

DC3′ applies generally; in particular it applies to large numbers of independent identical systems. In such cases we can calculate η by demanding that entropic updates be consistent with the weak law of large numbers.

Let the state for each individual system be described by a discrete variable $i = 1 \ldots m$.

First we treat the individual systems separately. The identical priors for the individual systems are q_i and the available information is that the potential posteriors p_i are subject, for example, to an expectation value constraint such as $\langle a \rangle = A$, where A is some specified value and $\langle a \rangle = \sum a_i p_i$. The preferred posterior P_i is found maximizing the η-entropy $S_\eta[p,q]$ subject to $\langle a \rangle = A$.

To treat the systems jointly we let the number of systems found in state i be n_i, and let $f_i = n_i/N$ be the corresponding frequency. The two descriptions are related by the law of large numbers: for large N the frequencies f_i converge

(in probability) to the desired posterior P_i while the sample average $\bar{a} = \sum a_i f_i$ converges (also in probability) to the expected value $\langle a \rangle = A$.

Now we consider the set of N systems treated jointly. The probability of a particular frequency distribution $f = (f_1 \ldots f_n)$ generated by the prior q is given by the multinomial distribution,

$$Q_N(f|q) = \frac{N!}{n_1! \ldots n_m!} q_1^{n_1} \ldots q_m^{n_m} \quad \text{with} \quad \sum_{i=1}^{m} n_i = N \ . \tag{6.85}$$

When the n_i are sufficiently large we can use Stirling's approximation,

$$\log n! = n \log n - n + \log \sqrt{2\pi n} + O(1/n) \ . \tag{6.86}$$

Then

$$\begin{aligned} \log Q_N(f|q) &\approx N \log N - N + \log \sqrt{2\pi N} \\ &\quad - \sum_i \left(n_i \log n_i - n_i + \log \sqrt{2\pi n_i} - n_i \log q_i \right) \\ &= -N \sum_i \frac{n_i}{N} \log \frac{n_i}{N q_i} - \sum_i \log \sqrt{\frac{n_i}{N}} - (N-1) \log \sqrt{2\pi N} \\ &= N S[f,q] - \sum_i \log \sqrt{f_i} - (N-1) \log \sqrt{2\pi N} \ , \end{aligned} \tag{6.87}$$

where $S[f,q]$ is the $\eta = 0$ entropy given by eq.(6.11). Therefore for large N can be written as

$$Q_N(f|q) \approx C_N (\prod_i f_i)^{-1/2} \exp(N S[f,q]) \tag{6.88}$$

where C_N is a normalization constant. The Gibbs inequality $S[f,q] \leq 0$, eq.(4.23), shows that for large N the probability $Q_N(f|q)$ shows an exceedingly sharp peak. The most likely frequency distribution is numerically equal to the probability distribution q_i. This is the weak law of large numbers. Equivalently, we can rewrite it as

$$\frac{1}{N} \log Q_N(f|q) \approx S[f,q] + r_N \ , \tag{6.89}$$

where r_N is a correction that vanishes (in probability) as $N \to \infty$. This means that finding the most probable frequency distribution is equivalent to maximizing the entropy $S[f,q]$.

The most probable frequency distribution f is q. The most probable frequency distribution that satisfies the constraint $\bar{a} = A$ is the distribution that maximizes $Q_N(f|q)$ subject to the constraint $\bar{a} = A$, which is equivalent to maximizing the entropy $S[f,q]$ subject to $\bar{a} = A$. In the limit of large N the frequencies f_i converge (in probability) to the desired posterior P_i while the sample average $\bar{a} = \sum a_i f_i$ converges (also in probability) to the expected value $\langle a \rangle = A$. The two procedures agree only when we choose $\eta = 0$. The reason the alternative η-entropies are discarded is clear: $\eta \neq 0$ is inconsistent with the law of large numbers.

[Csiszar 1984] and [Grendar 2001] have argued that the asymptotic argument above provides by itself a valid justification for the ME method of updating. An agent whose prior is q receives the information $\langle a\rangle = A$ which can be reasonably interpreted as a sample average $\bar{a} = A$ over a large ensemble of N trials. The agent's beliefs are updated so that the posterior P coincides with the most probable f distribution. This is quite compelling but, of course, as a justification of the ME method it is restricted to situations where it is natural to think in terms of ensembles with large N. This justification is not nearly as compelling for singular events for which large ensembles either do not exist or are too unnatural and contrived. From our point of view the asymptotic argument above does not by itself provide a fully convincing justification for the universal validity of the ME method but it does provide considerable inductive support. It serves as a valuable consistency check that must be passed by any inductive inference procedure that claims to be of *general* applicability.

6.5 Random remarks

6.5.1 On priors

All entropies are relative entropies. In the case of a discrete variable, if one assigns equal a priori probabilities, $q_i = 1$, one obtains the Boltzmann-Gibbs-Shannon entropy, $S[p] = -\sum_i p_i \log p_i$. The notation $S[p]$ has a serious drawback: it misleads one into thinking that S depends on p only. In particular, we emphasize that whenever $S[p]$ is used, the prior measure $q_i = 1$ has been implicitly assumed. In Shannon's axioms, for example, this choice is implicitly made in his first axiom, when he states that the entropy is a function of the probabilities $S = S(p_1...p_n)$ and nothing else, and also in his second axiom when the uniform distribution $p_i = 1/n$ is singled out for special treatment.

The absence of an explicit reference to a prior q_i may erroneously suggest that prior distributions have been rendered unnecessary and can be eliminated. It suggests that it is possible to transform information (*i.e.*, constraints) directly into posterior distributions in a totally objective and unique way. This was Jaynes' hope for the MaxEnt program. If this were true the old controversy, of whether probabilities are subjective or objective, would have been resolved in favor of complete objectivity. But the prior $q_i = 1$ is implicit in $S[p]$; the postulate of equal a priori probabilities or Laplace's "Principle of Insufficient Reason" still plays a major, though perhaps hidden, role. Any claims that probabilities assigned using maximum entropy will yield absolutely objective results are unfounded; not all subjectivity has been eliminated. *Just as with Bayes' theorem, what is objective here is the manner in which information is processed to update from a prior to a posterior, and not the prior probabilities themselves. And even then the updating is objective because we have agreed to adopt very specific criteria — this is objectivity by design.*

Choosing the prior density $q(x)$ can be tricky. Sometimes symmetry considerations can be useful in fixing the prior (three examples were given in section

4.5) but otherwise there is no fixed set of rules to translate information into a probability distribution except, of course, for Bayes' theorem and the ME method themselves.

What if the prior $q(x)$ vanishes for some values of x? $S[p,q]$ can be infinitely negative when $q(x)$ vanishes within some region $\mathcal{D}$. In other words, the ME method confers an overwhelming preference on those distributions $p(x)$ that vanish whenever $q(x)$ does. One must emphasize that this is as it should be; it is not a problem. As we saw in section 2.9.4 a similar situation also arises in the context of Bayes' theorem where a vanishing prior represents a tremendously serious commitment because no amount of data to the contrary would allow us to revise it. In both ME and Bayes updating we should recognize the implications of assigning a vanishing prior. Assigning a very low but non-zero prior represents a safer and less prejudiced representation of one's beliefs.

For more on the choice of priors see the review [Kass Wasserman 1996]; in particular for entropic priors see [Rodriguez 1990-2003, Caticha Preuss 2004]

6.5.2 Comments on other axiomatizations

One feature that distinguishes the axiomatizations proposed by various authors is how they justify maximizing a functional. In other words, why *maximum entropy*? In the approach of Shore and Johnson this question receives no answer; it is just one of the axioms. Csiszar provides a better answer. He derives the 'maximize a functional' rule from reasonable axioms of regularity and locality [Csiszar 1991]. In Skilling's and in the approach developed here the rule is not derived, but it does not go unexplained either: it is imposed by design, it is justified by the function that S is supposed to perform, namely, to achieve a transitive ranking.

Both Shore and Johnson and Csiszar require, and it is not clear why, that updating from a prior must lead to a unique posterior, and accordingly, there is a restriction that the constraints define a convex set. In Skilling's approach and in the one advocated here there is no requirement of uniqueness, we are perfectly willing to entertain situations where the available information points to several equally preferable distributions. To this subject we will return in chapter 8.

There is another important difference between the axiomatic approach presented by Csiszar and the design approach presented here. Our ME method is designed to be of universal applicability. As with all inductive procedures, in any particular instance of induction can turn out to be wrong — perhaps because, for example, not all relevant information has been taken into account — but this does not change the fact that ME is still the unique inductive inference method obeying rational design criteria. On the other hand Csiszar's version of the MaxEnt method is not designed to generalize beyond its axioms. His method was developed for linear constraints and therefore he does not feel justified in carrying out his *deductions* beyond the cases of linear constraints. In our case, the application to non-linear constraints is precisely the kind of *induction* the ME method was designed to perform.

It is interesting that if instead of axiomatizing the inference process, one axiomatizes the entropy itself by specifying those properties expected of a measure of separation between (possibly unnormalized) distributions one is led to a continuum of η-entropies, [Amari 1985]

$$S_\eta[p,q] = \frac{1}{\eta(\eta+1)} \int dx \left[(\eta+1)p - \eta q - p^{\eta+1}q^{-\eta}\right] \ , \tag{6.90}$$

labelled by a parameter η. These entropies are equivalent, for the purpose of updating, to the relative Renyi entropies [Renyi 1961, Aczel 1975]. The shortcoming of this approach is that it is not clear when and how such entropies are to be used, which features of a probability distribution are being updated and which preserved, or even in what sense do these entropies measure an amount of information. Remarkably, if one further requires that S_η be additive over independent sources of uncertainty, as one could reasonably expect from a measure, then the continuum in η is restricted to just the two values $\eta = 0$ and $\eta = -1$ which correspond to the logarithmic entropies $S[p,q]$ and $S[q,p]$.

For the special case when p is normalized and a uniform prior $q = 1$ we get (dropping the integral over q)

$$S_\eta = \frac{1}{\eta}\left(1 - \frac{1}{\eta+1}\int dx\, p^\eta\right) \ . \tag{6.91}$$

A related entropy

$$S'_\eta = \frac{1}{\eta}\left(1 - \int dx\, p^{\eta+1}\right) \tag{6.92}$$

has been proposed in [Tsallis 1988] (see section 5.5) and forms the foundation of a so-called non-extensive statistical mechanics. Clearly these two entropies are equivalent in that they generate equivalent variational problems – maximizing S_η is equivalent to maximizing S'_η. To conclude our brief remarks on the entropies S_η we point out that quite apart from the difficulty of achieving consistency with the law of large numbers, some the probability distributions obtained maximizing S_η may also be derived through a more standard use of MaxEnt or ME as advocated in these lectures (section 5.5).

6.6 Bayes' rule as a special case of ME

Since the ME method and Bayes' rule are both designed for updating probabilities, and both invoke a Principle of Minimal Updating, it is important to explore the relations between them. In section 6.2.3 we showed that ME is designed to include Bayes' rule as a special case. Here we would like to revisit this topic in greater depth, and, in particular to explore some variations and generalizations [Caticha Giffin 2006].

As described in section 2.9 the goal is to update our beliefs about $\theta \in \Theta$ (θ represents one or many parameters) on the basis of three pieces of information: (1) the prior information codified into a prior distribution $q(\theta)$; (2) the data

$x \in \mathcal{X}$ (obtained in one or many experiments); and (3) the known relation between θ and x given by the model as defined by the sampling distribution or likelihood, $q(x|\theta)$. The updating consists of replacing the *prior* probability distribution $q(\theta)$ by a *posterior* distribution $P(\theta)$ that applies after the data has been processed.

The crucial element that will allow Bayes' rule to be smoothly incorporated into the ME scheme is the realization that before the data information is available not only we do not know θ, we do not know x either. Thus, the relevant space for inference is not Θ but the product space $\Theta \times \mathcal{X}$ and the relevant joint prior is $q(x, \theta) = q(\theta)q(x|\theta)$. Let us emphasize two points: first, the likelihood function is *prior* information too; and second, we should emphasize that the information about how x is related to θ is contained in the *functional form* of the distribution $q(x|\theta)$ — for example, whether it is a Gaussian or a Cauchy distribution or something else – and not in the actual values of the arguments x and θ which are, at this point, still unknown.

Next we collect data and the observed values turn out to be x'. We must update to a posterior that lies within the family of distributions $p(x, \theta)$ that reflect the fact that x is now known to be x',

$$p(x) = \int d\theta\, p(\theta, x) = \delta(x - x') \ . \tag{6.93}$$

This data information constrains but is not sufficient to determine the joint distribution

$$p(x, \theta) = p(x)p(\theta|x) = \delta(x - x')p(\theta|x') \ . \tag{6.94}$$

Any choice of $p(\theta|x')$ is in principle possible. So far the formulation of the problem parallels section 2.9 exactly. We are, after all, solving the same problem. Next we apply the ME method and show that we get the same answer.

According to the ME method the selected joint posterior $P(x, \theta)$ is that which maximizes the entropy,

$$S[p, q] = -\int dx d\theta\ p(x, \theta) \log \frac{p(x, \theta)}{q(x, \theta)} \ , \tag{6.95}$$

subject to the appropriate constraints. Note that the information in the data, eq.(6.93), represents an *infinite* number of constraints on the family $p(x, \theta)$: for each value of x there is one constraint and one Lagrange multiplier $\lambda(x)$. Maximizing S, (6.95), subject to (6.93) and normalization,

$$\delta \left\{ S + \alpha \left[\int dx d\theta\ p(x, \theta) - 1 \right] + \int dx\, \lambda(x) \left[\int d\theta\ p(x, \theta) - \delta(x - x') \right] \right\} = 0 \ , \tag{6.96}$$

yields the joint posterior,

$$P(x, \theta) = q(x, \theta)\, \frac{e^{\lambda(x)}}{Z} \ , \tag{6.97}$$

where Z is a normalization constant, and the multiplier $\lambda(x)$ is determined from (6.93),

$$\int d\theta\ q(x, \theta) \frac{e^{\lambda(x)}}{Z} = q(x) \frac{e^{\lambda(x)}}{Z} = \delta(x - x') \ , \tag{6.98}$$

so that the joint posterior is

$$P(x,\theta) = q(x,\theta)\frac{\delta(x-x')}{q(x)} = \delta(x-x')q(\theta|x) \ , \tag{6.99}$$

The corresponding marginal posterior probability $P(\theta)$ is

$$P(\theta) = \int dx\, P(\theta,x) = q(\theta|x') = q(\theta)\frac{q(x'|\theta)}{q(x')} \ , \tag{6.100}$$

which is recognized as Bayes' rule. Thus, Bayes' rule is derivable from and therefore consistent with the ME method.

To summarize: the prior $q(x,\theta) = q(x)q(\theta|x)$ is updated to the posterior $P(x,\theta) = P(x)P(\theta|x)$ where $P(x) = \delta(x-x')$ is fixed by the observed data while $P(\theta|x') = q(\theta|x')$ remains unchanged. Note that in accordance with the philosophy that drives the ME method *one only updates those aspects of one's beliefs for which corrective new evidence has been supplied.*

I conclude with a few simple examples that show how ME allows generalizations of Bayes' rule. The general background for these generalized Bayes problems is the familiar one: We want to make inferences about some variables θ on the basis of information about other variables x and of a relation between them.

Bayes updating with uncertain data

As before, the prior information consists of our prior beliefs about θ given by the distribution $q(\theta)$ and a likelihood function $q(x|\theta)$ so the joint prior $q(x,\theta)$ is known. But now the information about x is much more limited. The data is uncertain: x is not known. The marginal posterior $p(x)$ is no longer a sharp delta function but some other known distribution, $p(x) = P_D(x)$. This is still an infinite number of constraints

$$p(x) = \int d\theta\, p(\theta,x) = P_D(x) \ , \tag{6.101}$$

that are easily handled by ME. Maximizing S, (6.95), subject to (6.101) and normalization, leads to

$$P(x,\theta) = P_D(x)q(\theta|x) \ . \tag{6.102}$$

The corresponding marginal posterior,

$$P(\theta) = \int dx\, P_D(x)q(\theta|x) = q(\theta)\int dx\, P_D(x)\frac{q(x|\theta)}{q(x)} \ , \tag{6.103}$$

is known as Jeffrey's rule which we met earlier in section 2.9.

Bayes updating with information about x moments

Now we have even less information about the "data" x: the marginal distribution $p(x)$ is not known. All we know about $p(x)$ is an expected value

$$\langle f\rangle = \int dx\, p(x)f(x) = F \ . \tag{6.104}$$

Maximizing S, (6.95), subject to (6.104) and normalization,

$$\delta\left\{S+\alpha\left[\int dxd\theta\ p(x,\theta)-1\right]+\lambda\int dxd\theta\ p(x,\theta)f(x)-F\right\}=0\ , \tag{6.105}$$

yields the joint posterior,

$$P(x,\theta)=q(x,\theta)\,\frac{e^{\lambda f(x)}}{Z}\ , \tag{6.106}$$

where the normalization constant Z and the multiplier λ are obtained from

$$Z=\int dx\ q(x)e^{\lambda f(x)}\quad\text{and}\quad\frac{d\log Z}{d\lambda}=F\ . \tag{6.107}$$

The corresponding marginal posterior is

$$P(\theta)=q(\theta)\int dx\ \frac{e^{\lambda f(x)}}{Z}q(x|\theta)\ . \tag{6.108}$$

These two examples (6.103) and (6.108) are sufficiently intuitive that one could have written them down directly without deploying the full machinery of the ME method, but they do serve to illustrate the essential compatibility of Bayesian and Maximum Entropy methods. Next we consider a slightly less trivial example.

Updating with data and information about θ moments

Here we follow [Giffin Caticha 2007]. In addition to data about x we have additional information about θ in the form of a constraint on the expected value of some function $f(\theta)$,

$$\int dxd\theta\,P(x,\theta)f(\theta)=\langle f(\theta)\rangle=F\ . \tag{6.109}$$

In the standard Bayesian practice it is possible to impose constraint information at the level of the prior, but this information need not be preserved in the posterior. What we do here that differs from the standard Bayes' rule is that we can require that the constraint (6.109) be satisfied by the posterior distribution.

Maximizing the entropy (6.95) subject to normalization, the data constraint (6.93), and the moment constraint (6.109) yields the joint posterior,

$$P(x,\theta)=q(x,\theta)\frac{e^{\lambda(x)+\beta f(\theta)}}{z}\ , \tag{6.110}$$

where z is a normalization constant,

$$z=\int dxd\theta\,e^{\lambda(x)+\beta f(\theta)}q(x,\theta)\,. \tag{6.111}$$

The Lagrange multipliers $\lambda(x)$ are determined from the data constraint, (6.93),

$$\frac{e^{\lambda(x)}}{z}=\frac{\delta(x-x')}{Zq(x')}\quad\text{where}\quad Z(\beta,x')=\int d\theta\,e^{\beta f(\theta)}q(\theta|x')\ , \tag{6.112}$$

so that the joint posterior becomes

$$P(x,\theta) = \delta(x - x')q(\theta|x')\frac{e^{\beta f(\theta)}}{Z} \ . \tag{6.113}$$

The remaining Lagrange multiplier β is determined by imposing that the posterior $P(x,\theta)$ satisfy the constraint (6.109). This yields an implicit equation for β,

$$\frac{\partial \log Z}{\partial \beta} = F \ . \tag{6.114}$$

Note that since $Z = Z(\beta, x')$ the multiplier β will depend on the observed data x'. Finally, the new marginal distribution for θ is

$$P(\theta) = q(\theta|x')\frac{e^{\beta f(\theta)}}{Z} = q(\theta)\frac{q(x'|\theta)}{q(x')}\frac{e^{\beta f(\theta)}}{Z} \ . \tag{6.115}$$

For $\beta = 0$ (no moment constraint) we recover Bayes' rule. For $\beta \neq 0$ Bayes' rule is modified by a "canonical" exponential factor yielding an effective likelihood function.

Updating with uncertain data and an unknown likelihood

The following example [Caticha 2010] derives and generalizes Zellner's Bayesian Method of Moments [Zellner 1997]. Usually the relation between x and θ is given by a known likelihood function $q(x|\theta)$ but suppose this relation is not known. This is the case when the joint prior is so ignorant that information about x tells us nothing about θ and vise versa; such a prior treats x and θ as statistically independent, $q(x,\theta) = q(x)q(\theta)$. Since we have no likelihood function the information about the relation between θ and the data x must be supplied elsewhere. One possibility is through a constraint. Suppose that in addition to normalization and the uncertain data constraint, eq.(6.101), we also know that the expected value over θ of a function $f(x,\theta)$ is

$$\langle f \rangle_x = \int d\theta \, p(\theta|x) f(x,\theta) = F(x) \ . \tag{6.116}$$

We seek a posterior $P(x,\theta)$ that maximizes (6.95). Introducing Lagrange multipliers α, $\lambda(x)$, and $\gamma(x)$,

$$\begin{aligned} 0 = \delta \Big\{ S + \alpha \Big[\int dx d\theta \ p(x,\theta) - 1 \Big] + \int dx \, \lambda(x) \Big[\int d\theta \ p(x,\theta) - P_D(x) \Big] \\ + \int dx \, \gamma(x) \Big[\int d\theta \ p(x,\theta) f(x,\theta) - P_D(x) F(x) \Big] \Big\} \ , \end{aligned} \tag{6.117}$$

the variation over $p(x,\theta)$ yields

$$P(x,\theta) = \frac{1}{\zeta} q(x) q(\theta) \, e^{\lambda(x) + \gamma(x) f(x,\theta)} \ , \tag{6.118}$$

where ζ is a normalization constant. The multiplier $\lambda(x)$ is determined from (6.101),

$$P(x) = \int d\theta \, P(\theta, x) = \frac{1}{\zeta} q(x) e^{\lambda(x)} \int d\theta \, q(\theta) \, e^{\gamma(x) f(x,\theta)} = P_D(x) \tag{6.119}$$

then,

$$P(x,\theta) = P_D(x)\frac{q(\theta)\, e^{\gamma(x)f(x,\theta)}}{\int d\theta'\, q(\theta')\, e^{\gamma(x)f(x,\theta')}} \tag{6.120}$$

so that

$$P(\theta|x) = \frac{P(x,\theta)}{P(x)} = \frac{q(\theta)\, e^{\gamma(x)f(x,\theta)}}{Z(x)} \quad \text{with} \quad Z(x) = \int d\theta'\, q(\theta')\, e^{\gamma(x)f(x,\theta')} \tag{6.121}$$

The multiplier $\gamma(x)$ is determined from (6.116)

$$\frac{1}{Z(x)}\frac{\partial Z(x)}{\partial \gamma(x)} = F(x)\ . \tag{6.122}$$

The corresponding marginal posterior is

$$P(\theta) = \int dx\, P_D(x) P(\theta|x) = q(\theta)\int dx\, P_D(x)\frac{e^{\gamma(x)f(x,\theta)}}{Z(x)}\ . \tag{6.123}$$

In the limit when the data are sharply determined $P_D(x) = \delta(x - x')$ the posterior takes the form of Bayes theorem,

$$P(\theta) = q(\theta)\,\frac{e^{\gamma(x')f(x',\theta)}}{Z(x')}\ , \tag{6.124}$$

where up to a normalization factor $e^{\gamma(x')f(x',\theta)}$ plays the role of the likelihood and the normalization constant Z plays the role of the evidence.

In conclusion, these examples demonstrate that the method of maximum entropy can fully reproduce the results obtained by the standard Bayesian methods and allows us to extend them to situations that lie beyond their reach such as when the likelihood function is not known.

6.7 Commuting and non-commuting constraints

The ME method allows one to process information in the form of constraints. When we are confronted with several constraints we must be particularly cautious. In what order should they be processed? Or should they be processed together? The answer depends on the problem at hand. (Here we follow [Giffin Caticha 2007].)

We refer to constraints as *commuting* when it makes no difference whether they are handled simultaneously or sequentially. The most common example is that of Bayesian updating on the basis of data collected in multiple experiments: for the purpose of inferring θ it is well-known that the order in which the observed data $x' = \{x'_1, x'_2, \ldots\}$ is processed does not matter. (See section 2.9.3) The proof that ME is completely compatible with Bayes' rule implies that data constraints implemented through δ functions, as in (6.93), commute. It is useful to see how this comes about.

When an experiment is repeated it is common to refer to the value of x in the first experiment and the value of x in the second experiment. This is a dangerous practice because it obscures the fact that we are actually talking about *two* separate variables. We do not deal with a single x but with a composite $x = (x_1, x_2)$ and the relevant space is $\mathcal{X}_1 \times \mathcal{X}_2 \times \Theta$. After the first experiment yields the value x_1', represented by the constraint $\mathcal{C}_1 : P(x_1) = \delta(x_1 - x_1')$, we can perform a second experiment that yields x_2' and is represented by a second constraint $\mathcal{C}_2 : P(x_2) = \delta(x_2 - x_2')$. These constraints $\mathcal{C}_1$ and $\mathcal{C}_2$ commute because they refer to *different* variables x_1 and x_2. An experiment, once performed and its outcome observed, cannot be *un-performed*; its result cannot be *un-observed* by a second experiment. Thus, imposing the second constraint does not imply a revision of the first.

In general constraints need not commute and when this is the case the order in which they are processed is critical. For example, suppose the prior is q and we receive information in the form of a constraint, $\mathcal{C}_1$. To update we maximize the entropy $S[p, q]$ subject to $\mathcal{C}_1$ leading to the posterior P_1 as shown in Figure 6.1. Next we receive a second piece of information described by the constraint $\mathcal{C}_2$. At this point we can proceed in essentially two different ways:

(a) Sequential updating. Having processed $\mathcal{C}_1$, we use P_1 as the current prior and maximize $S[p, P_1]$ subject to the new constraint $\mathcal{C}_2$. This leads us to the posterior P_a.

(b) Simultaneous updating. Use the original prior q and maximize $S[p, q]$ subject to both constraints C_1 and C_2 simultaneously. This leads to the posterior P_b.[8]

To decide which path (a) or (b) is appropriate we must be clear about how the ME method handles constraints. The ME machinery interprets a constraint such as $\mathcal{C}_1$ in a very mechanical way: all distributions satisfying $\mathcal{C}_1$ are in principle allowed and all distributions violating $\mathcal{C}_1$ are ruled out.

Updating to a posterior P_1 consists precisely in revising those aspects of the prior q that disagree with the new constraint $\mathcal{C}_1$. However, there is nothing final about the distribution P_1. It is just the best we can do in our current state of knowledge and we fully expect that future information may require us to revise it further. Indeed, when new information $\mathcal{C}_2$ is received we must reconsider whether the original $\mathcal{C}_1$ remains valid or not. Are *all* distributions satisfying the new $\mathcal{C}_2$ really allowed, even those that violate $\mathcal{C}_1$? If this is the case then the new $\mathcal{C}_2$ takes over and we update from P_1 to P_a. The constraint $\mathcal{C}_1$ may still retain some lingering effect on the posterior P_a through P_1, but in general $\mathcal{C}_1$ has now become obsolete.

Alternatively, we may decide that the old constraint $\mathcal{C}_1$ retains its validity. The new $\mathcal{C}_2$ is not meant to revise $\mathcal{C}_1$ but to provide an additional refinement of the family of allowed posteriors. If this is the case, then the constraint that

[8] At first sight it might appear that there exists a third possibility of simultaneous updating: (c) use P_1 as the current prior and maximize $S[p, P_1]$ subject to both constraints C_1 and C_2 simultaneously. Fortunately, and this is a valuable check for the consistency of the ME method, it is easy to show that case (c) is equivalent to case (b). Whether we update from q or from P_1 the selected posterior is P_b.

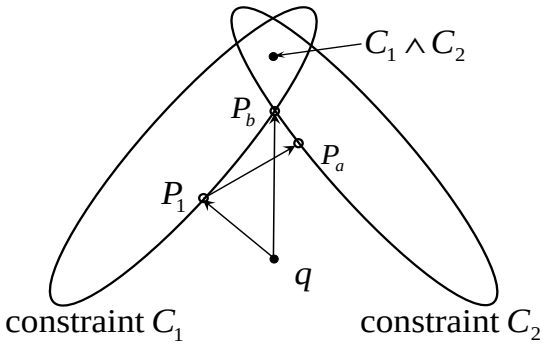


Figure 6.1: Illustrating the difference between processing two constraints C_1 and C_2 sequentially ($q \to P_1 \to P_a$) and simultaneously ($q \to P_b$ or $q \to P_1 \to P_b$).

correctly reflects the new information is not $\mathcal{C}_2$ but the more restrictive $\mathcal{C}_1 \wedge \mathcal{C}_2$. The two constraints should be processed simultaneously to arrive at the correct posterior P_b.

To summarize: sequential updating is appropriate when old constraints become obsolete and are superseded by new information; simultaneous updating is appropriate when old constraints remain valid. The two cases refer to different states of information and therefore *we expect* that they will result in different inferences. These comments are meant to underscore the importance of understanding what information is and how it is processed by the ME method; failure to do so will lead to errors that do not reflect a shortcoming of the ME method but rather a misapplication of it.

6.8 Conclusion

Any Bayesian account of the notion of information cannot ignore the fact that Bayesians are concerned with the beliefs of rational agents. The relation between information and beliefs must be clearly spelled out. The definition we have proposed — that information is that which constrains rational beliefs and therefore forces the agent to change its mind — is convenient for two reasons. First, the information/belief relation very explicit, and second, the definition is ideally suited for quantitative manipulation using the ME method.

Dealing with uncertainty requires that one solve two problems. First, one must represent a state of partial knowledge as a consistent web of interconnected beliefs. The instrument to do it is probability. Second, when new information becomes available the beliefs must be updated. The instrument for this is relative entropy. It is the only candidate for an updating method that is of universal applicability; that recognizes the value of prior information; and that recognizes the privileged role played by the notion of independence in science. The resulting general method — the ME method — can handle arbitrary priors and arbitrary constraints; it includes MaxEnt and Bayes' rule as special cases; and it provides its own criterion to assess the extent that non maximum-entropy distributions are ruled out.

The design of the ME method is essentially complete. However, the fact that ME operates by ranking distributions according to preference immediately raises questions about why should distributions that lie very close to the entropy maximum be totally ruled out; and if not ruled out completely, to what extent should they contribute to the inference. Do they make any difference? To what extent can we even distinguish similar distributions? Such matters discussed in the next two chapters significantly extend the utility of the ME method as a framework for inference.

Chapter 7

Information Geometry

A main concern of any theory of inference is to pick a probability distribution from a set of candidates and this immediately raises many questions. What if we had picked a neighboring distribution? What difference would it make? What makes two distributions similar? To what extent can we distinguish one distribution from another? Are there quantitative measures of distinguishability? The goal of this chapter is to address such questions by introducing methods of geometry. More specifically the goal will be to introduce a notion of "distance" between two probability distributions.

A parametric family of probability distributions — distributions $p_\theta(x)$ labeled by parameters $\theta = (\theta^1 \dots \theta^n)$ — forms a statistical manifold, namely, a space in which each point, labeled by coordinates θ, represents a probability distribution $p_\theta(x)$. Generic manifolds do not come with a pre-installed notion of distance; such additional structure has to be purchased separately in the form of a metric (that is, the metric tensor). Statistical manifolds are, however, an exception. One of the main goals of this chapter is to show that statistical manifolds possess a uniquely natural notion of distance — the so-called information metric. And the metric does not merely come as some optional software that is conveniently pre-installed; it is part of the hardware and it is inevitable. Geometry is intrinsic to statistical manifolds.

We will not develop the subject in all its possibilities — for a more extensive treatment see [Amari 1985, Amari Nagaoka 2000] — but we do wish to emphasize one specific result. Having a notion of distance means we have a notion of volume and this in turn implies that there is a unique and objective notion of a prior distribution that is uniform over the space of parameters — equal volumes are assigned equal prior probabilities. Whether such uniform distributions are maximally non-informative, or whether they define ignorance, or whether they reflect the actual prior beliefs of any rational agent, are all important issues but they are quite beside the specific point that we want to make, namely, that they are uniform — and this is not a matter of subjective judgment but of objective mathematical proof.

The distance $d\ell$ between two neighboring points θ and $\theta + d\theta$ is given by

Pythagoras' theorem, which written in terms of a metric tensor g_{ab}, is[1]

$$d\ell^2 = g_{ab} d\theta^a d\theta^b \ . \tag{7.1}$$

The singular importance of the metric tensor g_{ab} derives from a theorem due to N. Čencov that states that the metric g_{ab} on the manifold of probability distributions is essentially unique: up to an overall scale factor there is only one metric that takes into account the fact that these are not distances between simple structureless dots but between probability distributions [Cencov 1981].

7.1 Examples of statistical manifolds

An n-dimensional manifold $\mathcal{M}$ is a smooth, possibly curved, space that is locally like $\mathbb{R}^n$. What this means is that one can set up a coordinate frame (that is a map $\mathcal{M} \to \mathbb{R}^n$) so that each point $\theta \in \mathcal{M}$ is identified or labelled by its coordinates, $\theta = (\theta^1 \dots \theta^n)$.

A statistical manifold is a manifold in which each point θ represents a probability distribution $p_\theta(x)$. Thus, a statistical manifold is a family of probability distributions $p_\theta(x)$ that depend on n parameters $\theta = (\theta^1 \dots \theta^n)$; the distributions are labelled by the parameters θ. As we shall later see a very convenient notation is $p_\theta(x) = p(x|\theta)$.

The multinomial distributions are given by

$$p(\{n_i\}|\theta) = \frac{N!}{n_1! n_2! \dots n_m!} (\theta^1)^{n_1} (\theta^2)^{n_2} \dots (\theta^m)^{n_m} \ , \tag{7.2}$$

where $\theta = (\theta^1, \theta^2 \dots \theta^m)$, $N = \sum_{i=1}^m n_i$ and $\sum_{i=1}^m \theta^i = 1$. They form a statistical manifold of dimension $(m-1)$ called a simplex, S_{m-1}. The parameters $\theta = (\theta^1, \theta^2 \dots \theta^m)$ are a convenient choice of coordinates.

The multivariate Gaussian distributions with means μ^a, $a = 1 \dots n$, and variance σ^2,

$$p(x|\mu, \sigma) = \frac{1}{(2\pi\sigma^2)} \exp -\frac{1}{2\sigma^2} \sum_{a=1}^{n} (x^a - \mu^a)^2 \ , \tag{7.3}$$

form an $(n+1)$-dimensional statistical manifold with coordinates $\theta = (\mu^1, \dots, \mu^n, \sigma^2)$.

The canonical distributions, eq.(4.76),

$$p(i|F) = \frac{1}{Z} e^{-\lambda_k f_i^k} \ , \tag{7.4}$$

are derived by maximizing the Shannon entropy $S[p]$ subject to constraints on the expected values of n functions $f_i^k = f^k(x_i)$ labeled by superscripts $k = 1, 2, \dots n$,

$$\left\langle f^k \right\rangle = \sum_i p_i f_i^k = F^k \ . \tag{7.5}$$

[1] The use of superscripts rather than subscripts for the indices labelling coordinates is a standard and very convenient notational convention in differential geometry. We adopt the standard Einstein convention of summing over repeated indices whenever one appears as a superscript and the other as a subscript, that is, $g_{ab} f^{ab} = \sum_a \sum_b g_{ab} f^{ab}$.

They form an n-dimensional statistical manifold. As coordinates we can either use the expected values $F = (F^1 \dots F^n)$ or, equivalently, the Lagrange multipliers, $\lambda = (\lambda_1 \dots \lambda_n)$.

7.2 Vectors in curved spaces

In this section and the next we briefly review some basic notions of differential geometry. First we will be concerned with the notion of a *vector*. The treatment is not meant to be rigorous; the goal is to give an intuitive discussion of the basic ideas and to establish the notation.

Vectors as displacements

Perhaps the most primitive notion of a vector is associated to a displacement in space and is visualized as an arrow; other vectors such as velocities and accelerations are defined in terms of such displacements and from these one can elaborate further to define forces, fields and so on.

This notion of vector as a displacement is useful in flat spaces but it does not work in a curved space — a bent arrow is not useful. The appropriate generalization follows from the observation that smoothly curved spaces are locally flat — by which one means that within a sufficiently small region deviations from flatness can be neglected. Therefore one can imagine defining infinitesimal displacement vectors, say $d\vec{x}$, and defining other finite vectors, say $\vec{v}$, by appropriate multiplication by a suitably large number, say $\vec{v} = d\vec{x}/dt$.

Defined in this way vectors can no longer be thought of as "contained" in the original curved space. The set of vectors that one can define at any given point x of the curved manifold constitute the tangent space T_x at x. An immediate implication is that vectors at different locations x_1 and x_2 belong to different spaces, T_{x_1} and T_{x_2}, and therefore they cannot be added or subtracted or even compared without additional structure — we must provide criteria that "connect" the two tangent spaces and stipulate which vector at T_{x_2} corresponds to or is the same as a given vector in T_{x_1}.

The consequences for physics are enormous. Concepts that are familiar and basic in flat spaces cannot be defined in the curved spaces of general relativity. For example, the natural definition of the relative velocity of two distant objects involves taking the difference of their two velocities but the operation of subtraction is not available to us. The same happens with other concepts such as the total momentum or the total energy of an extended system of particles; the individual momenta live in different tangent spaces and there is no unique way to add them.

An objection to using displacements as the starting point to the construction of vectors is that it relies on our intuition of a curved space as being embedded in a flat space of larger dimension. Such larger spaces need not exist. So, while visualizing curved spaces in this way can be useful, it is also a good idea to pursue alternative approaches to the subject.

Vectors as tangents to curves

An alternative approach is to focus our attention directly on the velocities rather than on the displacements. Introduce a coordinate frame so that a point x has coordinates x^a with $a = 1 \dots n$. A parametrized curve $x(\lambda)$ is represented by n functions $x^a(\lambda)$ and the vector $\vec{v}$ tangent to the curve $x(\lambda)$ at the point labeled by λ is represented by the n-tuple of real numbers $\{dx^a/d\lambda\}$,

$$\vec{v} \sim (\frac{dx^1}{d\lambda} \dots \frac{dx^n}{d\lambda}) \ . \tag{7.6}$$

A shortcoming of such a definition is that it relies on the notion of coordinates. But coordinates depend on the choice of frame and therefore so do the components of vectors. When we change to new coordinates,

$$x^{a\prime} = f^{a\prime}(x^1 \dots x^n) \ , \tag{7.7}$$

the components of the vector in the new frame change accordingly; they are given by the chain rule,

$$\vec{v} \sim (\frac{dx^{1\prime}}{d\lambda} \dots \frac{dx^{n\prime}}{d\lambda}) \quad \text{where} \quad \frac{dx^{a\prime}}{d\lambda} = \frac{\partial x^{a\prime}}{\partial x^b}\frac{dx^b}{d\lambda} \tag{7.8}$$

The notation we adopt is quite standard: when an index appears repeated as an upper index and as a lower index it is understood that it is meant to be summed over; a derivative with respect to an upper index counts as a lower index, $\partial/\partial x^a = \partial_a$, and vice versa; and the primed frame is indicated by priming the indices, not the quantity — that is $x^{a\prime}$ and not x'^a.

In this approach a vector at the point x is defined as an n-tuple of real numbers $(v^1 \dots v^n)$ that under a change of coordinate frame transform according to

$$v^{a\prime} = X^{a\prime}_b v^b \quad \text{where} \quad X^{a\prime}_b = \frac{\partial x^{a\prime}}{\partial x^b} \ . \tag{7.9}$$

In other words, the different representations relative to different frames refer to the same vector; the vector itself is independent of the choice of coordinates.

The coordinate independence can be made more explicit by introducing the notion of a basis. A coordinate frame singles out n special vectors $\{\vec{e}_a\}$ defined so that the b component of $\vec{e}_a$ is

$$e^b_a = \delta^b_a \ . \tag{7.10}$$

More explicitly,

$$\vec{e}_1 \sim (1, 0 \dots 0), \ \vec{e}_2 \sim (0, 1, 0 \dots 0), \ \dots, \ \vec{e}_1 \sim (0, 0 \dots 1) \ . \tag{7.11}$$

Any vector $\vec{v}$ can be expressed in terms of the basis vectors,

$$\vec{v} = v^a \, \vec{e}_a \ . \tag{7.12}$$

The basis vectors in the primed frame are defined in the same way

$$e^{a\prime}_{b\prime} = \delta^{a\prime}_{b\prime} \ . \tag{7.13}$$

so that using eq.(7.9) we have

$$\vec{v} = v^{a\prime}\,\vec{e}_{a\prime} = X^{a\prime}_{b} v^{b}\,\vec{e}_{a\prime} = v^{b}\ \vec{e}_{b} \ , \tag{7.14}$$

where

$$\vec{e}_{b} = X^{a\prime}_{b}\ \vec{e}_{a\prime} \quad \text{or, equivalently} \quad \vec{e}_{a\prime} = X^{b}_{a\prime}\ \vec{e}_{b} \ . \tag{7.15}$$

Eq.(7.14) shows that while the components v^a and the basis vectors $\vec{e}_a$ both depend on the frame, the vector $\vec{v}$ itself is invariant, and eq.(7.15) shows that the invariance follows from the fact that components and basis vectors transform according to inverse matrices. Explicitly,

$$X^{a\prime}_{b} X^{b}_{c\prime} = \frac{\partial x^{a\prime}}{\partial x^{b}} \frac{\partial x^{b}}{\partial x^{c\prime}} = \frac{\partial x^{a\prime}}{\partial x^{c\prime}} = \delta^{a\prime}_{c\prime} \ . \tag{7.16}$$

Vectors as directional derivatives

There is yet a third way to introduce vectors. Let $\phi(x)$ be a scalar function and consider its derivative along the parametrized curve $x(\lambda)$ is given by the chain rule,

$$\frac{d\phi}{d\lambda} = \frac{\partial \phi}{\partial x^{a}} \frac{dx^{a}}{d\lambda} = \frac{\partial \phi}{\partial x^{a}} v^{a} \tag{7.17}$$

Note that $d\phi/d\lambda$ is independent of the choice of coordinates. Indeed, using the chain rule

$$\frac{d\phi}{d\lambda} = \frac{\partial \phi}{\partial x^{a}} v^{a} = \frac{\partial \phi}{\partial x^{a\prime}} \frac{\partial x^{a\prime}}{\partial x^{a}} v^{a} = \frac{\partial \phi}{\partial x^{a\prime}} v^{a\prime} \ . \tag{7.18}$$

But the function ϕ is arbitrary, therefore,

$$\frac{d}{d\lambda} = v^{a} \frac{\partial}{\partial x^{a}} \ . \tag{7.19}$$

Note further that the partial derivatives $\partial/\partial x^a$ transform exactly as the basis vectors, eq.(7.15)

$$\frac{\partial}{\partial x^{a}} = \frac{\partial x^{a\prime}}{\partial x^{a}} \frac{\partial}{\partial x^{a\prime}} = X^{a\prime}_{a} \frac{\partial}{\partial x^{a\prime}} \ , \tag{7.20}$$

so that there is a $1:1$ correspondence between the directional derivative $d/d\lambda$ and the vector $\vec{v}$ that is tangent to the curve $x(\lambda)$. Since mathematical objects are defined purely through the formal rules of manipulation it is legitimate to ignore the distinction between the two concepts and set

$$\vec{v} = \frac{d}{d\lambda} \quad \text{and} \quad \vec{e}_{a} = \frac{\partial}{\partial x^{a}} \tag{7.21}$$

Indeed, the "vector" $\partial/\partial x^a$ is the derivative along those curves parametrized by $\mu = x^a$ and defined by keeping the other coordinates constant, $x^b(\mu) = x^b(\mu_0)$ for $b \neq a$.

From a physical perspective, however, beyond the rules for formal manipulation mathematical objects are also assigned a meaning, an interpretation, and it is not clear that the two concepts, the derivative $d/d\lambda$ and the tangent vector $\vec{v}$, should be considered as physically identical. Nevertheless, we can still take advantage of the isomorphism to calculate using one picture while providing interpretations using the other.

7.3 Distance and volume in curved spaces

The notion of a distance between two points is not intrinsic to the manifold; it has to be supplied as an additional structure — the metric tensor. Statistical manifolds are a remarkable exception.

The basic intuition derives from the previous observation that curved spaces are locally flat: at any point x, within a sufficiently small region curvature effects can be neglected. The idea then is rather simple: within a very small region in the vicinity of a point x we can always transform from the original coordinates x^a to new coordinates $\hat{x}^\alpha = f^\alpha(x^1 \dots x^n)$ that we *declare* as being locally Cartesian (here denoted $\hat{x}$ and with Greek superscripts). An infinitesimal displacement is given by

$$d\hat{x}^\alpha = X^\alpha_a \, dx^a \quad \text{where} \quad X^\alpha_a = \frac{\partial \hat{x}^\alpha}{\partial x^a} \tag{7.22}$$

and the corresponding infinitesimal distance can be computed using Pythagoras theorem,

$$d\ell^2 = \delta_{\alpha\beta} d\hat{x}^\alpha d\hat{x}^\beta \ . \tag{7.23}$$

Changing back to the original frame

$$d\ell^2 = \delta_{\alpha\beta} d\hat{x}^\alpha d\hat{x}^\beta = \delta_{\alpha\beta} X^\alpha_a \, X^\beta_b \, dx^a dx^b \ . \tag{7.24}$$

Defining the quantities

$$g_{ab} \stackrel{\text{def}}{=} \delta_{\alpha\beta} X^\alpha_a \, X^\beta_b \ , \tag{7.25}$$

we can write the infinitesimal Pythagoras theorem in the generic initial frame as

$$d\ell^2 = g_{ab} dx^a dx^b \ . \tag{7.26}$$

The quantities g_{ab} are the components of the metric tensor. One can easily check that under a coordinate transformation g_{ab} transform according to

$$g_{ab} = X^{a\prime}_a X^{b\prime}_a g_{a\prime b\prime} \ , \tag{7.27}$$

so that the infinitesimal distance $d\ell$ is independent of the coordinate frame.

To find the finite length between two points along a curve $x(\lambda)$ one integrates along the curve,

$$\ell = \int_{\lambda_1}^{\lambda_2} d\ell = \int_{\lambda_1}^{\lambda_2} \left(g_{ab} \frac{dx^a}{d\lambda} \frac{dx^b}{d\lambda} \right)^{1/2} d\lambda \ . \tag{7.28}$$

Having decided on a measure of distance we can now also measure angles, areas, volumes and all sorts of other geometrical quantities. To find an expression for the n-dimensional volume element dV_n we use the same trick as before: Transform to locally Cartesian coordinates so that the volume element is simply given by the product

$$dV_n = d\hat{x}^1 d\hat{x}^2 \dots d\hat{x}^n \ , \tag{7.29}$$

and then transform back to the original coordinates x^a using eq.(7.22),

$$dV_n = \left|\frac{\partial \hat{x}}{\partial x}\right| dx^1 dx^2 \dots dx^n = |\det X^\alpha_a|\, d^n x \ . \tag{7.30}$$

This is the volume we seek written in terms of the coordinates x^a but we still have to calculate the Jacobian of the transformation, $|\partial\hat{x}/\partial x| = |\det X^\alpha_a|$. The transformation of the metric from its Euclidean form $\delta_{\alpha\beta}$ to g_{ab}, eq.(7.25), is the product of three matrices. Taking the determinant we get

$$g \stackrel{\text{def}}{=} \det(g_{ab}) = [\det X^\alpha_a]^2 \ , \tag{7.31}$$

so that

$$|\det (X^\alpha_a)| = g^{1/2} \ . \tag{7.32}$$

We have succeeded in expressing the volume element in terms of the metric $g_{ab}(x)$ in the original coordinates x^a. The answer is

$$dV_n = g^{1/2}(x) d^n x \ . \tag{7.33}$$

The volume of any extended region on the manifold is

$$V_n = \int dV_n = \int g^{1/2}(x) d^n x \ . \tag{7.34}$$

Example: A uniform distribution over such a curved manifold is one which assigns equal probabilities to equal volumes,

$$p(x) d^n x \propto g^{1/2}(x) d^n x \ . \tag{7.35}$$

Example: These ideas are also useful in flat spaces when dealing with non-Cartesian coordinates. The distance element of three-dimensional flat space in spherical coordinates (r, θ, ϕ) is

$$d\ell^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2 \ , \tag{7.36}$$

and the corresponding metric tensor is

$$(g_{ab}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2\theta \end{pmatrix} \ . \tag{7.37}$$

The volume element is the familiar expression

$$dV = g^{1/2} dr d\theta d\phi = r^2 \sin\theta \, dr d\theta d\phi \ . \tag{7.38}$$

7.4 Derivations of the information metric

The distance $d\ell$ between two neighboring distributions $p(x|\theta)$ and $p(x|\theta + d\theta)$ or, equivalently, between the two points θ and $\theta + d\theta$, is given by the metric g_{ab}. Our goal is to compute the tensor g_{ab} corresponding to $p(x|\theta)$. We give several different derivations because this serves to illuminate the meaning of the information metric, its interpretation, and ultimately, how it is to be used.

At this point a word of caution (and encouragement) might be called for. Of course it is possible to be confronted with sufficiently singular families of distributions that are not smooth manifolds and studying their geometry might seem a hopeless enterprise. Should we up on geometry? No. The fact that statistical manifolds can have complicated geometries does not detract from the value of the methods of information geometry any more than the existence of surfaces with rugged geometries detracts from the general value of geometry itself.

7.4.1 Derivation from distinguishability

We seek a quantitative measure of the extent that two distributions $p(x|\theta)$ and $p(x|\theta+d\theta)$ can be distinguished. The following argument is intuitively appealing [Rao 1945]. The advantage of this approach is the emphasis on interpretation — the metric measures distinguishability — the disadvantage is that the argument does not address the issue of uniqueness of the metric.

Consider the relative difference,

$$\Delta = \frac{p(x|\theta + d\theta) - p(x|\theta)}{p(x|\theta)} = \frac{\partial \log p(x|\theta)}{\partial \theta^a}\, d\theta^a. \tag{7.39}$$

The expected value of the relative difference, $\langle \Delta \rangle$, might seem a good candidate, but it does not work because it vanishes identically,

$$\langle \Delta \rangle = \int dx\, p(x|\theta)\, \frac{\partial \log p(x|\theta)}{\partial \theta^a}\, d\theta^a = d\theta^a\, \frac{\partial}{\partial \theta^a} \int dx\, p(x|\theta) = 0. \tag{7.40}$$

(Depending on the problem by the symbol $\int dx$ we mean to represent either discrete sums or integrals over one or more dimensions.) However, the variance does not vanish,

$$d\ell^2 = \langle \Delta^2 \rangle = \int dx\, p(x|\theta)\, \frac{\partial \log p(x|\theta)}{\partial \theta^a}\, \frac{\partial \log p(x|\theta)}{\partial \theta^b}\, d\theta^a d\theta^b\ . \tag{7.41}$$

This is the measure of distinguishability we seek; a small value of $d\ell^2$ means that the relative difference Δ is small and the points θ and $\theta + d\theta$ are difficult to distinguish. It suggests introducing the matrix g_{ab}

$$g_{ab} \stackrel{\text{def}}{=} \int dx\, p(x|\theta)\, \frac{\partial \log p(x|\theta)}{\partial \theta^a}\, \frac{\partial \log p(x|\theta)}{\partial \theta^b} \tag{7.42}$$

called the Fisher information *matrix* [Fisher 25], so that

$$d\ell^2 = g_{ab}\, d\theta^a d\theta^b \ . \tag{7.43}$$

Up to now no notion of distance has been introduced. Normally one says that the reason it is difficult to distinguish two points in say, the three dimensional space we seem to inhabit, is that they happen to be too close together. It is very tempting to invert this intuition and assert that the two points θ and $\theta+d\theta$ must be very close together because they are difficult to distinguish. Furthermore, note that being a variance, $d\ell^2 = \langle \Delta^2 \rangle$, the quantity $d\ell^2$ is positive and vanishes only when $d\theta$ vanishes. Thus it is natural to interpret g_{ab} as the metric tensor of a Riemannian space [Rao 1945]. This is the *information metric*. The recognition by Rao that g_{ab} is a metric in the space of probability distributions gave rise to the subject of information geometry [Amari 1985], namely, the application of geometrical methods to problems in inference and in information theory.

Other useful expressions for the information metric are

$$\begin{aligned} g_{ab} &= 4 \int dx \ \frac{\partial p^{1/2}(x|\theta)}{\partial \theta^a} \ \frac{\partial p^{1/2}(x|\theta)}{\partial \theta^b} \\ &= -4 \int dx \ p^{1/2}(x|\theta) \ \frac{\partial^2 p^{1/2}(x|\theta)}{\partial \theta^a \partial \theta^b} \ , \end{aligned} \tag{7.44}$$

and

$$g_{ab} = -\int dx \ p(x|\theta) \ \frac{\partial^2 \log p(x|\theta)}{\partial \theta^a \partial \theta^b} = -\langle \frac{\partial^2 \log p_\theta}{\partial \theta^a \partial \theta^b} \rangle \ . \tag{7.45}$$

The coordinates θ are quite arbitrary; one can freely relabel the points in the manifold. It is then easy to check that g_{ab} are the components of a tensor and that the distance $d\ell^2$ is an invariant, a scalar under coordinate transformations. Indeed, the transformation

$$\theta^{a\prime} = f^{a\prime}(\theta^1 \dots \theta^n) \tag{7.46}$$

leads to

$$d\theta^a = \frac{\partial \theta^a}{\partial \theta^{a\prime}} d\theta^{a\prime} \quad \text{and} \quad \frac{\partial}{\partial \theta^a} = \frac{\partial \theta^{a\prime}}{\partial \theta^a} \frac{\partial}{\partial \theta^{a\prime}} \tag{7.47}$$

so that, substituting into eq.(7.42),

$$g_{ab} = \frac{\partial \theta^{a\prime}}{\partial \theta^a} \frac{\partial \theta^{b\prime}}{\partial \theta^b} g_{a\prime b\prime} \tag{7.48}$$

7.4.2 Derivation from a Euclidean metric

Consider a discrete variable $i = 1 \dots m$. The possible probability distributions of i can be labelled by the probability values themselves: a probability distribution can be specified by a point p with coordinates $(p^1 \dots p^m)$. The corresponding statistical manifold is the simplex $\mathcal{S}_{m-1} = \{p = (p^1 \dots p^m) : \sum_i p^i = 1\}$.

Next we change to new coordinates $\psi^i = \left(p^i\right)^{1/2}$. In these new coordinates the equation for the simplex $\mathcal{S}_{m-1}$ — the normalization condition — reads $\sum \left(\psi^i\right)^2 = 1$, which we recognize as the equation of an $(m-1)$-sphere embedded in an m-dimensional Euclidean space $\mathbb{R}^m$, *provided* the ψ^i are interpreted as Cartesian coordinates. This suggests that we assign the simplest possible metric: the distance between the distribution $p(i|\psi)$ and its neighbor $p(i|\psi + d\psi)$ is the Euclidean distance in $\mathbb{R}^m$,

$$d\ell^2 = \sum_i \left(d\psi^i\right)^2 = \delta_{ij} d\psi^i d\psi^j \ . \tag{7.49}$$

Distances between more distant distributions are merely angles defined on the surface of the unit sphere $\mathcal{S}_{m-1}$. To express $d\ell^2$ in terms of the original coordinates $p^i = \left(p^i\right)^2$ substitute

$$d\psi^i = \frac{1}{2} \frac{dp^i}{(p^i)^{1/2}} \tag{7.50}$$

to get

$$d\ell^2 = \frac{1}{4} \sum_i \frac{\left(dp^i\right)^2}{p^i} = \frac{1}{4} \frac{\delta_{ij}}{p^i} dp^i dp^j \ . \tag{7.51}$$

Except for an overall constant this is the same information metric (7.43) we defined earlier! Indeed, consider an n-dimensional subspace ($n \leq m-1$) of the simplex $\mathcal{S}_{m-1}$ defined by $\psi^i = \psi^i(\theta^1, \ldots, \theta^n)$. The parameters θ^a, $i = 1 \ldots n$, can be used as coordinates on the subspace. The Euclidean metric on $\mathbb{R}^m$ induces a metric on the subspace. The distance between $p(i|\theta)$ and $p(i|\theta + d\theta)$ is

$$\begin{aligned} d\ell^2 &= \delta_{ij} d\psi^i d\psi^j = \delta_{ij} \frac{\partial \psi^i}{\partial \theta^a} d\theta^a \frac{\partial \psi^j}{\partial \theta^b} d\theta^b \\ &= \frac{1}{4} \sum_i p^i \frac{\partial \log p^i}{\partial \theta^a} \frac{\partial \log p^i}{\partial \theta^b} d\theta^a d\theta^b \ , \end{aligned} \tag{7.52}$$

which (except for the factor 1/4) we recognize as the discrete version of (7.42) and (7.43). This interesting result does not constitute a "derivation." There is a priori no reason why the coordinates ψ^i should be singled out as special and attributed a Euclidean metric. But perhaps it helps to lift the veil of mystery that might otherwise surround the strange expression (7.42).

7.4.3 Derivation from asymptotic inference

We have two very similar probability distributions. Which one do we prefer? To decide we collect data in N independent trials. Then the question becomes: To what extent does the data support one distribution over the other? This is a typical inference problem. To be explicit consider multinomial distributions specified by $p = (p_1 \ldots p_m)$. (Here it is slightly more convenient to revert to the notation where indices appear as subscripts.) Suppose the data consists of

the numbers $(n_1 \dots n_m)$ where n_i is the number of times that outcome i occurs. The corresponding frequencies are

$$f_i = \frac{n_i}{N} \quad \text{with} \quad \sum_{i=1}^{m} n_i = N \ . \tag{7.53}$$

The probability of a particular frequency distribution $f = (f_1 \dots f_n)$ is

$$P_N\,(f|p) = \frac{N!}{n_1! \dots n_m!} p_1^{n_1} \dots p_m^{n_m} \ . \tag{7.54}$$

For sufficiently large N and n_i we can use Stirling's approximation [see eq.(6.88)], to get

$$P_N\,(f|p) \approx C_N (\prod_i f_i)^{-1/2} \exp(NS[f,p]) \tag{7.55}$$

where C_N is a normalization constant and $S[f,p]$ is the entropy given by eq.(6.11). The Gibbs inequality $S[f,p] \leq 0$, eq.(4.23), shows that for large N the probability $P_N\,(f|p)$ shows an exceedingly sharp peak. The most likely f_i is p_i — this is the weak law of large numbers.

Now we come to the inference: the values of p best supported by the data f are inferred from Bayes rule,

$$P_N\,(p|f) \propto \exp(NS[f,p]) \ , \tag{7.56}$$

where we have used the fact that for large N the exponential e^{NS} dominates both the prior and the pre-factor $(\prod f_i)^{-1/2}$. For large N the data f_i supports the value $p_i = f_i$. But the distribution $P_N\,(p|f)$ is not infinitely sharp; there is some uncertainty. Distributions with parameters $p'_i = f_i + \delta p_i$ can only be distinguished from $p_i = f_i$ provided δp_i lies outside a small region of uncertainty defined roughly by

$$NS[p,p'] = NS[p,p+\delta p] \approx -1 \tag{7.57}$$

so that the probability $P_N\,(p'|f)$ is down by e^{-1} from the maximum. Expanding to second order,

$$S[p,p+\delta p] = -\sum_i p_i \log \frac{p_i}{p_i + \delta p_i} \approx -\frac{1}{2} \sum_i \frac{(\delta p_i)^2}{p_i} \tag{7.58}$$

Thus, the nearest that two neighboring points p and $p + \delta p$ can be and still be distinguishable in N trials is such that

$$\frac{N}{2} \sum_i \frac{(\delta p_i)^2}{p_i} \approx 1 \ . \tag{7.59}$$

As N increases the resolution δp with which we can distinguish neighboring distributions improves roughly as $1/\sqrt{N}$.

We can now define a "statistical" distance on the simplex $\mathcal{S}_{m-1}$. The argument below is given in [Wootters 1981]; see also [Balasubramanian 1997]. We

define the length of a curve between two given points by counting the number of distinguishable points that one can fit along the curve,

$$\text{Statistical length} = \ell = \lim_{N\to\infty} \frac{1}{\sqrt{N/2}} \begin{bmatrix} \text{number of distinguishable (in } N \text{ trials)} \\ \text{distributions that fit along the curve} \end{bmatrix} \tag{7.60}$$

Since the number of distinguishable points grows as $\sqrt{N}$ it is convenient to introduce a factor $1/\sqrt{N}$ so that there is a finite limit as $N \to \infty$. The factor $\sqrt{2}$ is purely conventional.

Remark: It is not actually necessary to include the $\sqrt{2/N}$ factor; this leads to a notion of statistical length ℓ_N defined on the space of N-trial multinomials. (See section 7.6.)

More explicitly, let the curve $p = p(\lambda)$ be parametrized by λ. The separation $\delta\lambda$ between two neighboring distributions that can barely be resolved in N trials is

$$\frac{N}{2}\sum_i \frac{1}{p_i}(\frac{dp_i}{d\lambda})^2\delta\lambda^2 \approx 1 \quad \text{or} \quad \delta\lambda \approx \left(\frac{N}{2}\sum_i \frac{1}{p_i}(\frac{dp_i}{d\lambda})^2\right)^{-1/2} . \tag{7.61}$$

The number of distinguishable distributions within the interval $d\lambda$ is $d\lambda/\delta\lambda$ and the corresponding statistical length, eq.(7.60), is

$$d\ell = \left(\sum_i \frac{1}{p_i}(\frac{dp_i}{d\lambda})^2\right)^{1/2} d\lambda = \left(\sum_i \frac{(dp_i)^2}{p_i}\right)^{1/2} \tag{7.62}$$

The length of the curve from λ_0 to λ_1 is

$$\ell = \int_{\lambda_0}^{\lambda_1} d\ell \quad \text{where} \quad d\ell^2 = \sum_i \frac{(dp_i)^2}{p_i} . \tag{7.63}$$

Thus, the width of the fluctuations is the unit used to define a local measure of "distance". To the extent that fluctuations are intrinsic to statistical problems the geometry they induce is unavoidably hardwired into the statistical manifolds. The statistical or distinguishability length differs from a possible Euclidean distance $d\ell_E^2 = \sum(dp_i)^2$ because the fluctuations are not uniform over the space $\mathcal{S}_{m-1}$ which affects our ability to resolve neighboring points.

Equation (7.63) agrees the previous definitions of the information metric. Consider the n-dimensional subspace ($n \leq m-1$) of the simplex $\mathcal{S}_{m-1}$ defined by $p_i = p_i(\theta^1, \ldots, \theta^n)$. The distance between two neighboring distributions in this subspace, $p(i|\theta)$ and $p(i|\theta + d\theta)$, is

$$d\ell^2 = \sum_{i=1}^m \frac{(\delta p_i)^2}{p_i} = \sum_{i,j=1}^m \frac{1}{p_i}\frac{\partial p_i}{\partial\theta^a}d\theta^a \frac{\partial p_j}{\partial\theta^b}d\theta^b = g_{ab}d\theta^a d\theta^b \tag{7.64}$$

where

$$g_{ab} = \sum_{i=1}^m p_i \frac{\partial \log p_i}{\partial\theta^a}\frac{\partial \log p_i}{\partial\theta^b} , \tag{7.65}$$

which is the discrete version of (7.42).

7.4.4 Derivation from relative entropy

The relation we uncovered above between the information metric and entropy, eq.(7.58), is not restricted to multinomials; it is quite general. Consider the entropy of one distribution $p(x|\theta')$ relative to another $p(x|\theta)$,

$$S(\theta', \theta) = -\int dx\, p(x|\theta') \log \frac{p(x|\theta')}{p(x|\theta)} \ . \tag{7.66}$$

We study how this entropy varies when $\theta' = \theta + d\theta$ is in the close vicinity of a given θ. As we had seen in section 4.2 – recall the Gibbs inequality $S(\theta', \theta) \leq 0$ with equality if and only if $\theta' = \theta$ — the entropy $S(\theta', \theta)$ attains an absolute maximum at $\theta' = \theta$. Therefore, the first nonvanishing term in the Taylor expansion about θ is second order in $d\theta$

$$S(\theta + d\theta, \theta) = \frac{1}{2} \frac{\partial^2 S(\theta', \theta)}{\partial \theta'^a \partial \theta'^b}\bigg|_{\theta'=\theta} d\theta^a d\theta^b + \ldots \leq 0 \ , \tag{7.67}$$

which suggests defining the distance $d\ell$ by

$$S(\theta + d\theta, \theta) = -\frac{1}{2} d\ell^2 \ . \tag{7.68}$$

The second derivative is

$$\begin{aligned} -\frac{\partial^2 S(\theta', \theta)}{\partial \theta'^a \partial \theta'^b} &= \frac{\partial}{\partial \theta'^a} \int dx \, [\log \frac{p(x|\theta')}{p(x|\theta)} + 1] \frac{\partial p(x|\theta')}{\partial \theta'^b} \\ &= \int dx \, [\frac{\partial \log p(x|\theta')}{\partial \theta'^a} \frac{\partial p(x|\theta')}{\partial \theta'^b} + \, [\log \frac{p(x|\theta')}{p(x|\theta)} + 1] \frac{\partial^2 p(x|\theta')}{\partial \theta'^a \partial \theta'^b}] \ , \end{aligned}$$

so that, evaluating at $\theta' = \theta$, gives the desired result,

$$-\frac{\partial S(\theta', \theta)}{\partial \theta'^a \partial \theta'^b}\bigg|_{\theta'=\theta} = \int dx\, p(x|\theta) \frac{\partial \log p(x|\theta)}{\partial \theta^a} \frac{\partial \log p(x|\theta)}{\partial \theta^b} = g_{ab} \ . \tag{7.69}$$

7.5 Uniqueness of the information metric

The most remarkable fact about the information metric is that it is essentially unique: except for an constant scale factor it is the only Riemannian metric that adequately takes into account the nature of the points of a statistical manifold, namely, that these points are not "structureless", that they are probability distributions. This theorem was first proved by N. Čencov within the framework of category theory [Cencov 1981]. The proof I give below follows the treatment in [Campbell 1986].

Markov embeddings

Consider a discrete variable $i = 1, \ldots, n$ and let the probability of any particular i be $\Pr(i) = p^i$. In practice the limitation to discrete variables is not very

serious because we can choose an n large enough to approximate a continuous distribution to any desirable degree. However, it is possible to imagine situations where the continuum limit is tricky — here we avoid such situations.

The set of numbers $p = (p^1, \ldots p^n)$ can be used as coordinates to define a point on a statistical manifold. In this particular case the manifold is the $(n-1)$-dimensional simplex $S_{n-1} = \{p = (p^1, \ldots p^n) : \sum p^i = 1\}$. The argument is, however, considerably simplified by considering instead the n-dimensional space of non-normalized distributions. This is the positive "octant" $R_n^+ = \{p = (p^1, \ldots p^n) : p^i > 0\}$. The boundary is explicitly avoided so that R_n^+ is an open set.

Next we introduce the notion of Markov mappings. The set of values of i can be grouped or partitioned into M disjoint subsets with $2 \leq M \leq n$. Let $A = 1 \ldots M$ label the subsets, then the probability of the Ath subset is

$$\Pr(A) = P^A = \textstyle\sum_{i \in A} p^i \ . \tag{7.70}$$

The space of these coarser probability distributions is the simplex $S_{M-1} = \{P = (P^1, \ldots P^M) : \sum P^A = 1\}$. The corresponding space of non-normalized distributions if the positive octant $R_M^+ = \{P = (P^1, \ldots P^M) : P^A > 0\}$.

Thus, the act of partitioning (or grouping, or coarse graining) has produced a mapping $G : R_n^+ \to R_M^+$ with $P = G(p)$ given by eq.(7.70). This is a many-to-one map; it has no inverse. An interesting map that runs in the opposite direction $R_M^+ \to R_n^+$ can be defined by introducing conditional probabilities. Let

$$q_A^i = \begin{cases} \Pr(i|A) & \text{if} \quad i \in A \\ 0 & \text{if} \quad i \notin A \end{cases} \tag{7.71}$$

with

$$\sum_i q_A^i = \sum_{i \in A} \Pr(i|A) = 1 \ . \tag{7.72}$$

Thus, to each choice of the set of numbers $\{q_A^i\}$ we can associate a one-to-one map $Q : R_m^+ \to R_n^+$ with $p = Q(P)$ defined by

$$p^i = q_A^i P^A \ . \tag{7.73}$$

This is a sum over A but since $q_A^i = 0$ unless $i \in A$ only one term in the sum is non-vanishing and the map is clearly invertible. These Q maps, called *Markov mappings*, define an embedding of R_M^+ into R_n^+. Markov mappings preserve normalization,

$$\textstyle\sum_i p^i = \sum_i q_A^i P^A = \sum_A P^A \ . \tag{7.74}$$

Example: A coarse graining map G for the case of $R_3^+ \to R_2^+$ is

$$G(p^1, p^2, p^3) = (p^1, p^2 + p^3) = (P^1, P^2) \ . \tag{7.75}$$

One Markov map Q running in the opposite direction $R_2^+ \to R_3^+$ could be

$$Q(P^1, P^2) = (P^1, \frac{1}{3}P^2, \frac{2}{3}P^3) = (p^1, p^2, p^3) \ . \tag{7.76}$$

This particular map is defined by setting all $q^i_A = 0$ except $q^1_1 = 1$, $q^2_3 = 1/3$, and $q^3_2 = 2/3$.

Example: We can use binomial distributions to analyze the act of tossing a coin (the outcomes are either heads or tails) or, equally well, the act of throwing a die (provided we only care about outcomes that are either even or odd). This amounts to embedding the space of coin distributions (which are binomials, R^+_M with $M = 2$) as a subspace of the space of die distributions (which are multinomials, R^+_n with $n = 6$).

To minimize confusion between the two spaces we will use lower case symbols to refer to the original larger space R^+_n and upper case symbols to refer to the coarse grained embedded space R^+_M.

Having introduced the notion of Markov embeddings we can now state the basic idea behind Campbell's argument. For a fixed choice of $\{q^i_A\}$, that is for a fixed Markov map Q, the distribution P and its image $p = Q(P)$ represent exactly the same information. In other words, whether we talk about heads/tails outcomes in coins or about even/odd outcomes in dice, binomials are binomials. Therefore the map Q is invertible. The Markov image $Q(S_{M-1})$ of the simplex S_{M-1} in S_{n-1} is statistically "identical" to S_{M-1},

$$Q(S_{M-1}) = S_{M-1} \tag{7.77}$$

in the sense that it is just as easy or just as difficult to distinguish the two distributions P and $P + dP$ as it is to distinguish their images $p = Q(P)$ and $p + dp = Q(P + dP)$. Whatever geometrical relations are assigned to distributions in S_{M-1}, exactly the same geometrical relations should be assigned to the corresponding distributions in $Q(S_{M-1})$. Thus Markov mappings are not just embeddings, they are congruent embeddings; distances between distributions in R^+_m should match the distances between the corresponding images in R^+_n.

Our goal is to find the Riemannian metrics that are invariant under Markov mappings. It is easy to see why imposing such invariance is extremely restrictive: The fact that distances computed in R^+_M must agree with distances computed in subspaces of R^+_n introduces a constraint on the allowed metric tensors; but we can always embed R^+_M in spaces of larger and larger dimension which imposes an infinite number of constraints. It could very well have happened that no Riemannian metric survives such restrictive conditions; it is quite remarkable that some do and it is even more remarkable that (up to an uninteresting scale factor) the surviving Riemannian metric is unique.

The invariance of the metric is conveniently expressed as an invariance of the inner product: inner products among vectors in R^+_M should coincide with the inner products among the corresponding images in R^+_n. Let vectors tangent to R^+_M be denoted by

$$\vec{V} = V^A \frac{\partial}{\partial P^A} = V^A \vec{E}_A \ , \tag{7.78}$$

where $\{\bar{E}_A\}$ is a coordinate basis. The inner product of two such vectors is

$$\langle \vec{V}, \vec{U} \rangle_M = g^{(M)}_{AB} V^A U^B \tag{7.79}$$

where the metric is

$$g^{(M)}_{AB} \stackrel{\text{def}}{=} \langle \vec{E}_A, \vec{E}_B \rangle_M \ . \tag{7.80}$$

Similarly, vectors tangent to R^+_n are denoted by

$$\vec{v} = v^i \frac{\partial}{\partial p^i} = v^i \vec{e}_i \ , \tag{7.81}$$

and the inner product of two such vectors is

$$\langle \vec{v}, \vec{u} \rangle_n = g^{(n)}_{ij} v^i u^j \tag{7.82}$$

where

$$g^{(n)}_{ij} \stackrel{\text{def}}{=} \langle \vec{e}_i, \vec{e}_j \rangle_n \ . \tag{7.83}$$

The images of vectors $\vec{V}$ tangent to R^+_m under Q are obtained from eq.(7.73)

$$Q_* \frac{\partial}{\partial P^A} = \frac{\partial p^i}{\partial P^A} \frac{\partial}{\partial p^i} = q^i_A \frac{\partial}{\partial p^i} \quad \text{or} \quad Q_* \vec{E}_A = q^i_A \vec{e}_i \ , \tag{7.84}$$

which leads to

$$Q_* \vec{V} = \vec{v} \quad \text{with} \quad v^i = q^i_a V^a \ . \tag{7.85}$$

Therefore, the invariance or isometry we want to impose is expressed as

$$\langle \vec{V}, \vec{U} \rangle_M = \langle Q_* \vec{V}, Q_* \vec{U} \rangle_n = \langle \vec{v}, \vec{u} \rangle_n \ . \tag{7.86}$$

The Theorem

Let $\langle \ , \ \rangle_M$ be the inner product on R^+_M for any $M \in \{2, 3, \ldots\}$. The theorem states that the metric is invariant under Markov embeddings if and only if

$$g^{(M)}_{AB} = \langle \bar{e}_A, \bar{e}_B \rangle_M = \alpha(|P|) + |P| \beta(|P|) \frac{\delta_{AB}}{P^A} \ , \tag{7.87}$$

where $|P| \stackrel{\text{def}}{=} \sum_A P^A$, and α and β are smooth (C^∞) functions with $\beta > 0$ and $\alpha + \beta > 0$. The proof is given in the next section.

The metric above refers to the positive cone R^+_M but ultimately we are interested in the metric induced on the simplex S_{M-1} defined by $|P| = 1$. In order to find the induced metric we first show that vectors that are tangent to the simplex S_{M-1} are such that

$$|V| \stackrel{\text{def}}{=} \sum_A V^A = 0 \ . \tag{7.88}$$

Indeed, consider the derivative of any function $f = f(|P|)$ defined on R^+_M along the direction defined by $\vec{V}$,

$$0 = V^A \frac{\partial f}{\partial P^A} = V^A \frac{df}{d|P|} \frac{\partial |P|}{\partial P^A} = |V| \frac{df}{d|P|} \ , \tag{7.89}$$

where we used $\partial|P|/\partial P^A = 1$. Therefore $|V| = 0$.

Next consider the inner product of any two vectors $\vec{V}$ and $\vec{U}$,

$$\begin{aligned}\langle \vec{V}, \vec{U} \rangle_M &= \sum_{AB} V^A U^B \left(\alpha(|P|) + |P|\beta(|P|) \frac{\delta_{AB}}{P^A} \right) \\ &= \alpha(|P|)|V||U| + |P|\beta(|P|) \sum_A \frac{V^A U^A}{P^A} \ . \end{aligned} \tag{7.90}$$

For vectors tangent to the simplex S_{M-1} this simplifies to

$$\langle \vec{V}, \vec{U} \rangle_M = \beta(1) \sum_A \frac{V^A U^A}{P^A} \ . \tag{7.91}$$

Therefore the choice of the function $\alpha(|P|)$ is irrelevant and the corresponding metric on S_{M-1} is determined up to a multiplicative constant $\beta(1) = \beta$

$$g_{AB} = \beta \frac{\delta_{AB}}{P^A} \ . \tag{7.92}$$

It is trivial to check that this is exactly the information metric that was heuristically suggested earlier. Indeed, changing to new coordinates $\psi^A = (P^A)^{1/2}$ reduces the metric to its Euclidean form

$$d\ell^2 = g_{AB}\ dP^A dP^B = 4\beta\ \delta_{AB}\ d\psi^A d\psi^B, \tag{7.93}$$

which, incidentally, shows that when the function $\alpha(|P|) = 0$, the geometry of the space R_M^+ is Euclidean. Furthermore, transforming to a generic coordinate frame $\psi^A = \psi^A(\theta^1, \ldots, \theta^M)$ yields

$$d\ell^2 = g_{ab} d\theta^a d\theta^b \tag{7.94}$$

with

$$g_{ab} = \beta \sum_A P^A \frac{\partial \log P^A}{\partial \theta^a} \frac{\partial \log P^A}{\partial \theta^b} \ . \tag{7.95}$$

The Proof

The strategy is to consider special cases of Markov embeddings to determine what kind of constraints the impose on the metric. First we consider the consequences of invariance under the family of Markov maps Q' that embed R_M^+ into itself. In this case $n = M$ and the action of Q' is to permute coordinates. A simple example in which just two coordinates are permuted is

$$\begin{aligned}(p^1, \ldots p^a, \ldots p^b, \ldots p^M) &= Q'(P^1, \ldots P^M) \\ &= (P^1, \ldots P^b, \ldots P^a, \ldots P^m) \end{aligned} \tag{7.96}$$

The required q_A^i are

$$q_A^a = \delta_A^b \,, \quad q_A^b = \delta_A^a \quad \text{and} \quad q_A^i = \delta_A^i \quad \text{for} \quad A \neq a, b \ , \tag{7.97}$$

so that eq.(7.84), $Q'_*\vec{E}_A = q^i_A \vec{e}_i$, gives

$$Q'_*\vec{E}_a = \vec{e}_b\,, \quad Q'_*\vec{E}_b = \vec{e}_a \quad \text{and} \quad Q'_*\vec{E}_A = \vec{e}_A \quad \text{for} \quad A \neq a, b\,. \tag{7.98}$$

The invariance

$$\langle \vec{E}_A, \vec{E}_B \rangle_M = \langle Q'_*\vec{E}_A, Q'_*\vec{E}_B \rangle_M \tag{7.99}$$

yields,

$$g^{(M)}_{aA}(P) = g^{(M)}_{bA}(p) \quad \text{and} \quad g^{(M)}_{bA}(P) = g^{(M)}_{aA}(p) \quad \text{for} \quad A \neq a, b \tag{7.100}$$

$$g^{(M)}_{aa}(P) = g^{(M)}_{bb}(p) \quad \text{and} \quad g^{(M)}_{bb}(P) = g^{(M)}_{aa}(p) \tag{7.101}$$

$$g^{(M)}_{AB}(P) = g^{(M)}_{AB}(p) \quad \text{for} \quad A, B \neq a, b\,.$$

These conditions are useful for points along the line along the center of R^+_M, $P^1 = P^2 = \ldots = P^M$. Let $P_c = (c/M, \ldots, c/M)$ with $c = |P_c|$; we have $p_c = Q'(P_c) = P_c$. Using eqs.(7.100) and (7.101) for all choices of the pairs (a, b) implies

$$\begin{aligned} g^{(M)}_{AA}(P_c) &= F_M(c) \\ g^{(M)}_{AB}(P_c) &= G_M(c) \quad \text{for} \quad A \neq B\,, \end{aligned} \tag{7.102}$$

where F_M and G_M are some unspecified functions.

Next we consider the family of Markov maps $Q'' : R^+_M \to R^+_{kM}$ with $k \geq 2$

$$\begin{aligned} Q''(P^1, \ldots P^M) &= (p^1, \ldots p^{kM}) \\ &= (\underbrace{\frac{P^1}{k}, \ldots \frac{P^1}{k}}_{k \text{ times}}, \underbrace{\frac{P^2}{k}, \ldots \frac{P^2}{k}}_{k \text{ times}}, \ldots, \underbrace{\frac{P^M}{k}, \ldots \frac{P^M}{k}}_{k \text{ times}})\,. \end{aligned} \tag{7.103}$$

Q'' is implemented by choosing

$$q^i_A = \begin{cases} 1/k & \text{if} \quad i \in \{k(A-1)+1, \ldots kA\} \\ 0 & \text{if} \quad i \notin \{k(A-1)+1, \ldots kA\} \end{cases} \tag{7.104}$$

Under the action of Q'' vectors are transformed according to eq.(7.84),

$$Q''_*\vec{E}_A = q^i_A \vec{e}_i = \frac{1}{k}\left(\vec{e}_{k(A-1)+1} + \ldots + \vec{e}_{kA}\right) \tag{7.105}$$

so that the invariance

$$\langle \vec{E}_A, \vec{E}_B \rangle_M = \langle Q''_*\vec{E}_A, Q''_*\vec{E}_B \rangle_{kM} \tag{7.106}$$

yields,

$$g^{(M)}_{AB}(P) = \frac{1}{k^2} \sum_{i,\ j=k(A-1)+1}^{kA} g^{(kM)}_{ij}(p)\,. \tag{7.107}$$

Along the center lines, $P_c = (c/M, \ldots, c/M)$ and $p_c = (c/kM, \ldots, c/kM)$, equations (7.102) and (7.107) give

$$F_M(c) = \frac{1}{k} F_{kM}(c) + \frac{k-1}{k} G_{kM}(c) \tag{7.108}$$

and

$$G_M(c) = G_{kM}(c) \ . \tag{7.109}$$

But this holds for all values of M and k, therefore $G_M(c) = \alpha(c)$ where α is a function independent of M. Furthermore, eq.(7.108) can be rewritten as

$$\frac{1}{M} \left[F_M(c) - \alpha(c) \right] = \frac{1}{kM} \left[F_{kM}(c) - \alpha(c) \right] = \beta(c) \ , \tag{7.110}$$

where $\beta(c)$ is a function independent of the integer M. Indeed, for any two integers M_1 and M_2 we have

$$\frac{1}{M_1} \left[F_{M_1}(c) - \alpha(c) \right] = \frac{1}{M_1 M_2} \left[F_{M_1 M_2}(c) - \alpha(c) \right] = \frac{1}{M_2} \left[F_{M_2}(c) - \alpha(c) \right] \ . \tag{7.111}$$

Therefore,

$$F_M(c) = \alpha(c) + M\beta(c) \ , \tag{7.112}$$

and for points along the center line,

$$g^{(M)}_{AB}(P_c) = \alpha(c) + M\beta(c)\delta_{AB} \ . \tag{7.113}$$

So far the invariance under the special Markov embeddings Q' and Q'' has allowed us to find the metric of R^+_M for arbitrary M but only along the center line $P = P_c$ for any $c > 0$. To find the metric $g^{(M)}_{AB}(P)$ at any arbitrary $P \in R^+_M$ we show that it is possible to cleverly choose the embedding $Q''' : R^+_M \to R^+_n$ so that the image of P can be brought arbitrarily close to the center line of R^+_n, $Q'''(P) \approx p_c$, where the metric is known. Indeed, consider the embeddings $Q''' : R^+_M \to R^+_n$ defined by

$$Q'''(P^1, \ldots P^M) = (\underbrace{\frac{P^1}{k_1}, \ldots \frac{P^1}{k_1}}_{k_1 \text{ times}}, \underbrace{\frac{P^2}{k_2}, \ldots \frac{P^2}{k_2}}_{k_2 \text{ times}}, \ldots, \underbrace{\frac{P^M}{k_M}, \ldots \frac{P^M}{k_M}}_{k_m \text{ times}}) \ . \tag{7.114}$$

Q''' is implemented by choosing

$$q^i_A = \begin{cases} 1/k_A & \text{if} \quad i \in \{(k_1 + \ldots + k_{A-1} + 1), (k_1 + \ldots + k_{A-1} + 2) \ldots, (k_1 + \ldots + k_A)\} \\ 0 & \text{if} \quad i \notin \{(k_1 + \ldots + k_{A-1} + 1), (k_1 + \ldots + k_{A-1} + 2) \ldots, (k_1 + \ldots + k_A)\} \end{cases} \tag{7.115}$$

Next note that any point P in R^+_M can be arbitrarily well approximated by points of the "rational" form

$$P = \left(\frac{ck_1}{n}, \frac{ck_2}{n}, \ldots \frac{ck_M}{n} \right) \ , \tag{7.116}$$

where the ks are positive integers and $\sum k_A = n$ and $|P| = c$. For these rational points the action of Q''' is

$$Q'''(P^1, \ldots P^M) = q^i_A P^A = \left(\frac{c}{n}, \frac{c}{n}, \ldots \frac{c}{n}\right) = p_c \tag{7.117}$$

which lies along the center line of R^+_n where the metric is known, eq.(7.113).

The action of Q''' on vectors, eq.(7.84), gives

$$Q'''_* \vec{E}_A = q^i_A \vec{e}_i = \frac{1}{k_A}\left(\vec{e}_{k_1+\ldots+k_{A-1}+1} + \ldots + \vec{e}_{k_1+\ldots+k_A}\right) \ . \tag{7.118}$$

Using eq.(7.113) the invariance

$$\langle \vec{E}_A, \vec{E}_B \rangle_M = \langle Q'''_* \vec{E}_A, Q'''_* \vec{E}_B \rangle_n \tag{7.119}$$

yields, for $A = B$,

$$\begin{aligned} g^{(M)}_{AA}(P) &= \frac{1}{(k_A)^2} \sum_{i,\ j=k_1+\ldots+k_{A-1}+1}^{k_1+\ldots+k_A} g^{(n)}_{ij}(p_c) \\ &= \frac{1}{(k_A)^2} \sum_{i,\ j=k_1+\ldots+k_{A-1}+1}^{k_1+\ldots+k_A} [\alpha(c) + n\beta(c)\delta_{ij}] \\ &= \frac{1}{(k_A)^2}\left[(k_A)^2 \alpha(c) + k_A n \beta(c)\right] \\ &= \alpha(c) + \frac{n}{k_A}\beta(c) = \alpha(c) + \frac{c\beta(c)}{P^A} \ , \end{aligned} \tag{7.120}$$

where we used eq.(7.116), $P^A = ck_A/n$. Similarly, for $A \neq B$,

$$g^{(M)}_{AB}(P) = \frac{1}{k_A k_B} \sum_{i=k_1+\ldots+k_{A-1}+1}^{k_1+\ldots+k_A} \ \sum_{j=k_1+\ldots+k_{B-1}+1}^{k_1+\ldots+k_B} g^{(n)}_{ij}(p_c) \tag{7.121}$$

$$= \frac{1}{k_A k_B} k_A k_B \alpha(c) = \alpha(c) \ . \tag{7.122}$$

Therefore,

$$g^{(M)}_{AB} = \langle \vec{E}_A, \vec{E}_B \rangle_M = \alpha(c) + c\beta(c)\frac{\delta_{AB}}{P^A} \ , \tag{7.123}$$

with $c = |P|$. This almost concludes the proof.

The sign conditions on α and β follow from the positive-definiteness of inner products. Using eq.(7.90),

$$\langle \vec{V}, \vec{V} \rangle = \alpha |V|^2 + |P|\beta \sum_A \frac{\left(V^A\right)^2}{P^A}, \tag{7.124}$$

we see that for vectors with $|V| = 0$, $\langle \vec{V}, \vec{V} \rangle \geq 0$ implies that $\beta > 0$, while for vectors with $V^A = KP^A$, where K is any constant we have

$$\langle \vec{V}, \vec{V} \rangle = K^2 |P|^2 (\alpha + \beta) > 0 \Rightarrow \alpha + \beta > 0 \; . \tag{7.125}$$

Conversely, we show that if these sign conditions are satisfied then $\langle \vec{V}, \vec{V} \rangle \geq 0$ for all vectors. Using Cauchy's inequality,

$$\left(\sum_i x_i^2 \right) \left(\sum_i y_i^2 \right) \geq \left(\sum_i \| x_i y_i \| \right)^2 , \tag{7.126}$$

where $\|.\|$ denotes the modulus, we have

$$\left(\sum_A P^A \right) \left(\sum_B \frac{\left(V^B \right)^2}{P^B} \right) \geq \left(\sum_A \left\| V^A \right\| \right)^2 \geq \left(\sum_A V^A \right)^2 \; . \tag{7.127}$$

Therefore,

$$\langle \vec{V}, \vec{V} \rangle = \alpha |V|^2 + |P| \beta \sum_A \frac{\left(V^A \right)^2}{P^A} \geq |V|^2 (\alpha + \beta) \geq 0 \; , \tag{7.128}$$

with equality if and only if all $V^A = 0$.

We have just proved that for invariance under Markov embeddings it is necessary that the metrics be of the form (7.123). It remains to prove the converse, that this condition is sufficient. This is much easier. Indeed,

$$\begin{aligned} \langle Q_* \vec{E}_A, Q_* \vec{E}_B \rangle_n &= q_A^i q_B^j \langle \bar{e}_i, \bar{e}_j \rangle_n \\ &= \sum_{ij} q_A^i q_B^j \left[\alpha(|p|) + |p| \beta(|p|) \frac{\delta_{ij}}{p^i} \right] \; . \end{aligned} \tag{7.129}$$

But as noted earlier, Markov mappings $p^i = q_A^i P^A$ are such that $\sum_i q_A^i = 1$ and they preserve normalization $|P| = |p|$, therefore

$$\langle Q_* \vec{E}_A, Q_* \vec{E}_B \rangle_n = \alpha(|P|) + |P| \beta(|P|) \sum_i \frac{q_A^i q_B^i}{p^i} \; . \tag{7.130}$$

Furthermore, since $q_A^i = 0$ unless $i \in A$,

$$\sum_i \frac{q_A^i q_B^i}{p^i} = \delta_{AB} \sum_i \frac{q_A^i}{P^A} = \frac{\delta_{AB}}{P^A} \; . \tag{7.131}$$

which finally leads to

$$\langle Q_* \vec{E}_A, Q_* \vec{E}_B \rangle_n = \alpha(|P|) + |P| \beta(|P|) \frac{\delta_{AB}}{P^A} = \langle \bar{e}_A, \bar{e}_B \rangle_M \tag{7.132}$$

which concludes the proof.

7.6 The metric for some common distributions

Multinomial distributions

The statistical manifold of multinomials,

$$P_N\,(n|\theta) = \frac{N!}{n_1!\dots n_m!}\theta_1^{n_1}\dots\theta_m^{n_m}\ , \tag{7.133}$$

where

$$n = (n_1\dots n_m) \quad \text{with} \quad \sum_{i=1}^{m} n_i = N \quad \text{and} \quad \sum_{i=1}^{m}\theta_i = 1\,, \tag{7.134}$$

is the simplex $\mathcal{S}_{m-1}$. The metric is given by eq.(7.65),

$$g_{ij} = \sum_n P_N \frac{\partial \log P_N}{\partial\theta_i}\frac{\partial \log P_N}{\partial\theta_j} \quad \text{where} \quad 1 \le i,j \le m-1\ . \tag{7.135}$$

The result is

$$g_{ij} = \left\langle (\frac{n_i}{\theta_i} - \frac{n_m}{\theta_m})(\frac{n_j}{\theta_j} - \frac{n_m}{\theta_m}) \right\rangle\ , \tag{7.136}$$

which, on computing the various correlations, gives

$$g_{ij} = \frac{N}{\theta_i}\delta_{ij} + \frac{N}{\theta_m} \quad \text{where} \quad 1 \le i,j \le m-1\ . \tag{7.137}$$

A somewhat simpler expression can be obtained by extending the range of the indices to include $i,j = m$. This is done as follows. The distance $d\ell$ between neighboring distributions is

$$d\ell^2 = \sum_{i,j=1}^{m-1} (\frac{N}{\theta_i}\delta_{ij} + \frac{N}{\theta_m}) d\theta_i d\theta_j\ . \tag{7.138}$$

Using

$$\sum_{i=1}^{m}\theta_i = 1 \implies \sum_{i=1}^{m} d\theta_i = 0\ . \tag{7.139}$$

the second sum can be written as

$$\frac{N}{\theta_m}\sum_{i,j=1}^{m-1} d\theta_i \sum_{i,j=1}^{m-1} d\theta_j = \frac{N}{\theta_m}(d\theta_m)^2\ . \tag{7.140}$$

Therefore,

$$d\ell^2 = \sum_{i,j=1}^{m} g_{ij} d\theta_i d\theta_j \quad \text{with} \quad g_{ij} = \frac{N}{\theta_i}\delta_{ij}\ . \tag{7.141}$$

Remark: As we saw in the previous section, eq.(7.95), the information metric is defined up to an overall multiplicative factor. This arbitrariness amounts to a choice of units. We see here that the distance $d\ell$ between N-trial multinomials contains a factor $\sqrt{N}$. It is a matter of convention whether we decide to include

such factors or not — that is, whether we want to adopt the same length scale when discussing two different statistical manifolds such as $\mathcal{S}_{m-1}^{(N)}$ and $\mathcal{S}_{m-1}^{(N')}$.

A uniform distribution over the simplex $\mathcal{S}_{m-1}$ is one which assigns equal probabilities to equal volumes,

$$P(\theta)d^{m-1}\theta \propto g^{1/2}d^{m-1}\theta \quad \text{with} \quad g = \frac{N^{m-1}}{\theta_1\theta_2\dots\theta_m} \tag{7.142}$$

In the particular case of binomial distributions $m = 2$ with $\theta_1 = \theta$ and $\theta_2 = 1 - \theta$ the results above become

$$g = g_{11} = \frac{N}{\theta(1-\theta)} \tag{7.143}$$

so that the uniform distribution over θ (with $0 < \theta < 1$) is

$$P(\theta)d\theta \propto d\ell = [\frac{N}{\theta(1-\theta)}]^{1/2}d\theta \ . \tag{7.144}$$

Canonical distributions

Let z denote the microstates of a system (*e.g.*, points in phase space) and let $m(z)$ be the underlying measure (*e.g.*, a uniform density on phase space). The space of macrostates is a statistical manifold: each macrostate is a canonical distribution (see sections 4.9 and 5.4) obtained by maximizing entropy $S[p, m]$ subject to n constraints $\langle f^a \rangle = F^a$ for $a = 1 \dots n$, plus normalization,

$$p(z|F) = \frac{1}{Z(\lambda)}m(z)e^{-\lambda_a f^a(z)} \quad \text{where} \quad Z(\lambda) = \int dz\, m(z)e^{-\lambda_a f^a(z)} \ . \tag{7.145}$$

The set of numbers $F = (F^1 \dots F^n)$ determines one point $p(z|F)$ on the statistical manifold so we can use the F^a as coordinates.

First, here are some useful facts about canonical distributions. The Lagrange multipliers λ_a are implicitly determined by

$$\langle f^a \rangle = F^a = -\frac{\partial \log Z}{\partial \lambda_a} \, , \tag{7.146}$$

and it is straightforward to show that a further derivative with respect to λ_b yields the covariance matrix. Indeed,

$$-\frac{\partial F^a}{\partial \lambda_b} = \frac{\partial}{\partial \lambda_b}(\frac{1}{Z}\frac{\partial Z}{\partial \lambda_a}) = -\frac{1}{Z^2}\frac{\partial Z}{\partial \lambda_a}\frac{\partial Z}{\partial \lambda_b} + \frac{1}{Z}\frac{\partial^2 Z}{\partial \lambda_a \partial \lambda_b} \tag{7.147}$$

$$= -F^a F^b + \langle f^a f^b \rangle \ . \tag{7.148}$$

Therefore

$$C^{ab} \stackrel{\text{def}}{=} \langle (f^a - F^a)(f^b - F^b) \rangle = -\frac{\partial F^a}{\partial \lambda_b} \ . \tag{7.149}$$

Furthermore, using the chain rule

$$\delta_a^c = \frac{\partial \lambda_a}{\partial \lambda_c} = \frac{\partial \lambda_a}{\partial F^b} \frac{\partial F^b}{\partial \lambda_c} \ , \tag{7.150}$$

we see that the matrix

$$C_{ab} = -\frac{\partial \lambda_a}{\partial F^b} \tag{7.151}$$

is the inverse of the covariance matrix,

$$C_{ab} C^{bc} = \delta_a^c \ .$$

The information metric is

$$\begin{aligned} g_{ab} &= \int dz\, p(z|F)\, \frac{\partial \log p(z|F)}{\partial F^a}\, \frac{\partial \log p(z|F)}{\partial F^b} \\ &= \frac{\partial \lambda_c}{\partial F^a} \frac{\partial \lambda_d}{\partial F^b} \int dz\, p\, \frac{\partial \log p}{\partial \lambda_c}\, \frac{\partial \log p}{\partial \lambda_d} \ . \end{aligned} \tag{7.152}$$

Using eqs.(7.145) and (7.146),

$$\frac{\partial \log p(z|F)}{\partial \lambda_c} = F^c - f^c(z) \tag{7.153}$$

therefore,

$$g_{ab} = C_{ca} C_{db} C^{cd} \implies g_{ab} = C_{ab} \ , \tag{7.154}$$

so that the metric tensor g_{ab} is the inverse of the covariance matrix C^{ab}.

Instead of F^a we could use the Lagrange multipliers λ_a themselves as coordinates. Then the information metric is the covariance matrix,

$$g^{ab} = \int dz\, p(z|\lambda)\, \frac{\partial \log p(z|\lambda)}{\partial \lambda_a}\, \frac{\partial \log p(z|\lambda)}{\partial \lambda_b} = C^{ab} \ . \tag{7.155}$$

The distance $d\ell$ between neighboring distributions can then be written in either of two equivalent forms,

$$d\ell^2 = g_{ab} dF^a dF^b = g^{ab} d\lambda_a d\lambda_b \ . \tag{7.156}$$

The uniform distribution over the space of macrostates assigns equal probabilities to equal volumes,

$$P(F) d^n F \propto C^{-1/2} d^n F \quad \text{or} \quad P'(\lambda) d^n \lambda \propto C^{1/2} d^n \lambda \ , \tag{7.157}$$

where $C = \det C^{ab}$.

Gaussian distributions

Gaussian distributions are a special case of canonical distributions — they maximize entropy subject to constraints on mean values and correlations. Consider Gaussian distributions in D dimensions,

$$p(x|\mu, C) = \frac{c^{1/2}}{(2\pi)^{D/2}} \exp\left[-\frac{1}{2} C_{ij}(x^i - \mu^i)(x^j - \mu^j)\right] \ , \tag{7.158}$$

where $1 \leq i \leq D$, C_{ij} is the inverse of the correlation matrix, and $c = \det C_{ij}$. The mean values μ^i are D parameters μ^i, while the symmetric C_{ij} matrix is an additional $\frac{1}{2}D(D+1)$ parameters. Thus the dimension of the statistical manifold is $D + \frac{1}{2}D(D+1)$.

Calculating the information distance between $p(x|\mu, C)$ and $p(x|\mu + d\mu, C + dC)$ is a matter of keeping track of all the indices involved. Skipping all details, the result is

$$d\ell^2 = g_{ij} d\mu^i d\mu^j + g^{ij}_k dC_{ij} d\mu^k + g^{ij\,kl} dC_{ij} dC_{kl} \ , \tag{7.159}$$

where

$$g_{ij} = C_{ij} \ , \quad g^{ij}_k = 0 \ , \quad \text{and} \quad g^{ij\,kl} = \frac{1}{4}(C^{ik}C^{jl} + C^{il}C^{jk}) \ , \tag{7.160}$$

where C^{ik} is the correlation matrix, that is, $C^{ik}C_{kj} = \delta^i_j$. Therefore,

$$d\ell^2 = C_{ij} dx^i dx^j + \frac{1}{2} C^{ik} C^{jl} dC_{ij} dC_{kl} \ . \tag{7.161}$$

To conclude we consider a couple of special cases. For Gaussians that differ only in their means the information distance between $p(x|\mu, C)$ and $p(x|\mu + d\mu, C)$ is obtained setting $dC_{ij} = 0$, that is,

$$d\ell^2 = C_{ij} dx^i dx^j \ , \tag{7.162}$$

which is an instance of eq.(7.154).

Finally, for spherically symmetric Gaussians,

$$p(x|\mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left[-\frac{1}{2\sigma^2} \delta_{ij}(x^i - \mu^i)(x^j - \mu^j)\right] \ . \tag{7.163}$$

The covariance matrix and its inverse are both diagonal and proportional to the unit matrix,

$$C_{ij} = \frac{1}{\sigma^2}\delta_{ij} \ , \quad C^{ij} = \sigma^2 \delta^{ij} \ , \quad \text{and} \quad c = \sigma^{-2D} \ . \tag{7.164}$$

Using

$$dC_{ij} = d\frac{1}{\sigma^2}\delta_{ij} = -\frac{2\delta_{ij}}{\sigma^3} d\sigma \tag{7.165}$$

in eq.(7.161), the induced information metric is

$$d\ell^2 = \frac{1}{\sigma^2}\delta_{ij}d\mu^i d\mu^j + \frac{1}{2}\sigma^4\delta^{ik}\delta^{jl}\frac{2\delta_{ij}}{\sigma^3}d\sigma\frac{2\delta_{kl}}{\sigma^3}d\sigma \tag{7.166}$$

which, using

$$\delta^{ik}\delta^{jl}\delta_{ij}\delta_{kl} = \delta^k_j\delta^j_k = \delta^k_k = D \ , \tag{7.167}$$

simplifies to

$$d\ell^2 = \frac{\delta_{ij}}{\sigma^2}d\mu^i d\mu^j + \frac{2D}{\sigma^2}(d\sigma)^2 \ . \tag{7.168}$$

Chapter 8

Entropy IV: Entropic Inference

There is one last issue that must be addressed before one can claim that the design of the method of entropic inference is more or less complete. Higher entropy represents higher preference but there is nothing in the previous arguments to tell us by how much. Suppose the maximum of the entropy function is not particularly sharp, are we really confident that distributions with entropy close to the maximum are totally ruled out? We want a quantitative measure of the extent to which distributions with lower entropy are ruled out. Or, to phrase this question differently: We can rank probability distributions p relative to a prior q according to the relative entropy $S[p,q]$ but any monotonic function of the relative entropy will accomplish the same goal. Does twice the entropy represent twice the preference or four times as much? Can we quantify 'preference'? The discussion below follows [Caticha 2000].

8.1 Deviations from maximum entropy

The problem is to update from a prior $q(x)$ given information specified by certain constraints. The constraints specify a family of candidate distributions $p_\theta(x) = p(x|\theta)$ which can be conveniently labelled with a finite number of parameters θ^i, $i = 1 \ldots n$. Thus, the parameters θ are coordinates on the statistical manifold specified by the constraints. The distributions in this manifold are ranked according to their entropy $S[p_\theta, q] = S(\theta)$ and the chosen posterior is the distribution $p(x|\theta_0)$ that maximizes the entropy $S(\theta)$.

The question we now address concerns the extent to which $p(x|\theta_0)$ should be preferred over other distributions with lower entropy or, to put it differently: To what extent is it rational to believe that the selected value ought to be the entropy maximum θ_0 rather than any other value θ? This is a question about the probability $p(\theta)$ of various values of θ.

The original problem which led us to design the maximum entropy method

was to assign a probability to x; we now see that the full problem is to assign probabilities to both x and θ. We are concerned not just with $p(x)$ but rather with the joint distribution $p_J(x,\theta)$; the universe of discourse has been expanded from $\mathcal{X}$ (the space of xs) to the product space $\mathcal{X}\times\Theta$ (Θ is the space of parameters θ).

To determine the joint distribution $p_J(x,\theta)$ we make use of essentially the only method at our disposal — the ME method itself — but this requires that we address the standard two preliminary questions: first, what is the prior distribution, what do we know about x and θ before we receive information about the constraints? And second, what is this new information that constrains the allowed $p_J(x,\theta)$?

This first question is the more subtle one: when we know absolutely nothing about the θs we know neither their physical meaning nor whether there is any relation to the xs. A joint prior that reflects this lack of correlations is a product, $q_J(x,\theta) = q(x)\mu(\theta)$. We will assume that the prior over x is known — it is the same prior we had used when we updated from $q(x)$ to $p(x|\theta_0)$. But we are not totally ignorant about the θs: we know that they label points on some as yet unspecified statistical manifold Θ. Then there exists a natural measure of distance in the space Θ. It is given by the information metric g_{ij} introduced in the previous chapter and the corresponding volume elements are given by $g^{1/2}(\theta)d^n\theta$, where $g(\theta)$ is the determinant of the metric. The uniform prior for θ, which assigns equal probabilities to equal volumes, is proportional to $g^{1/2}(\theta)$ and therefore we choose $\mu(\theta) = g^{1/2}(\theta)$. Therefore, the joint prior is $q_J(x,\theta) = q(x)g^{1/2}(\theta)$.

Next we tackle the second question: what are the constraints on the allowed joint distributions $p_J(x,\theta)$? Consider the space of all joint distributions. To each choice of the functional form of $p(x|\theta)$ (for example, whether we talk about Gaussians, Boltzmann-Gibbs distributions, or something else) there corresponds a different subspace defined by distributions of the form $p_J(x,\theta) = p(\theta)p(x|\theta)$. The crucial constraint is that which specifies the subspace by specifying the particular functional form of $p(x|\theta)$. This defines the meaning to the θs and also fixes the prior $g^{1/2}(\theta)$ on the relevant subspace.

To select the preferred joint distribution $P(x,\theta)$ we maximize the joint entropy $\mathcal{S}[p_J,q_J]$ over all distributions of the form $p_J(x,\theta) = p(\theta)p(x|\theta)$ by varying with respect to $p(\theta)$ with $p(x|\theta)$ fixed. It is convenient to write the entropy as

$$\begin{aligned}\mathcal{S}[p_J,q_J] &= -\int dx\, d\theta\, p(\theta)p(x|\theta)\, \log\frac{p(\theta)p(x|\theta)}{g^{1/2}(\theta)q(x)} \\ &= S[p,g^{1/2}] + \int d\theta\, p(\theta)S(\theta),\end{aligned} \tag{8.1}$$

where

$$S[p,g^{1/2}] = -\int d\theta\, p(\theta)\log\frac{p(\theta)}{g^{1/2}(\theta)} \tag{8.2}$$

and

$$S(\theta) = -\int dx\, p(x|\theta)\log\frac{p(x|\theta)}{q(x)}. \tag{8.3}$$

The notation shows that $S[p, g^{1/2}]$ is a functional of $p(\theta)$ while $S(\theta)$ is a function of θ (it is also a functional of $p(x|\theta)$). Maximizing (8.1) with respect to variations $\delta p(\theta)$ such that $\int d\theta\, p(\theta) = 1$, yields

$$0 = \int d\theta \left(-\log \frac{p(\theta)}{g^{1/2}(\theta)} + S(\theta) + \log\zeta \right) \delta p(\theta)\,, \tag{8.4}$$

where the required Lagrange multiplier has been written as $1 - \log\zeta$. Therefore the probability that the value of θ should lie within the small volume $g^{1/2}(\theta)d^n\theta$ is

$$P(\theta)d^n\theta = \frac{1}{\zeta}\, e^{S(\theta)} g^{1/2}(\theta) d^n\theta \quad \text{with} \quad \zeta = \int d^n\theta\, g^{1/2}(\theta)\, e^{S(\theta)}. \tag{8.5}$$

Equation (8.5) is the result we seek. It tells us that, as expected, the preferred value of θ is the value θ_0 that maximizes the entropy $S(\theta)$, eq.(8.3), because this maximizes the scalar probability density $\exp S(\theta)$. But it also tells us the degree to which values of θ away from the maximum are ruled out.

Remark: The density $\exp S(\theta)$ is a scalar function and the presence of the Jacobian factor $g^{1/2}(\theta)$ makes Eq.(8.5) manifestly invariant under changes of the coordinates θ^i in the space Θ.

8.2 The ME method

Back in section 6.2.4 we summarized the method of maximum entropy as follows:

The ME method: *We want to update from a prior distribution q to a posterior distribution when there is new information in the form of constraints $\mathcal{C}$ that specify a family $\{p\}$ of allowed posteriors. The posterior is selected through a ranking scheme that recognizes the value of prior information and the privileged role of independence. Within the family $\{p\}$ the preferred posterior P is that which maximizes the relative entropy $S[p, q]$ subject to the available constraints. No interpretation for $S[p, q]$ is given and none is needed.*

The discussion of the previous section allows us to refine our understanding of the method. ME is not an all-or-nothing recommendation to pick the single distribution that maximizes entropy and reject all others. The ME method is more nuanced: in principle all distributions within the constraint manifold ought to be included in the analysis; they contribute in proportion to the exponential of their entropy and this turns out to be significant in situations where the entropy maximum is not particularly sharp.

Going back to the original problem of updating from the prior $q(x)$ given information that specifies the manifold $\{p(x|\theta)\}$, the preferred update within the family $\{p(x|\theta)\}$ is $p(x|\theta_0)$, but to the extent that other values of θ are not totally ruled out, a better update is obtained marginalizing the joint posterior $P_J(x, \theta) = P(\theta)p(x|\theta)$ over θ,

$$P(x) = \int d^n\theta\, P(\theta)p(x|\theta) = \int d^n\theta\, g^{1/2}(\theta) \frac{e^{S(\theta)}}{\zeta} p(x|\theta)\ . \tag{8.6}$$

In situations where the entropy maximum at θ_0 is very sharp we recover the old result,

$$P(x) \approx p(x|\theta_0) \ . \tag{8.7}$$

When the entropy maximum is not very sharp eq.(8.6) is the more honest update.

The discussion in section 8.1 is itself an application of the same old ME method discussed in section 6.2.4, not on the original space $\mathcal{X}$, but on the enlarged product space $\mathcal{X} \times \Theta$. Thus, adopting the improved posterior (8.6) does not reflect a renunciation of the old ME method — only a refinement. To the summary description of the ME method above we can add the single line:

> *The ME method can be deployed to assess its own limitations and to take the appropriate corrective measures.*

Remark: Physical applications of the extended ME method are ubiquitous. For macroscopic systems the preference for the distribution that maximizes $S(\theta)$ can be overwhelming but for small systems such fluctuations about the maximum are common. Thus, violations of the second law of thermodynamics can be seen everywhere — provided we know where to look. Indeed, as we shall see in the next section, eq.(8.5) agrees with Einstein's theory of thermodynamic fluctuations and extends it beyond the regime of small fluctuations. Another important application, to be developed in chapter 9, is quantum mechanics — the ultimate theory of small systems.

We conclude this section by pointing out that there are a couple of interesting points of analogy between the pair of {maximum likelihood, Bayesian} methods and the corresponding pair of {MaxEnt, ME} methods. The first point is that maximizing the likelihood function $L(\theta|x) \overset{\text{def}}{=} p(x|\theta)$ selects a single preferred value of θ but no measure is given of the extent to which other values of θ are ruled out. The method of maximum likelihood does not provide us with a distribution for θ – the likelihood function $L(\theta|x)$ is not a probability distribution for θ. Similarly, maximizing entropy as prescribed by the MaxEnt method yields a single preferred value of the label θ but MaxEnt fails to address the question of the extent to which other values of θ are ruled out. The second point of analogy is that neither the maximum likelihood nor the MaxEnt methods are capable of handling information contained in prior distributions, while both Bayesian and ME methods can. The latter analogy is to be expected since neither the maximum likelihood nor the MaxEnt methods were designed for updating probabilities.

8.3 An application to fluctuations

The starting point for the standard formulation of the theory of fluctuations in thermodynamic systems (see [Landau 1977, Callen 1985]) is Einstein's inversion of Boltzmann's formula $S = k \log W$ to obtain the probability of a fluctuation in the form $W \sim \exp S/k$. A careful justification, however, reveals a number of

approximations which, for most purposes, are legitimate and work very well. A re-examination of fluctuation theory from the point of view of ME is, however, valuable. Our general conclusion is that the ME point of view allows exact formulations; in fact, it is clear that deviations from the canonical predictions can be expected, although in general they will be negligible. Other advantages of the ME approach include the explicit covariance under changes of coordinates, the absence of restrictions to the vicinity of equilibrium or to large systems, and the conceptual ease with which one deals with fluctuations of both the extensive as well as their conjugate intensive variables. [Caticha 2000]

This last point is an important one: within the canonical formalism (section 5.4) the extensive variables such as energy are uncertain while the intensive ones such as the temperature or the Lagrange multiplier β are fixed parameters, they do not fluctuate. There are, however, several contexts in which it makes sense to talk about fluctuations of the conjugate variables. Below we discuss the standard scenario of an open system that can exchange say, energy, with its environment.

Consider the usual setting of a thermodynamical system with microstates labelled by z. Let $m(z)dz$ be the number of microstates within the range dz. According to the postulate of "equal a priori probabilities" we choose a uniform prior distribution proportional to the density of states $m(z)$. The canonical ME distribution obtained by maximizing $S[p,m]$ subject to constraints on the expected values $\left\langle f^k \right\rangle = F^k$ of relevant variables $f^k(z)$, is

$$p(z|F) = \frac{1}{Z(\lambda)}\, m(z)\, e^{-\lambda_k f^k(z)} \quad \text{with} \quad Z(\lambda) = \int dz\, m(z)\, e^{-\lambda_k f^k(z)} \ , \tag{8.8}$$

and the corresponding entropy is

$$S(F) = \log Z(\lambda) + \lambda_k F^k \ . \tag{8.9}$$

Fluctuations of the variables $f^k(z)$ or of any other function of the microstate z are usually computed in terms of the various moments of $p(z|F)$. Within this context all expected values such as the constraints $\left\langle f^k \right\rangle = F^k$ and the entropy $S(F)$ itself are fixed; they do not fluctuate. The corresponding conjugate variables, the Lagrange multipliers $\lambda_k = \partial S/\partial F^k$, eq.(4.87), do not fluctuate either.

The standard way to make sense of λ fluctuations is to couple the system of interest to a second system, a bath, and allow exchanges of the quantities f^k. All quantities referring to the bath will be denoted by primes: the microstates are z', the density of states is $m'(z')$, and the variables are $f'^k(z')$, etc. Even though the overall expected value $\left\langle f^k + f'^k \right\rangle = F^k_T$ of the combined system plus bath is fixed, the individual expected values $\left\langle f^k \right\rangle = F^k$ and $\left\langle f'^k \right\rangle = F'^k = F^k_T - F^k$ are allowed to fluctuate. The ME distribution $p_0(z,z')$ that best reflects the prior information contained in $m(z)$ and $m'(z')$ updated by information on the total F^k_T is

$$p_0(z,z') = \frac{1}{Z_0}\, m(z)m'(z')\, e^{-\lambda_{0\alpha}\left(f^k(z)+f'^k(z')\right)}. \tag{8.10}$$

But distributions of lower entropy are not totally ruled out; to explore the possibility that the quantities F_T^k are distributed between the two systems in a less than optimal way we consider the joint distributions $p_J(z, z', F)$ constrained to the form

$$p_J(z, z', F) = p(F)p(z|F)p(z'|F_T - F), \tag{8.11}$$

where $p(z|F)$ is the canonical distribution in eq.(8.8), its entropy is eq.(8.9) and analogous expressions hold for the primed quantities.

We are now ready to write down the probability that the value of F fluctuates into a small volume $g^{1/2}(F)dF$. From eq.(8.5) we have

$$P(F)dF = \frac{1}{\zeta}\, e^{S_T(F)} g^{1/2}(F)dF, \tag{8.12}$$

where ζ is a normalization constant and the entropy $S_T(F)$ of the system plus the bath is

$$S_T(F) = S(F) + S'(F_T - F). \tag{8.13}$$

The formalism simplifies considerably when the bath is large enough that exchanges of F do not affect it, and λ' remains fixed at λ_0. Then

$$S'(F_T - F) = \log Z'(\lambda_0) + \lambda_{0k}\left(F_T^k - F^k\right) = \text{const} - \lambda_{0k}F^k. \tag{8.14}$$

It remains to calculate the determinant $g(F)$ of the information metric given by eq.(7.69),

$$g_{ij} = -\frac{\partial^2 S_T(\dot{F}, F)}{\partial\dot{F}^i\partial\dot{F}^j} = -\frac{\partial^2}{\partial\dot{F}^i\partial\dot{F}^j}\left[S(\dot{F}, F) + S'(F_T - \dot{F}, F_T - F)\right] \tag{8.15}$$

where the dot indicates that the derivatives act on the first argument. The first term on the right is

$$\begin{aligned}\frac{\partial^2 S(\dot{F}, F)}{\partial\dot{F}^i\partial\dot{F}^j} &= -\frac{\partial^2}{\partial\dot{F}^i\partial\dot{F}^j}\int dz\, p(z|\dot{F}) \log\frac{p(z|\dot{F})}{m(z)}\frac{m(z)}{p(z|F)} \\ &= \frac{\partial^2 S(F)}{\partial F^i\partial F^j} + \int dz\, \frac{\partial^2 p(z|F)}{\partial F^i\partial F^j}\log\frac{p(z|F)}{m(z)}\ .\end{aligned} \tag{8.16}$$

To calculate the integral on the right use eq.(8.8) written in the form

$$\log\frac{p(z|F)}{m(z)} = -\log Z(\lambda) - \lambda_k f^k(z)\ , \tag{8.17}$$

so that the integral vanishes,

$$-\log Z(\lambda)\frac{\partial^2}{\partial F^i\partial F^j}\int dz\, p(z|F)\ - \lambda_k\frac{\partial^2}{\partial F^i\partial F^j}\int dz\, p(z|F) f^k(z) = 0\ . \tag{8.18}$$

Similarly,

$$\begin{aligned}\frac{\partial^2}{\partial\dot{F}^i\partial\dot{F}^j}S'(F_T - \dot{F}, F_T - F) &= \frac{\partial^2 S'(F_T - F)}{\partial F^i\partial F^j} \\ &\quad + \int dz'\, \frac{\partial^2 p(z'|F_T - F)}{\partial F^i\partial F^j}\log\frac{p(z'|F_T - F)}{m'(z')}\end{aligned} \tag{8.19}$$

and here, using eq.(8.14), both terms vanish. Therefore

$$g_{ij} = -\frac{\partial^2 S(F)}{\partial F^i \partial F^j} \ . \tag{8.20}$$

We conclude that the probability that the value of F fluctuates into a small volume $g^{1/2}(F)dF$ becomes

$$p(F)dF = \frac{1}{\zeta}\ e^{S(F)-\lambda_{0k}F^k} g^{1/2}(F)dF \ . \tag{8.21}$$

This equation is exact.

An important difference with the usual theory stems from the presence of the Jacobian factor $g^{1/2}(F)$. This is required by coordinate invariance and can lead to small deviations from the canonical predictions. The quantities $\langle\lambda_k\rangle$ and $\left\langle F^k\right\rangle$ may be close but will not in general coincide with the quantities λ_{0k} and F_0^k at the point where the scalar probability density attains its maximum. For most thermodynamic systems however the maximum is very sharp. In its vicinity the Jacobian can be considered constant, and one obtains the usual results [Landau 1977], namely, that the probability distribution for the fluctuations is given by the exponential of a Legendre transform of the entropy.

The remaining difficulties are purely computational and of the kind that can in general be tackled systematically using the method of steepest descent to evaluate the appropriate generating function. Since we are not interested in variables referring to the bath we can integrate Eq.(8.11) over z', and use the distribution $P(z,F) = p(F)p(z|F)$ to compute various moments. As an example, the correlation between $\delta\lambda_i = \lambda_i - \langle\lambda_i\rangle$ and $\delta f^j = f^j - \left\langle f^j\right\rangle$ or $\delta F^j = F^j - \left\langle F^j\right\rangle$ is

$$\left\langle \delta\lambda_i \delta f^j \right\rangle = \left\langle \delta\lambda_i \delta F^j \right\rangle = -\frac{\partial \left\langle\lambda_i\right\rangle}{\partial \lambda_{0j}} + (\lambda_{0i} - \langle\lambda_i\rangle)\left(F_0^j - \left\langle F^j\right\rangle\right). \tag{8.22}$$

When the differences $\lambda_{0i} - \langle\lambda_i\rangle$ or $F_0^j - \left\langle F^j\right\rangle$ are negligible one obtains the usual expression,

$$\left\langle \delta\lambda_i \delta f^j \right\rangle \approx -\delta_i^j \ . \tag{8.23}$$

8.4 Avoiding pitfalls – II

Over the years a number of objections and paradoxes have been raised against the method of maximum entropy. Some were discussed in chapter 4. Here we discuss some objections of the type discussed in [Shimony 1985] and [Seidenfeld 1986]; see also [van Fraasen 1981 and 1986].[1] I believe some of these objections were quite legitimate at the time they were raised. They uncovered conceptual limitations with the old MaxEnt as it was understood at the time. I also believe that in the intervening decades our understanding of entropic inference has evolved to the point that all these concerns can now be addressed satisfactorily.

[1] Other objections raised by these authors, such as the compatibility of Bayesian and entropic methods, have been addressed elsewhere in these lectures.

8.4.1 The three-sided die

To set the stage for the issues involved consider a three-sided die. The die has three faces labeled by the number of spots $i = 1, 2, 3$ with probabilities $\{\theta_1, \theta_2, \theta_3\} = \theta$. The space of distributions is the simplex $\mathcal{S}_2$ with $\sum_i \theta_i = 1$. A fair die is one for which $\theta = \theta_C = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ which lies at the very center of the simplex. The expected number of spots for a fair die is $\langle i \rangle = 2$. Having $\langle i \rangle = 2$ is no guarantee that the die is fair but if $\langle i \rangle \neq 2$ the die is necessarily biased.

Next we consider three cases characterized by different states of information. First we have a situation of complete ignorance. See Fig.8-1(a). Nothing is known about the die; we do not know that it is fair but on the other hand there is nothing that induces us to favor one face over another. On the basis of this minimal information we can use MaxEnt: maximize

$$S(\theta) = -\sum_i \theta_i \log \theta_i \tag{8.24}$$

subject to $\sum_i \theta_i = 1$. The maximum entropy distribution is $\theta_{ME} = \theta_C$.

The second case involves more information: we are told that $r = \langle i \rangle = 2$. This constraint is shown in Fig.8-1(b) as a vertical dashed line that includes distributions θ other than θ_C. Therefore $r = 2$ does not imply that the die is fair. However, maximizing the entropy $S(\theta)$ subject to $\sum_i \theta_i = 1$ and $\langle i \rangle = 2$ leads us to assign $\theta'_{ME} = \theta_C$.

Finally, the third case involves even more information: we are told that the die is fair. Maximizing $S(\theta)$ subject to the constraint $\theta = \theta_C$ yields, of course, $\theta''_{ME} = \theta_C$. This is shown in Fig.8-1(c).

The fact that MaxEnt assigns the same probability to the three cases suggests that the three situations are epistemically identical — which they obviously are not — and thereby casts doubt on the validity of entropic methods in general. Indeed, failing to see a distinction where there actually is one is a prime source of paradoxes.

A more refined analysis, however, shows that — despite the fact that MaxEnt assigns the same $\theta_{ME} = \theta_C$ in all three cases — the fluctuations about θ_C are different. Indeed, the fact that the maximum of the entropy $S(\theta)$ at θ_C is not particularly sharp indicates that a full-blown ME analysis is called for. For case (a) of complete ignorance, the probability that θ lies in any small region $d^2\theta = d\theta_1 d\theta_2$ of the simplex is given by eq.(8.5),

$$P_a(\theta) d\theta_1 d\theta_2 \propto e^{S(\theta)} g^{1/2}(\theta) d\theta_1 d\theta_2 \quad \text{with} \quad g(\theta) = \frac{1}{\theta_1 \theta_2 \theta_3}. \tag{8.25}$$

The maximum of $P_a(\theta)$ is indeed at the center θ_C but the distribution is broad and extends over the whole simplex.

The ME distribution for case (b) is formally similar to case (a),

$$P_b(\theta_2) d\theta_2 \propto e^{S(\theta)} g^{1/2}(\theta_2) d\theta_2 \quad \text{with} \quad g(\theta_2) = \frac{1}{\theta_2(1-\theta_2)}. \tag{8.26}$$

The maximum of $P_b(\theta_2)$ is also at the center θ_C but the distribution is confined to the vertical line defined by $\theta_1 = \theta_3 = (1-\theta_2)/2$ in Fig.8.1(b); the probability over the rest of the simplex is strictly zero.

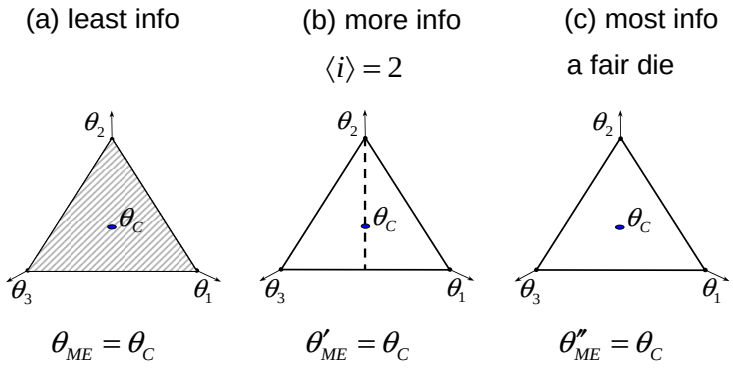


Figure 8.1: Three different states of information concerning a three-sided die. **(a)** Absolute ignorance: the distribution assigned by MaxEnt is $\theta_C = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. **(b)** We know that $r = \langle i \rangle = 2$: the MaxEnt distribution is also θ_C. **(c)** The die is known to be fair: we know that $\theta = \theta_C$. Despite the fact that MaxEnt assigns the same $\theta = \theta_C$ in all three cases the fluctuations about θ_C are different.

Finally, in case (c) the distribution is concentrated at the single central point θ_C,

$$P_c(\theta) = \delta(\theta - \theta_C) \ , \tag{8.27}$$

and there is absolutely no room for fluctuations.

To summarize: complete ignorance about $i = 1, 2, 3$ with full knowledge of $\theta = \theta_C = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is not the same as complete ignorance about both $i = 1, 2, 3$ and $\theta = \{\theta_1, \theta_2, \theta_3\}$. An assessment of 'complete ignorance' can be perfectly legitimate but to avoid confusion we must be very specific about what it is that we are being ignorant about.

8.4.2 Understanding ignorance

Ignorance, like the vacuum, is not a trivial concept.[2] Further opportunities for confusion arise when we consider constraints $\langle i\rangle = r$ with $r \neq 2$. In Fig.8-2 the constraint $\langle i\rangle = r = 1$ is shown as a vertical dashed line. Maximizing $S(\theta)$ subject to $\langle i\rangle = 1$ and normalization leads to the point at the intersection where the $r = 1$ line crosses the dotted line. The dotted curve is the set of MaxEnt distributions $\theta_{ME}(r)$ as r spans the range from 1 to 3.

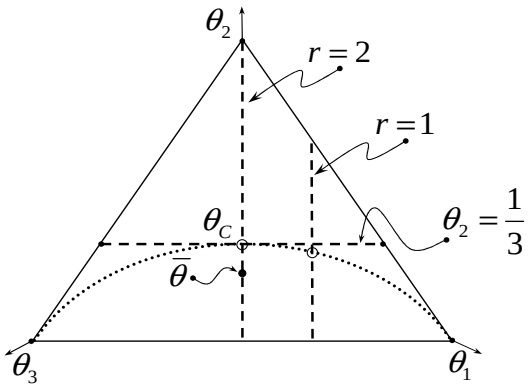


Figure 8.2: The MaxEnt solution for the constraint $\langle i\rangle = r$ for different values of r leads to the dotted line.If r is unknown averaging over r should lead to the distribution at the point θ marked by $\bar{\theta}$.

It is tempting (but ill advised) to pursue the following line of thought: We have a die but we do not know much about it. We do know, however, that the quantity $\langle i\rangle$ must have some value, call it r, about which we are ignorant too. Now, the most ignorant distribution given r is the MaxEnt distribution $\theta_{ME}(r)$. But r is itself unknown so a more honest θ assignment is an average over r,

$$\bar{\theta} = \int dr\, p(r)\theta_{ME}(r) \ , \tag{8.28}$$

[2]The title for this section is borrowed from Rodriguez's paper on the two-envelope paradox [Rodriguez 1988]. Other papers of his on the general subject of ignorance and geometry (see the bibliography) are highly recommended for the wealth of insights they contain.

where $p(r)$ reflects our uncertainty about r. It may, for example, make sense to pick a uniform distribution over r but the precise choice is not important for our purposes. The point is that since the MaxEnt dotted curve is concave the point $\bar{\theta}$ necessarily lies below θ_C so that $\bar{\theta}_2 < 1/3$. And we have a paradox: we started admitting complete ignorance and through a process that claims to express full ignorance at every step we reach the conclusion that the die is biased against $i = 2$. Where is the mistake?

The first clue is symmetry: We started with a situation that treats the outcomes $i = 1, 2, 3$ symmetrically and end up with a distribution that is biased against $i = 2$. The symmetry must have been broken somewhere and it is clear that this happened at the moment the constraint on $\langle i \rangle = r$ was imposed — this is shown as *vertical* lines on the simplex. Had we chosen to express our ignorance not in terms of the unknown value of $\langle i \rangle = r$ but in terms of some other function $\langle f(i) \rangle = s$ then we could have easily broken the symmetry in some other direction. For example, let $f(i)$ be a cyclic permutation of i,

$$f(1) = 2, \quad f(2) = 3, \quad \text{and} \quad f(3) = 1 \ , \tag{8.29}$$

then repeating the analysis above would lead us to conclude that $\bar{\theta}_3 < 1/3$, which represents a die biased against $i = 3$. Thus, the question becomes: What leads to choose a constraint on $\langle i \rangle$ rather than a constraint on $\langle f \rangle$ when we are equally ignorant about both?

The discussion in section 4.10 is relevant here. There we identified four epistemically different situations:

(A) The ideal case: We know that $\langle f \rangle = F$ and we know that it captures all the information that happens to be relevant to the problem at hand.

(B) The important case: We know that $\langle f \rangle$ captures all the information that happens to be relevant to the problem at hand but its actual numerical value F is not known.

(C) The predictive case: There is nothing special about the function f except that we happen to know its expected value, $\langle f \rangle = F$. In particular, we do not know whether information about $\langle f \rangle$ is complete or whether it is at all relevant to the problem at hand.

(D) The extreme ignorance case: We know neither that $\langle f \rangle$ captures relevant information nor its numerical value F. There is nothing that singles out one function f over any other.

The paradox with the three-sided die arises because two epistemically different situations, case B and case D have been confused. On one hand, the unknown die is meant to reflect a situation of complete ignorance, case D. We do not know whether it is the constraint $\langle i \rangle$ or any other function $\langle f \rangle$ that captures relevant information; and their numerical values are also unknown. There is nothing to single out $\langle i \rangle$ or $\langle f \rangle$ and therefore the correct inference consists of maximizing

S imposing the only constraint we *actually* know, namely, normalization. The result is as it should be — a uniform distribution ($\theta_{ME} = \theta_C$).

On the other hand, the argument that led to the assignment of $\bar{\theta}$ in eq.(8.28) turns out to be actually correct when applied to an epistemic situation of type B. Imposing the constraint $\langle i \rangle = r$ when r is unknown and then averaging over r represents a situation in which *we know something*. We have some knowledge that singles out $\langle i \rangle$ — and not any other $\langle f \rangle$ — as the function that captures information that is relevant to the die. There is some ignorance here — we do not know r — but this is not extreme ignorance. We can summarize as follows: knowing that the die is biased against $i = 2$ but not knowing by how much is not the same as not knowing anything.

A different instance of the same paradox is discussed in [Shimony 1985]. A physical system can be in any of n microstates labeled $i = 1 \ldots n$. When we know absolutely nothing about the system maximizing entropy subject to the single constraint of normalization leads to a uniform probability distribution, $p_u(i) = 1/n$. A different (misleading) way to express complete ignorance is to argue that the expected energy $\langle \varepsilon \rangle$ must have some value E about which we are ignorant. Maximizing entropy subject to both $\langle \varepsilon \rangle = E$ and normalization leads to the usual Boltzmann distributions,

$$p(i|\beta) = \frac{e^{-\beta\varepsilon_i}}{Z(\beta)} \quad \text{where} \quad Z(\beta) = \sum_i e^{-\beta\varepsilon_i} \ . \tag{8.30}$$

Since the inverse temperature $\beta = \beta(E)$ is itself unknown we must average over β,

$$p_t(i) = \int d\beta \, p(\beta) p(i|\beta) \ . \tag{8.31}$$

To the extent that both distributions reflect complete ignorance we must have

$$p_u(i) = p_t(i) \qquad \text{((wrong))}$$

which can only happen provided

$$p(\beta) = \delta(\beta) \quad \text{or} \quad \beta = 0 \ . \tag{8.32}$$

Indeed, setting the Lagrange multiplier $\beta = 0$ in $p(i|\beta)$ amounts to maximizing entropy without imposing the energy constraint and this leads to the uniform distribution $p_u(i)$. But now we have a paradox: The first way of expressing complete ignorance about the system implies we are ignorant about its temperature. In fact, we do not even know that it has a temperature at all, much less that it has a single uniform temperature. But we also have a second way of expressing ignorance and if impose that the two agree we are led to conclude that β has the precise value $\beta = 0$; we have concluded that the system is infinitely hot — ignorance is hell.

The paradox is dissolved once we realize that, just as with the die problem, we have confused two epistemically different situations — types D and B above: Knowing nothing about a system is not the same as merely not knowing its

temperature — while knowing full well that it is in thermal equilibrium and that it actually has a temperature.

It may be worthwhile to rephrase this important point in different words. If $\mathcal{I}$ is the space of microstates and β is some unknown arbitrary quantity in some space $\mathcal{B}$ the rules of probability theory allow us to write

$$p(i) = \int d\beta\, p(i,\beta) \quad \text{where} \quad p(i,\beta) = p(\beta)p(i|\beta)\ . \tag{8.33}$$

Paradoxes will easily arise if we fail to distinguish a situation of complete ignorance from a situation where the conditional probability $p(i|\beta)$ — which is what gives meaning to the parameter β — is known. Or, to put it in yet another way: complete ignorance over the space $\mathcal{I}$ is not the same as complete ignorance over the full space $\mathcal{I} \times \mathcal{B}$.

Chapter 9

Entropic Dynamics: Time and Quantum Theory

Law without Law: *"The only thing harder to understand than a law of statistical origin would be a law that is not of statistical origin, for then there would be no way for it — or its progenitor principles — to come into being."*

Two tests: *"No test of these views looks like being someday doable, nor more interesting and more instructive, than a derivation of the structure of quantum theory... No prediction lends itself to a more critical test than this, that every law of physics, pushed to the extreme, will be found statistical and approximate, not mathematically perfect and precise."*

J. A. Wheeler[1]

Quantum mechanics involves probabilities in a fundamental way and, therefore, it is a theory of inference. But this has not always been clear. The controversy revolves around the interpretation of the quantum state — the wave function. Does it represent the actual real state of the system — its *ontic* state — or does it represent a state of knowledge about the system — an *epistemic* state? The problem has been succinctly stated by Jaynes: "Our present QM formalism is a peculiar mixture describing in part realities in Nature, in part incomplete human information about Nature — all scrambled up by Heisenberg and Bohr into an omelette that nobody has seen how to unscramble." [Jaynes 1990]

The ontic interpretations have been the most common. From the very beginning, Schrödinger's original waves were meant to be real material waves — although formulating the theory in configuration space immediately introduced problems. Then the "orthodox" interpretation (sometimes but not always called

[1] [Wheeler Zurek 1983, p. 203 and 210]

the Copenhagen interpretation) gradually took over. As crystallized in the standard textbooks (including the early classics by Dirac and von Neumann) it regards the quantum state as a complete objective specification of the properties of the system — a concept that is totally divorced from the state of belief of a rational agent. The conceptual problems that plagued the orthodox interpretation motivated the creation of alternatives such as the de Broglie-Bohm pilot wave theory and Everett's many worlds interpretation. In both the wave function represents a real state of affairs. On the other side, the epistemic interpretation has had a growing number of advocates starting, most prominently, with Einstein and Heisenberg [Heisenberg 1958].[2]

Faced with this controversy, Jaynes also understood where to start looking for a solution: "We suggest that the proper tool for incorporating human information into science is simply probability theory — not the currently taught 'random variable' kind, but the original 'logical inference' kind of James Bernoulli and Laplace" which he explains "is often called Bayesian inference" and is "supplemented by the notion of information entropy". Bohr, Heisenberg, Einstein and other founders of quantum theory might have agreed. They were keenly aware of the epistemological and pragmatic elements in quantum mechanics (see e.g., [Stapp 1972]) but they wrote at a time when the language and the tools of quantitative epistemology — Bayesian and entropic methods — had not yet been sufficiently developed.

But interpreting quantum theory is not merely a matter of postulating the mathematical formalism and then appending an interpretation to it. For the epistemic view of quantum states to be satisfactory it is not sufficient to state that wave functions are tools for codifying our beliefs. It is also necessary to show that the particular ways in which quantum states are calculated and manipulated are in complete agreement with the highly constrained ways in which probabilities are to be manipulated, computed, and updated. Let us be more explicit: it is not sufficient to accept that $|\psi|^2$ represents a state of knowledge; we must also provide an epistemic interpretation for the phase of the wave function. Furthermore, we must show that changes or updates of the epistemic ψ — which include both unitary time evolution according to the Schrödinger equation and the projection postulate during measurement — are nothing but instances of entropic updating (including Bayes rule as a special case). There is no room for alternative "quantum" probabilities obeying alternative forms of Bayesian inference.

Our goal is to derive quantum theory (including its classical mechanics limit) as an example of entropic inference. In essence, we want to do for quantum mechanics what Jaynes did for statistical mechanics.

The wave function will be explicitly epistemic — which means neither fully objective nor fully subjective. In earlier chapters we argued against a sharp subjective/objective dichotomy. The point is that probabilities will unavoidably retain a subjective character but they are useful only to the extent that

[2]For more recent advocates of the epistemic interpretation see [Ballentine 1970, 1998; Caves et al 2007; Harrigan Spekkens 2010; Friedrich 2011] and references therein. For criticism of the epistemic view see *e.g.* [Zeh 2002; Ferrero et al 2004; Marchildon 2004]

the subjectivity is controlled, tempered by some objectivity. The injection of objectivity occurs through updating: posteriors are more useful than priors. We update to enhance objectivity — otherwise, why bother? The wave function inherits this dual subjective/objective character. In other words, being epistemic does not preclude some measure of objectivity. Wave functions capture information that, although incomplete and perhaps uncompletable, is undeniably *relevant*.

A central feature of the entropic dynamics (ED) model developed below is the privileged role we assign to position over and above all other observables. Strictly, position is the only *observable*. This is one important difference from other approaches that also emphasize notions of information.[3] ED shows formal similarities with another position-based model — the stochastic mechanics developed by Nelson [Nelson 1966, 1967, 1985][4] — but there are important conceptual differences. Stochastic mechanics operates at the ontological level; its goal is a realistic interpretation of quantum theory as arising from a deeper, possibly non-local, but essentially classical "reality". In Nelson's own words: "... what stochastic mechanics is all about: it is an attempt to build a naively realistic picture of physical phenomena, an objective representation of physical processes without reference to any observer" [Nelson 1986]. In contrast, in the ED model there is no underlying classical dynamics that rules over a subquantum world. ED operates almost completely at the epistemological level:

> *The laws of quantum mechanics are not laws of nature; they are rules for processing relevant information about nature.*

An important feature is that ED is also a model for time. Indeed, the rules of Bayesian and entropic inference are silent on the matter of time; they are completely atemporal. This means that the process of developing a dynamics of change driven by entropy will require constructing a notion of time. Such an "entropic" time is a book-keeping device designed to keep track of the accumulation of change. It deserves to be called 'time' because it includes (a) something one might identify as "instants"; (b) a sense in which these instants can be "ordered"; and (c) a convenient concept of "duration" measuring the separation between instants. The welcome new feature is that entropic time is intrinsically directional; an arrow of time is generated automatically. As we shall see, for the pragmatic purpose of predicting the empirically observable correlations among particles nothing more "physical" than entropic time is needed.

In this chapter we introduce the ED model for time and quantum theory following [Caticha 2010]. Except for side comments comparing ED to other approaches to quantum theory, the treatment will be (hopefully) self-contained.[5]

[3] For a very incomplete list where more references can be found see *e.g.*, [Wootters 1981; Caticha 1998, 2006; Brukner Zeilinger 2002; Fuchs 2002; Spekkens 2007; Goyal et al 2010; Hardy 2011].

[4] See also [Guerra 1981, Guerra Morato 1983] and references therein.

[5] While I do not assume that the reader has had an extensive prior education in quantum mechanics I doubt very much that readers who are totally innocent in these matters will have made it this far into the book.

In the next chapter we discuss the quantum measurement problem and the introduction of observables other than position [Johnson Caticha 2011], including the definition of momentum and the corresponding uncertainty relations [Nawaz Caticha 2011].

But before we proceed with our subject I must emphasize that the struggle to overcome the conceptual difficulties with quantum theory has engendered a literature that is too vast to even consider reviewing here. Excellent sources for the earlier work are found in [Jammer 1966, 1974; Wheeler Zurek 1983]; for more recent work see, *e.g.* [Jaegger 2009]. Even within the more restricted subject of developing quantum theory without appealing to the notion of a "quantum" probability there exist many proposals. Some share with Nelson's approach the foundation of an underlying classical dynamics with an additional stochastic element. The sub-quantum dynamics is variously described by a classical action principle, or a Liouville equation, or even Newton's law. The additional stochastic element has been introduced in a variety of ways:[6] through an extra momentum fluctuation [Hall Reginatto 2002]; a hidden non-equilibrium thermodynamics [Groessing 2008, 2009]; Brownian fluctuations caused by energy exchanges with the surrounding vacuum [Fritsche Haug 2009]; coarse graining an underlying dynamics that is reparametrization-invariant and ergodic [Elze 2002, 2003]; tracing out certain inaccessible degrees of freedom [Smolin 2006; Wetterich 2010]; through explicit dissipation ['t Hooft 1988]; and also as the statistical mechanics of a particular class of matrix models [Adler 2004]. In contrast, the ED described here does not assume any underlying mechanics whether classical, deterministic, or stochastic. Both quantum dynamics and its classical limit are derived as examples of entropic inference.

9.1 The statistical model

Just as in any other problem of entropic inference we must first identify the microstates that are the subject of our inference; we must also identify prior probabilities; and finally, we must identify the constraints that represent the information that is relevant to our problem.

Consider particles living in flat three-dimensional space. The particles have definite positions x.[7] For a single particle the configuration space $\mathcal{X}$ is Euclidean with metric

$$\gamma_{ab} = \frac{\delta_{ab}}{\sigma^2} \ . \tag{9.1}$$

(The reason for the scale factor σ^2 will become clear once we generalize to N particles below and in section 9.7.) Our first assumption is that

[6] This list is not meant to be exhaustive; it merely provides an entry point to the literature. More recent work by these authors and related work can be found at arxiv.org.

[7] In this work entropic dynamics is developed as a model for the quantum mechanics of particles. The same framework can be deployed to construct models for the quantum mechanics of fields, in which case it is the fields that are "real" and have well defined albeit unknown values.

In addition to the particles there exist some extra variables y that live in a space $\mathcal{Y}$ and are subject to an uncertainty that depends on the location x of the particles and is described by some unspecified probability distribution $p(y|x)$.

The number and nature of the extra variables $y \in \mathcal{Y}$ and the origin of their uncertainty need not be specified. The assumption that there exist other variables out there in the world does not appear excessive or unnatural. It is a strength of this model that our conclusions hold irrespective of any detailed assumptions about the y variables.[8] As we shall see it is the entropy of the distributions $p(y|x)$ that plays a significant role in defining the dynamics of x; the finer details of $p(y|x)$ turn out to be irrelevant.

For a single particle the *statistical manifold* $\mathcal{M}$ of distributions $p(y|x)$ is three-dimensional: for each x there is a corresponding $p(y|x)$. Each distribution $p(y|x) \in \mathcal{M}$ can be conveniently labeled by its corresponding x so that the label x denotes both a regular point in the configuration space $\mathcal{X}$ and also its corresponding "point" in the statistical manifold $\mathcal{M}$. For later reference, the entropy $S(x)$ of $p(y|x)$ relative to an underlying measure $q(y)$ of the space $\mathcal{Y}$ is[9]

$$S(x) = -\int dy\, p(y|x) \log \frac{p(y|x)}{q(y)} \ . \tag{9.2}$$

This entropy $S(x)$ is a natural scalar field on both the configuration space $\mathcal{X}$ and the statistical manifold $\mathcal{M}$.

The peculiar features of quantum mechanics such as non-local correlations and entanglement will arise naturally provided the theory for N particles is formulated on the $3N$-dimensional configuration space $\mathcal{X}_N$. Accordingly, to complete the specification of the model we need to describe $\mathcal{X}_N$ and its corresponding statistical manifold $\mathcal{M}_N$. The generalization is straightforward. For N particles the y variable distributions are $p(y|x)$ where now the position $x \in \mathcal{X}_N$ is given by x^A and the index A now takes $3N$ values. More explicitly $x = (x^{a_1}, x^{a_2} \ldots)$ where $a_1 = 1, 2, 3$ denotes the first particle, $a_2 = 4, 5, 6$ denotes the second particle, and so on. The $3N$-dimensional configuration space $\mathcal{X}_N$ remains flat but it is not, in general, isotropic. For example, for $N = 2$ particles the metric, written in block matrix form, is

$$\gamma_{AB} = \begin{bmatrix} \delta_{a_1 b_1}/\sigma_1^2 & 0 \\ 0 & \delta_{a_2 b_2}/\sigma_2^2 \end{bmatrix} \ . \tag{9.3}$$

We shall later see that this choice of an anisotropic configuration space leads to a theory of particles with different masses. For particles that are identical the appropriate configuration space is isotropic with $\sigma_1 = \sigma_2 = \ldots = \sigma$.

To summarize, the first basic assumption is that there exist particles which have definite albeit unknown positions x and there existence of some extra

[8] The y variables will be referred to as extra variables or just y variables. In section 9.10 we shall argue that they are not hidden variables.

[9] This is a multidimensional integral over all y variables; for simplicity we write dy instead of $d^n y$.

variables y subject to an x-dependent uncertainty described by some unspecified distributions $p(y|x)$. The statistical manifold $\mathcal{M}_N$ and the entropy field $S(x)$ are convenient inference tools introduced to explore the implications of this assumption.

9.2 Entropic dynamics

The second basic assumption is that

> *Small changes from one state to another are possible and do, in fact, happen. Large changes are assumed to result from the accumulation of many small changes.*

We do not explain why changes happen but, given the information that they occur, our problem is to venture a guess about what to expect. Consider a single particle (the generalization to several particles is immediate and will be carried out in section 9.7) that moves away from an initial position x to an unknown final position x'. All we know about x' is that it is near x. What can we say about x'? Since x and x' represent probability distributions we see that this is precisely the kind of problem the method of maximum entropy (ME) has been designed to solve, namely, to update from a prior distribution to a posterior distribution selected from within a specified set. As in all ME problems success hinges on appropriate choices of the entropy, prior distribution, and constraints.

Since neither the new x' nor the new variables y' are known what we want is the joint distribution $P(x', y'|x)$ and the relevant space is $\mathcal{X} \times \mathcal{Y}$. To find it maximize the appropriate (relative) entropy,

$$\mathcal{S}[P, Q] = -\int dx'dy'\, P(x', y'|x) \log \frac{P(x', y'|x)}{Q(x', y'|x)} \ . \tag{9.4}$$

The relevant information is introduced through the prior $Q(x', y'|x)$ and the constraints that specify the family of acceptable posteriors $P(x', y'|x)$.

The prior

We select a prior that represents a state of extreme ignorance: the relation between x' and y' is not known; knowledge of x' tells us nothing about y' and vice versa. Such ignorance is represented by a product, $Q(x', y'|x) = Q(x'|x)Q(y'|x)$. Furthermore we take the distributions $Q(y'|x)dy'$ and $Q(x'|x)d^3x'$ to be uniform, that is, proportional to the respective volume elements which are respectively given by $dv_x = \gamma^{1/2}d^3x$ [where $\gamma = \det \gamma_{ab}$, see eq.(9.1)] and by $dv_y = q(y)dy$ where the measure $q(y)$ need not be specified further. Therefore, up to an irrelevant proportionality constant, the joint prior is

$$Q(x', y'|x) = \gamma^{1/2} q(y') \ . \tag{9.5}$$

The constraints

Next we specify the constraints. Write the posterior as

$$P(x', y'|x) = P(x'|x)P(y'|x', x) \tag{9.6}$$

and consider the two factors separately. First we require that x' and y' be related to each other in a very specific way, namely that $P(y'|x', x) = p(y'|x') \in \mathcal{M}$ — the uncertainty in y' depends only on x', and not on the previous position x. Therefore, our first constraint is that the joint posterior be of the form

$$P(x', y'|x) = P(x'|x)p(y'|x') \ . \tag{9.7}$$

The second constraint concerns the factor $P(x'|x)$ and represents the fact that actual physical changes do not happen discontinuously: we require that x' be an infinitesimally short distance away from x. Let $x'^a = x^a + \Delta x^a$. We require that the expectation

$$\left\langle \Delta\ell^2(x', x) \right\rangle = \left\langle \gamma_{ab} \Delta x^a \Delta x^b \right\rangle = \Delta\bar{\ell}^2 \tag{9.8}$$

be some small but for now unspecified numerical value $\Delta\bar{\ell}^2$ which could in principle depend on x. (This is just as in the statistical mechanics of equilibrium, section 4.10, where a constraint on the expected energy $\langle\varepsilon\rangle$ is recognized as codifying relevant information but its numerical value, $\langle\varepsilon\rangle = E$, is not known.)

Entropy maximization

Having specified the prior and the constraints the ME method takes over. Substituting the prior (9.5) and the constraint (9.7) into the joint entropy (9.4) gives

$$\mathcal{S}[P, Q] = -\int dx' \, P(x'|x) \log \frac{P(x'|x)}{\gamma^{1/2}} + \int dx' \, P(x'|x) S(x') \ , \tag{9.9}$$

where $S(x)$ is given in eq.(9.2). Next we vary $P(x'|x)$ to maximize $\mathcal{S}[P, Q]$ subject to (9.8) and normalization. The result is

$$P(x'|x) = \frac{1}{\zeta(x, \alpha)} e^{S(x') - \frac{1}{2}\alpha(x)\Delta\ell^2(x', x)} \ , \tag{9.10}$$

where

$$\zeta(x, \alpha) = \int dx' \, e^{S(x') - \frac{1}{2}\alpha(x)\Delta\ell^2(x', x)} \ , \tag{9.11}$$

and the Lagrange multiplier $\alpha(x)$ is determined from the constraint eq.(9.8),

$$\frac{\partial}{\partial\alpha} \log \zeta(x, \alpha) = -\frac{1}{2}\Delta\bar{\ell}^2 \ . \tag{9.12}$$

The distribution (9.10) is not merely a local maximum or a stationary point, it yields the absolute maximum of the relative entropy $\mathcal{S}[P, Q]$ subject to the constraints (9.7) and (9.8). The proof follows the standard argument originally due to Gibbs (see section 4.9).

Analysis

The probability of a step from x to x', eq.(9.10), represents a compromise between three conflicting tendencies. One, which can be traced to the uniform prior $Q(x'|x) = \gamma^{1/2}$ and is represented by the first integral in (9.9), is to make $P(x'|x)$ spread as uniformly as possible. Another, induced by the second integral in (9.9), contributes the entropy term in the exponent of $P(x'|x)$ and favors a single giant step to the distribution $p(y'|x')$ that maximizes the entropy $S(x')$. And last, the constraint on $\langle \Delta\ell^2 \rangle$ leads to the $\Delta\ell^2(x', x)$ term in the exponent of $P(x'|x)$ and favors values of x' that are close to x. Large α means short steps. The compromise in eq.(9.10) leads to short steps in essentially random directions with a small anisotropic bias along the entropy gradient.

Next we seek a more useful expression for $P(x'|x)$ for large α by expanding the exponent about its maximum. Let $x'^a = x^a + \Delta x^a$. The exponent is maximum at $x' = \bar{x}$ such that

$$\frac{\partial}{\partial x'^a}\left[S(x') - \frac{\alpha}{2}\gamma_{ab}\Delta x^a \Delta x^b\right]_{\bar{x}} = 0 \quad \text{or} \quad \partial_a S = \alpha\gamma_{ab}\,\Delta\bar{x}^b \ , \tag{9.13}$$

where $\partial_a = \partial/\partial x^a$. Therefore,

$$\Delta\bar{x}^a = \bar{x}^a - x^a = \frac{1}{\alpha}\gamma^{ab}\partial_b S(x) \ . \tag{9.14}$$

Then the exponent in (9.10) becomes

$$\begin{aligned} S(x') - \frac{\alpha}{2}\gamma_{ab}\Delta x^a \Delta x^b &= S(x) + \partial_a S\,\Delta x^a + \frac{1}{2}\partial_a\partial_b S\,\Delta x^a \Delta x^b - \frac{\alpha}{2}\gamma_{ab}\Delta x^a \Delta x^b \\ &= S(x) + \alpha\gamma_{ab}\,\Delta\bar{x}^b \Delta x^a - \frac{\alpha}{2}\gamma_{ab}\Delta x^a \Delta x^b \end{aligned} \tag{9.15}$$

where the term $\partial_a\partial_b S\,\Delta x^a \Delta x^b$ can be dropped because for large α it is negligible relative to the other terms. Therefore,

$$S(x') - \frac{\alpha}{2}\gamma_{ab}\Delta x^a \Delta x^b = S(x) - \frac{\alpha}{2}\gamma_{ab}\left(\Delta x^a - \Delta\bar{x}^a\right)\left(\Delta x^b - \Delta\bar{x}^b\right) + \frac{\alpha}{2}\gamma_{ab}\Delta\bar{x}^a \Delta\bar{x}^b \ . \tag{9.16}$$

Thus, for large α the transition probability, eq.(9.10), becomes a Gaussian distribution,

$$P(x'|x) \approx \frac{1}{Z(x)}\exp\left[-\frac{\alpha(x)}{2\sigma^2}\delta_{ab}\left(\Delta x^a - \Delta\bar{x}^a\right)\left(\Delta x^b - \Delta\bar{x}^b\right)\right] \ . \tag{9.17}$$

The first and third terms on the right of eq.(9.16) are independent of x' and they have been absorbed into a new normalization $Z(x)$. Thus, the displacement Δx^a can be expressed as an expected drift plus a fluctuation,

$$\Delta x^a = \Delta\bar{x}^a + \Delta w^a \ , \tag{9.18}$$

where

$$\langle \Delta x^a \rangle = \Delta\bar{x}^a = \frac{\sigma^2}{\alpha(x)}\delta^{ab}\partial_b S(x) \ , \tag{9.19}$$

$$\langle \Delta w^a \rangle = 0 \quad \text{and} \quad \left\langle \Delta w^a \Delta w^b \right\rangle = \frac{\sigma^2}{\alpha(x)} \delta^{ab} \ . \tag{9.20}$$

The particle tends to drift along the entropy gradient. Note that as $\alpha \to \infty$ the steps get correspondingly smaller but the fluctuations become dominant: the drift is $\Delta \bar{x} \sim O(\alpha^{-1})$ while the fluctuations are much larger $\Delta w \sim O(\alpha^{-1/2})$. This implies that as $\alpha \to \infty$ the trajectory is continuous but not differentiable — just as in Brownian motion.

We can now return to the unfinished business of choosing $\Delta \bar{\ell}^2$ in eq.(9.8) which is equivalent to choosing the multiplier $\alpha(x)$. We invoke a symmetry argument. We just saw that in the limit of infinitesimally short steps the relevant dynamics is dominated by the fluctuations Δw. In order that the dynamics reflect the translational symmetry of the configuration space $\mathcal{X}$ we choose $\alpha(x)$ so that the fluctuations $\left\langle \Delta w^a \Delta w^b \right\rangle$ in eq.(9.20) be independent of x. Therefore $\alpha(x) =$ constant.

9.3 Entropic time

Our goal is to derive laws of physics as an application of inference methods but the latter make no reference to time so additional assumptions are needed. *The foundation to any notion of time is dynamics.* We introduce time as a convenient book-keeping device to keep track of the accumulation of small changes.

In this section we show how a dynamics driven by entropy naturally leads to an "entropic" notion of time. Our task here is to develop a model that includes (a) something one might identify as an "instant", (b) a sense in which these instants can be "ordered", (c) a convenient concept of "duration" measuring the separation between instants. A welcome bonus is that the model incorporates an intrinsic directionality — an evolution from past instants towards future instants. Thus, an arrow of time does not have to be externally imposed but is generated automatically. This set of concepts constitutes what we will call "entropic time".

Important questions such as the relation between entropic time, in which instants are ordered through the sequence of inference steps, and an externally imposed structure of a presumably "physical" time will be discussed later (section 9.8) after the dynamics has been more fully developed.

9.3.1 Time as a sequence of instants

In entropic dynamics change is given, at least for infinitesimally short steps, by the transition probability $P(x'|x)$ in eq.(9.17). For finite steps the relevant piece of information is that large changes occur *only* as the result of a continuous succession of very many small changes.

Consider the nth step. In general we will be uncertain about both its initial and the final positions, x and x'. This means we must deal with the joint probability $P(x', x)$. Using $P(x', x) = P(x'|x)P(x)$ and integrating over x, we

get

$$P(x') = \int dx\, P(x'|x)P(x) \ . \tag{9.21}$$

It is important to emphasize that this equation is a direct consequence of the laws of probability — no assumptions of a physical nature have been made. However, when we consider the transition probability from x to x', given by $P(x'|x)$, it is implicitly assumed that the three initial coordinates $x = (x^1, x^2, x^3)$ occur all at one instant and similarly that the three final coordinates $x' = (x'^1, x'^2, x'^3)$ also occur all at another instant. Thus, if $P(x)$ happens to be the probability of different values of x at *an "initial" instant of entropic time* t, then it is tempting to interpret $P(x')$ as the probability of values of x' at *a "later" instant of entropic time* $t' = t + \Delta t$. Accordingly, we write $P(x) = \rho(x, t)$ and $P(x') = \rho(x', t')$ so that

$$\rho(x', t') = \int dx\, P(x'|x)\rho(x, t) \ . \tag{9.22}$$

Nothing in the laws of probability that led to eq.(9.21) forces this interpretation on us — this is an independent assumption about what constitutes time in our model. We use eq.(9.22) to define what we mean by an instant:

> *An instant is defined by a probability distribution $\rho(x)$. If the distribution $\rho(x, t)$ refers to a certain instant t, then the distribution $\rho(x', t')$ in eq.(9.22) defines what we mean by the "next" instant, $t' = t + \Delta t$.*

Thus, eq.(9.22) allows entropic time to be constructed, step by step, as a succession of instants.

We can phrase this idea somewhat differently. Once we have decided what is the information that is relevant for predicting future behavior we imagine all that information codified into a single instant and we use this to define what we mean by the "present" instant:

> *Given the present the future is independent of the past.*

Remark: An equation such as (9.22) is commonly employed to define Markovian behavior in which case it is known as the Chapman-Kolmogorov equation. Markovian processes are such that specifying the state of the system at time t is sufficient to fully determine its state after time t — no additional information about times before t is needed. We make no Markovian assumptions. We are concerned with a different problem. We do not use (9.22) to define Markovian processes; we use it to define time.

9.4 Duration: a convenient time scale

Having introduced the notion of successive instants we now have to specify the interval Δt between them. Successive instants are connected through the transition probability $P(x'|x)$. Specifying the interval of time Δt between successive instants amounts to tuning the steps or, equivalently, the multiplier $\alpha(x, t)$. To

model a time that, like Newtonian time, flows "equably" everywhere, that is, at the same rate at all places and times we define Δt as being independent of x, and such that every Δt is as long as the previous one. Inspection of the actual dynamics as given in eq.(9.17-9.20) shows that this is achieved if we choose $\alpha(x,t)$ so that

$$\alpha(x,t) = \frac{\tau}{\Delta t} = \text{constant} \ , \tag{9.23}$$

where τ is a constant introduced so that Δt has units of time. As already anticipated in the previous section, it is the translational symmetry of the configuration space $\mathcal{X}$ expressed as the "equable" flow of time that leads us to impose uniformity on the expected step sizes $\Delta\bar{\ell}$ and the corresponding multipliers α. This completes the implementation of entropic time. In the end, however, the only justification for any definition of duration is that it simplifies the description of motion, and indeed, the transition probability in eq.(9.17) becomes

$$P(x'|x) \approx \frac{1}{Z(x)} \exp\left[-\frac{\tau}{2\sigma^2\Delta t}\delta_{ab}\left(\Delta x^a - \Delta\bar{x}^a\right)\left(\Delta x^b - \Delta\bar{x}^b\right)\right] \ , \tag{9.24}$$

which we recognize as a standard Wiener process. A displacement $\Delta x = x' - x$ is given by

$$\Delta x^a = b^a(x)\Delta t + \Delta w^a \ , \tag{9.25}$$

where the drift velocity $b^a(x)$ and the fluctuation Δw^a are

$$\langle \Delta x^a \rangle = b^a \Delta t \quad \text{with} \quad b^a(x) = \frac{\sigma^2}{\tau}\,\delta^{ab}\partial_b S(x) \ , \tag{9.26}$$

$$\langle \Delta w^a \rangle = 0 \quad \text{and} \quad \left\langle \Delta w^a \Delta w^b \right\rangle = \frac{\sigma^2}{\tau}\Delta t\,\delta^{ab} \ . \tag{9.27}$$

The constant $\sigma^2/2\tau$ plays the role of the diffusion constant in Brownian motion. The formal similarity to Nelson's stochastic mechanics [Nelson 1966] is evident. An important difference concerns the expression of the drift velocity as the gradient of a scalar function: unlike stochastic mechanics, here eq.(9.26) has been derived rather than postulated, and $S(x)$ is not merely an uninterpreted auxiliary scalar function—it turns out to be the entropy of the y variables.

9.4.1 The directionality of entropic time

Time constructed according to eq.(9.22) is remarkable in yet another respect: the inference implied by $P(x'|x)$ in eq.(9.17) incorporates an intrinsic directionality in entropic time: there is an absolute sense in which $\rho(x,t)$ is prior and $\rho(x',t')$ is posterior.

Suppose we wanted to find a time-reversed evolution. We would write

$$\rho(x,t) = \int dx'\, P(x|x')\rho(x',t') \, . \tag{9.28}$$

This is perfectly legitimate but the transition probability $P(x|x')$ cannot be obtained from eq.(9.17) by merely exchanging x and x'. Indeed, according to the rules of probability theory $P(x|x')$ is related to eq.(9.17) by Bayes' theorem,

$$P(x|x') = \frac{P(x)}{P(x')} P(x'|x) \ . \tag{9.29}$$

In other words, one of the two transition probabilities, either $P(x|x')$ or $P(x'|x)$, *but not both*, can be given by the maximum entropy distribution eq.(9.17). The other is related to it by Bayes' theorem. I hesitate to say that this is what breaks the time-reversal symmetry because the symmetry was never there in the first place. There is no symmetry between prior and posterior; there is no symmetry between the inferential past and the inferential future.

An interesting consequence of the time asymmetry is that the mean velocities towards the future and from the past do not coincide. Let us be more specific. Equation (9.26) gives the *mean velocity to the future* or *future drift*,

$$\begin{aligned} b^a(x) &= \lim_{\Delta t\to 0^+} \frac{\langle x^a(t+\Delta t)\rangle_{x(t)} - x^a(t)}{\Delta t} \\ &= \lim_{\Delta t\to 0^+} \frac{1}{\Delta t}\int dx'\, P(x'|x)\Delta x^a \ , \end{aligned} \tag{9.30}$$

where $x = x(t)$, $x' = x(t+\Delta t)$, and $\Delta x^a = x'^a - x^a$. Note that the expectation in (9.30) is conditional on the earlier position $x = x(t)$. One can also define a *mean velocity from the past* or *past drift*,

$$b^a_*(x) = \lim_{\Delta t\to 0^+} \frac{x^a(t) - \langle x^a(t-\Delta t)\rangle_{x(t)}}{\Delta t} \tag{9.31}$$

where the expectation is conditional on the later position $x = x(t)$. Shifting the time by Δt, b^a_* can be equivalently written as

$$\begin{aligned} b^a_*(x') &= \lim_{\Delta t\to 0^+} \frac{x^a(t+\Delta t) - \langle x^a(t)\rangle_{x(t+\Delta t)}}{\Delta t} \\ &= \lim_{\Delta t\to 0^+} \frac{1}{\Delta t}\int dx\, P(x|x')\Delta x^a \ , \end{aligned} \tag{9.32}$$

with the same definition of Δx^a as in eq.(9.30).

The two mean velocities, to the future b^a, and from the past b^a_*, do not coincide. The connection between them is well known [Nelson 1966, 1985],

$$b^a_*(x,t) = b^a(x) - \frac{\sigma^2}{\tau}\partial^a \log\rho(x,t) \ , \tag{9.33}$$

where[10] $\partial^a = \delta^{ab}\partial_b$ and $\rho(x,t) = P(x)$. What might not be widely appreciated is that eq.(9.33) is a straightforward consequence of Bayes' theorem, eq. (9.29).

[10] From now on we will raise and lower indices with the Euclidean metric δ_{ab}.

(For a related idea see [Jaynes 1989].) To derive eq.(9.33) expand $P(x')$ about x in (9.29) to get

$$P(x|x') = \left[1 - \frac{\partial \log \rho(x,t)}{\partial x^b}\,\Delta x^b + \ldots\right] P(x'|x)\,. \tag{9.34}$$

Multiply $b^a_*(x')$ in eq.(9.32) by a smooth test function $f(x')$ and integrate,

$$\int dx'\, b^a_*(x') f(x') = \frac{1}{\Delta t}\int dx' \int dx\, P(x|x')\Delta x^a f(x')\ . \tag{9.35}$$

(The limit $\Delta t \to 0^+$ is understood.) On the right hand side expand $f(x')$ about x and use (9.34),

$$\frac{1}{\Delta t}\int dx' \int dx\, P(x'|x)[\Delta x^a f(x) - \Delta x^a \Delta x^b \frac{\partial \log \rho(x,t)}{\partial x^b} f(x) + \Delta x^a \Delta x^c \frac{\partial f}{\partial x^c} + \ldots]\,. \tag{9.36}$$

Next interchange the orders of integration and take $\Delta t \to 0^+$ using eq.(9.27),

$$\langle \Delta x^a \Delta x^b \rangle = \langle \Delta w^a \Delta w^b \rangle = \frac{\sigma^2}{\tau}\Delta t\, \delta^{ab}\ . \tag{9.37}$$

On integration by parts the third term of (9.36) vanishes and we get

$$\int dx\, b^a_*(x) f(x) = \int dx\, \left[b^a(x) - \frac{\sigma^2}{\tau}\delta^{ab}\partial_b \log \rho(x,t)\right] f(x)\,, \tag{9.38}$$

Since $f(x)$ is arbitrary we get (9.33).

The puzzle of the arrow of time has a long history (see *e.g.* [Price 1996; Zeh 2001]). The standard question is how can an arrow of time be derived from underlying laws of nature that are symmetric? Entropic dynamics offers a new perspective because it does not assume any underlying laws of nature — whether they be symmetric or not — and its goal is not to explain the asymmetry between past and future. The asymmetry is the inevitable consequence of entropic inference. From the point of view of entropic dynamics the challenge does not consist in explaining the arrow of time, but rather in explaining how it comes about that despite the arrow of time some laws of physics turn out to be reversible. Indeed, even when the derived laws of physics — in our case, the Schrödinger equation — turns out to be fully time-reversible, *entropic time itself only flows forward.*

9.5 Accumulating changes

Time has been introduced as a useful device to keep track of the accumulation of small changes. The technique to do this is well known from diffusion theory [Chandrasekhar 1943]. Small changes given by (9.25-9.27) accumulate according to the Fokker-Planck equation (FP) which we now proceed to derive.

9.5.1 Derivation of the Fokker-Planck equation

The result of building up a finite change from an initial time t_0 up to time t leads to the density

$$\rho(x,t) = \int dx_0\, P(x,t|x_0,t_0)\rho(x_0,t_0) \ , \tag{9.39}$$

where the finite-time transition probability, $P(x,t|x_0,t_0)$, is constructed by iterating the infinitesimal changes described in eq.(9.22),

$$P(x,t+\Delta t|x_0,t_0) = \int dz\, P(x,t+\Delta t|z,t)P(z,t|x_0,t_0) \ . \tag{9.40}$$

For small times Δt the distribution $P(x,t+\Delta t|z,t)$, given in eq. (9.24), is very sharply peaked at $x=z$. In fact, as $\Delta t \to 0$ we have

$$P(x,t+\Delta t|z,t) \to \delta(x-z) \ . \tag{9.41}$$

Such singular behavior cannot be handled directly by Taylor expanding in z about the point x. Instead one follows an indirect procedure. Multiply by a smooth test function $f(x)$ and integrate over x,

$$\begin{aligned}\int dx\, P(x,t+\Delta t|x_0,t_0)f(x) &= \int dx \int dz\, P(x,t+\Delta t|z,t)P(z,t|x_0,t_0)f(x) \\ &= \int dz \left[\int dx\, P(x,t+\Delta t|z,t)f(x)\right] P(z,t|x_0,t_0) \ .\end{aligned} \tag{9.42}$$

The test function $f(x)$ is assumed sufficiently smooth precisely so that it can be expanded about z. Then as $\Delta t \to 0$ the integral in the brackets, including all terms that contribute to order Δt, is

$$\begin{aligned}[\cdots] &= \int dx\, P(x,t+\Delta t|z,t)\left(f(z) + \frac{\partial f}{\partial z^a}(x^a - z^a) + \right. \\ &\quad \left. + \frac{1}{2}\frac{\partial^2 f}{\partial z^a \partial z^b}(x^a - z^a)(x^b - z^b) + \ldots\right) \\ &= f(z) + \Delta t\, b^a(z)\frac{\partial f}{\partial z^a} + \frac{1}{2}\Delta t\, \frac{\sigma^2}{\tau}\delta^{ab}\frac{\partial^2 f}{\partial z^a \partial z^b} + \ldots\end{aligned} \tag{9.43}$$

where we used eqs.(9.25-9.27),

$$\begin{aligned}\lim_{\Delta t\to 0^+} \frac{1}{\Delta t}\int dx\, P(x,t+\Delta t|z,t)(x^a - z^a) &= b^a(z) \ , \\ \lim_{\Delta t\to 0^+} \frac{1}{\Delta t}\int dx\, P(x,t+\Delta t|z,t)(x^a - z^a)(x^b - z^b) &= \frac{\sigma^2}{\tau}\delta^{ab} \ .\end{aligned} \tag{9.44}$$

Substituting (9.43) into the right hand side of (9.42),

$$\int dz \left[f(z) + \Delta t\, b^a(z)\frac{\partial f}{\partial z^a} + \frac{1}{2}\Delta t\, \frac{\sigma^2}{\tau}\delta^{ab}\frac{\partial^2 f}{\partial z^a \partial z^b}\right] P(z,t|x_0,t_0) \ , \tag{9.45}$$

dividing by Δt, and integrating by parts, (9.42) becomes

$$\int dx \, \frac{1}{\Delta t} \left[P(x, t+\Delta t|x_0, t_0) - P(x, t|x_0, t_0) \right] f(x) = \\ \int dx \left[-\frac{\partial}{\partial x^a} \left(b^a(x) P(x, t|x_0, t_0) \right) + \frac{\sigma^2}{2\tau} \nabla^2 P(x, t|x_0, t_0) \right] f(x) \,, \tag{9.46}$$

where $\nabla^2 = \delta^{ab} \partial^2 / \partial x^a \partial x^b$. Finally we let $\Delta t \to 0$ and since $f(x)$ is arbitrary we get the Fokker-Planck equation for the finite-time transition probability,

$$\frac{\partial}{\partial t} P(x, t|x_0, t_0) = -\frac{\partial}{\partial x^a} \left(b^a(x) P(x, t|x_0, t_0) \right) + \frac{\sigma^2}{2\tau} \nabla^2 P(x, t|x_0, t_0) \,. \tag{9.47}$$

The corresponding evolution equation for the density $\rho(x, t)$, which is what we will henceforth call the Fokker-Planck equation (FP), is obtained differentiating eq.(9.39) with respect to t,

$$\frac{\partial \rho(x, t)}{\partial t} = \int dx_0 \, \frac{\partial P(x, t|x_0, t_0)}{\partial t} \rho(x_0, t_0) \,. \tag{9.48}$$

Using eqs.(9.47) and (9.39),

$$\frac{\partial \rho(x, t)}{\partial t} = \int dx_0 \, \left[-\frac{\partial}{\partial x^a} \left(b^a P(x, t|x_0, t_0) \right) + \frac{\sigma^2}{2\tau} \nabla^2 P(x, t|x_0, t_0) \right] \rho(x_0, t_0) \,, \tag{9.49}$$

leads to the FP equation,

$$\partial_t \rho = -\partial_a \left(b^a \rho \right) + \frac{\sigma^2}{2\tau} \nabla^2 \rho \,, \tag{9.50}$$

where $\partial_a = \partial / \partial x^a$ and $\nabla^2 = \delta^{ab} \partial^2 / \partial x^a \partial x^b$.

9.5.2 The current and osmotic velocities

The FP equation can be rewritten as a continuity equation,

$$\partial_t \rho = -\partial_a \left(v^a \rho \right) \,, \tag{9.51}$$

where the velocity of the probability flow or *current velocity* is

$$v^a = b^a - \frac{\sigma^2}{2\tau} \delta^{ab} \frac{\partial_a \rho}{\rho} \,. \tag{9.52}$$

It is convenient to introduce the *osmotic velocity*

$$u^a \stackrel{\text{def}}{=} -\frac{\sigma^2}{\tau} \partial^a \log \rho^{1/2} \,, \tag{9.53}$$

so that

$$v^a = b^a + u^a. \tag{9.54}$$

The interpretation is straightforward: The drift b^a represents the tendency of the probability ρ to flow up the entropy gradient while u^a represents the tendency to flow down the density gradient. The situation is analogous to Brownian motion where the drift velocity is the response to the gradient of an external potential, while u^a is a response to the gradient of concentration or chemical potential—the so-called *osmotic force*.[11] The osmotic contribution to the probability flow is the actual diffusion current,

$$\rho u^a = -\frac{\sigma^2}{2\tau}\partial^a\rho \ , \tag{9.55}$$

which can be recognized as Fick's law, with a diffusion coefficient given by $\sigma^2/2\tau$.

Since both the future drift b^a and the osmotic velocity u^a are gradients, it follows that the current velocity is a gradient too. For later reference, from (9.26) and (9.53),

$$v^a = \frac{\sigma^2}{\tau}\,\partial^a\phi \ , \tag{9.56}$$

where

$$\phi(x,t) = S(x) - \log\rho^{1/2}(x,t). \tag{9.57}$$

With these results entropic dynamics reaches a certain level of completion: We figured out what small changes to expect — they are given by $P(x'|x)$ — and time was introduced to keep track of how these small changes accumulate; the net result is diffusion according to the FP equation (9.50).

9.6 Non-dissipative diffusion

But quantum mechanics is not *just* diffusion. The description so far has led us to the density $\rho(x,t)$ as the important dynamical object but to construct a wave function, $\Psi = \rho^{1/2}e^{i\phi}$, we need a second degree of freedom, the phase ϕ. The problem is that as long as the geometry of the statistical manifold $\mathcal{M}$ is rigidly fixed there is no logical room for additional degrees of freedom. Note that the function ϕ introduced in eqs.(9.56) and (9.57) does not represent an independent degree of freedom. The natural solution is to remove this constraint. We can take $S(x,t)$ to be the new independent degree of freedom but eq.(9.56) suggests that a more convenient and yet equivalent choice is

$$\phi(x,t) = S(x,t) - \log\rho^{1/2}(x,t) \ . \tag{9.58}$$

Thus the dynamics will consist of the coupled evolution of $\rho(x,t)$ and $\phi(x,t)$.

[11] The definition of osmotic velocity adopted in [Nelson 1966] and other authors differs from ours by a sign. Nelson takes the osmotic velocity to be the velocity imparted by an external force that is needed to balance the osmotic force (due to concentration gradients) in order to attain equilibrium. For us the osmotic velocity is the velocity imparted by the osmotic force itself.

To specify the dynamics of the manifold $\mathcal{M}$ we follow [Nelson 1979] and impose that the dynamics be non-dissipative, that is, we require the conservation of a certain functional $E[\rho, S]$ to be specified below that we will proceed to call the "energy". Thus, we make the following assumption:

> *The statistical manifold $\mathcal{M}$ participates in the dynamics: the particles react back on the y variables so that their entropy $S(x,t)$ becomes a time-dependent field. The dynamics between $\rho(x,t)$ and $S(x,t)$ is such that there is a conserved energy,* $E[\rho, S] = \text{constant}$.

At first sight it might appear that imposing that some energy $E[\rho, S]$ be conserved is natural because it agrees with our classical preconceptions of what physics ought to be like. But classical intuitions are not a good guide here. In the more sophisticated approaches to physics energy is taken to be whatever happens to be conserved as a result of invariance under translations in time. But our dynamics has hardly been defined yet; what, then, is "energy" and why should it be conserved in the first place? Furthermore, if we go back to eq.(9.25) we see that it is the kind of equation (a Langevin equation) that characterizes a Brownian motion in the limit of *infinite* friction. Thus, the explanation of quantum theory in terms of a sub-quantum classical mechanics would require that particles be subjected to infinite friction and suffer zero dissipation at the same time. Such a strange sub-quantum mechanics could hardly be called 'classical'.

Remark: In the 19th century the effort to model the wave properties of light led Fresnel and other researchers to postulate an underlying medium, the ether. In order to acccount for the purely transverse nature of light polarization the ether had to have very unusual and contradictory properties. It had to be simultaneously infinitely rigid to prevent any longitudinal polarization and also infinitely tenuous to allow the free motion of material bodies. Maxwell's contribution was to produce a model of light that focused attention elsewhere — light is just a peculiar configuration of electromagnetic fields. By ignoring the ether Maxwell demonstrated that it was effectively superfluous, which eventually led to its being altogether discarded by Einstein. The situation with a sub-quantum stochastic mechanics is somewhat analogous: it is assumed that quantum fluctuations are caused by some physical sub-quantum agent, either Nelson's background field [Nelson 1985] or Smolin's hidden variables [Smolin 2006]. Whatever this unusual agent might be it must simultaneously allow for both infinite friction and zero dissipation. On the other hand, entropic dynamics, being a purely epistemic model, is silent on the matter of whether there is some physical agent causing the particles to fluctuate. What fluctuates in ED are not necessarily the particles themselves — for all we know, they might — but our beliefs about where the particles are most likely to be found.

9.6.1 Manifold dynamics

The energy functional $E[\rho, S]$ is chosen to be the expectation of a *local* "energy" function $\varepsilon(x,t)$, that is,

$$E[\rho, S] = \int dx\, \rho(x,t)\, \varepsilon(x,t) \ , \tag{9.59}$$

where $\varepsilon(x,t)$ depends on $\rho(x,t)$ and $S(x,t)$ and their derivatives. This is more conveniently expressed in terms of $\phi(x,t)$,

$$\varepsilon(x,t) = \varepsilon(\rho, \partial\rho, \phi, \partial\phi; x) \ . \tag{9.60}$$

The specific form of ε is chosen to be invariant under time reversal [Smolin 2006]. Under time reversal $t \to -t$ we have

$$b^a \to -b^a_* \,, \quad v^a \to -v^a \,, \quad u^a \to u^a \,. \tag{9.61}$$

If we require that the velocities enter in rotationally invariant terms, then for low velocities we need only include velocity terms in v^2 and u^2, therefore

$$\varepsilon(\rho, \partial\rho, \phi, \partial\phi; x) = A\gamma_{ab}v^a v^b + B\gamma_{ab}u^a u^b + V(x) \ , \tag{9.62}$$

where A and B are constants, γ_{ab} is given by (9.1), and $V(x)$ represents an external potential. If ε has units of energy then A/σ^2 and B/σ^2 have units of mass. Let us define new constants

$$m = \frac{2A}{\sigma^2} \quad \text{and} \quad \mu = \frac{2B}{\sigma^2} \ , \tag{9.63}$$

which we will call the "current mass" and the "osmotic mass". Then

$$\varepsilon = \frac{1}{2}mv^2 + \frac{1}{2}\mu u^2 + V(x) \ . \tag{9.64}$$

It is further convenient to combine the constant τ, which sets the units of time, with A into yet a new constant η,

$$\eta = \frac{2A}{\tau} \quad \text{so that} \quad \frac{\sigma^2}{\tau} = \frac{\eta}{m} \ . \tag{9.65}$$

η relates the units of mass or energy with those of time. Then the current and osmotic velocities, eqs.(9.56) and (9.53) are

$$mv_a = \eta\, \partial_a\phi \quad \text{and} \quad mu_a = -\eta\partial_a \log\rho^{1/2} \ , \tag{9.66}$$

and the energy (9.62) becomes

$$\varepsilon = \frac{\eta^2}{2m}(\partial_a\phi)^2 + \frac{\mu\eta^2}{8m^2}(\partial_a \log\rho)^2 + V \ . \tag{9.67}$$

Remark: This energy is unusual in several respects. First and foremost, note that unlike classical mechanics the energy ε is not a property of the particle.

It explicitly involves ρ and S and thus ε is rather an epistemic concept; it is property of our state of knowledge. This is a topic to which we will return at some length in the next chapter: in the ED model the particles are understood to have actual positions but other quantities such as energy or momentum are not attributes of the particles themselves; they are epistemic constructs. Second, as seen in both eqs.(9.64) and (9.67) beyond the analogues of classical kinetic and potential energies the energy ε contains an extra term, the osmotic or "quantum" potential.

When the potential is static, $\dot{V} = 0$, energy is conserved, $\dot{E} = 0$. Otherwise we impose that energy increase at the rate

$$\dot{E} = \int dx\, \rho \dot{V} \ . \tag{9.68}$$

Next take the time derivative of (9.67)

$$\frac{dE}{dt} = \int d^3x \left[\frac{\partial(\rho\varepsilon)}{\partial\rho}\dot{\rho} + \frac{\partial(\rho\varepsilon)}{\partial(\partial_a\rho)}\partial_a\dot{\rho} + \rho\frac{\partial\varepsilon}{\partial\phi}\dot{\phi} + \rho\frac{\partial(\varepsilon)}{\partial(\partial_a\phi)}\partial_a\dot{\phi} \right] \ . \tag{9.69}$$

Use

$$\frac{\partial\varepsilon}{\partial\phi} = 0 \quad \text{and} \quad \frac{\partial(\varepsilon)}{\partial(\partial_a\phi)} = \eta v^a, \tag{9.70}$$

and integrate by parts to get,

$$\frac{dE}{dt} = \int d^3x \left(\frac{\partial(\rho\varepsilon)}{\partial\rho} - \partial_a \frac{\partial(\rho\varepsilon)}{\partial(\partial_a\rho)} + \eta\dot{\phi} \right) \dot{\rho} \ . \tag{9.71}$$

Now, any instant t can be taken as the initial instant for evolution into the future. We impose that the energy E be conserved for any arbitrary choice of initial conditions, namely $\rho(x,t)$ and $\phi(x,t)$, which implies an arbitrary choice of $\dot{\rho}$. Therefore,

$$\eta\dot{\phi} + \frac{\partial(\rho\varepsilon)}{\partial\rho} - \partial_a \frac{\partial(\rho\varepsilon)}{\partial(\partial_a\rho)} = 0 \ . \tag{9.72}$$

Substituting eq.(9.67) for ε we get,

$$\dot{\phi} + \varepsilon - \frac{\mu\eta^2}{4m^2}\frac{\nabla^2\rho}{\rho} = 0 \ , \tag{9.73}$$

or,

$$\eta\dot{\phi} + \frac{\eta^2}{2m}(\partial_a\phi)^2 + V - \frac{\mu\eta^2}{2m^2}\frac{\nabla^2\rho^{1/2}}{\rho^{1/2}} = 0 \ . \tag{9.74}$$

While the continuity equation, eq.(9.51) with (9.56), is

$$\dot{\rho} = -\partial_a\left(\rho v^a\right) = -\frac{\eta}{m}\partial^a\left(\rho\partial_a\phi\right) = -\frac{\eta}{m}\left(\partial^a\rho\partial_a\phi + \rho\nabla^2\phi\right) \ , \tag{9.75}$$

Equations (9.74) and (9.75) are the coupled dynamical equations we seek. They describe entropic diffusion and energy conservation. The evolution of $\phi(x,t)$,

eq.(9.74), is determined by $\rho(x,t)$; and the evolution of $\rho(x,t)$, eq.(9.75), is guided by $\phi(x,t)$. The evolving geometry of the manifold $\mathcal{M}$ enters through $\phi(x,t)$.

Incidentally, taking the expectation of eq.(9.73) and integrating by parts and discarding surface terms at infinity leads to a particularly elegant expression for the energy,

$$E = \int d^3x\, \rho\varepsilon = -\int d^3x\, \rho\dot{\phi} \ . \tag{9.76}$$

9.6.2 Classical limits

Before proceeding further we note that writing $S_{HJ} = \eta\phi$ in equations (9.66) and (9.74) and taking the limit $\eta \to 0$ with S_{HJ}, m, and μ fixed leads to

$$mv_a = \partial_a S_{HJ} \quad \text{and} \quad u_a = 0 \ , \tag{9.77}$$

and to

$$\dot{S}_{HJ} + \frac{1}{2m}(\partial_a S_{HJ})^2 + V = 0 \ , \tag{9.78}$$

which is identical with the Hamilton-Jacobi equation — this is the equation of motion in the Hamilton-Jacobi formulation of classical mechanics (see *e.g.* [Landau Lifshitz 1993]). The particle's energy and momentum are given by

$$E = -\frac{\partial S_{HJ}}{\partial t} \quad \text{and} \quad p_a = \frac{\partial S_{HJ}}{\partial x^a} \ , \tag{9.79}$$

which suggests that the constant m be interpreted as the inertial mass. Furthermore, eq.(9.26) and (9.54) with $u_a = 0$ tell us that $S_{HJ} \to \eta S$ so that,

> *Up to a proportionality constant the Hamilton-Jacobi function $S_{HJ}(x,t)$ is the entropy $S(x,t)$ of the y variables.*

Thus, the classical particle is expected to move along the entropy gradient. Furthermore, eq.(9.27),

$$\langle \Delta w^a \rangle = 0 \quad \text{and} \quad \left\langle \Delta w^a \Delta w^b \right\rangle = \frac{\eta}{m}\Delta t\, \delta^{ab} \to 0 \ , \tag{9.80}$$

says that the fluctuations about the expected trajectory vanish. We conclude that

> *In the limit $\eta \to 0$ entropic dynamics reproduces classical mechanics with classical trajectories following the gradient of the entropy $S(x,t)$ of the y variables.*

A similar classical limit can also be attained for fixed η provided the mass m is sufficiently large. In ED this effect can be already seen at the level of the short step transition probability, eq.(9.24). A more realistic example (see

[Johnson 2011]) involves a macrocopic body composed of N particles of masses m_n, $n = 1 \dots N$. Using eq.(9.65) the short step transition for N particles is

$$P(x'|x) = \frac{1}{Z_N} \exp\left[-\sum_{n=1}^{N} \frac{m_n}{2\eta\Delta t} \delta_{ij} \left(\Delta x_n^i - \Delta \bar{x}_n^i\right) \left(\Delta x_n^j - \Delta \bar{x}_n^j\right) \right] , \tag{9.81}$$

where $i, j = 1, 2, 3$. We are interested in the motion of the center of mass,

$$X^i = \frac{1}{M} \sum_{n=1}^{N} m_n x^i \quad \text{where} \quad M = N\bar{m} = \sum_{n=1}^{N} m_n \ . \tag{9.82}$$

The probability that the center of mass goes from X to $X' = X + \Delta X$ is

$$P(X'|X) = \int d^{3N}x' \, P(x'|x) \delta \left(\Delta X - \frac{1}{M} \sum_n m_n \Delta x^i \right) \ . \tag{9.83}$$

This integral is a straighforward instance of the central limit theorem (see section 2.8.2). The result is

$$P(X'|X) \propto \exp\left[-\frac{M}{2\eta\Delta t} \delta_{ij} \left(\Delta X^i - \Delta \bar{X}^i\right) \left(\Delta X^j - \Delta \bar{X}^j\right) \right] , \tag{9.84}$$

where, using eq.(9.26), the expected step $\Delta \bar{X}^i$ is

$$\Delta \bar{X}^i = \frac{1}{M} \sum_{n=1}^{N} m_n \Delta \bar{x}^i = \frac{\eta \Delta t}{N\bar{m}} \sum_{n=1}^{N} \frac{\partial S}{\partial x_n^i} \ . \tag{9.85}$$

The actual displacement is

$$\Delta X^i = \Delta \bar{X}^i + \Delta W^i \ , \tag{9.86}$$

with fluctuations ΔW^i such that

$$\left\langle \Delta W^i \Delta W^j \right\rangle = \frac{\eta}{N\bar{m}} \Delta t \, \delta^{ij} \ . \tag{9.87}$$

Therefore whereas the expected drift $\Delta \bar{X}^i$ is of order N^0 (because the N terms in the sum offset the $1/N$) the fluctuations are of order $N^{-1/2}$. For large N the fluctuations become negligible and the body follows the the smooth trajectory predicted by classical mechanics. Indeed, set $\eta S = S_{HJ}$ and eq.(9.85) becomes

$$M \frac{d\bar{X}^i}{dt} = \sum_{n=1}^{N} \frac{\partial S_{HJ}}{\partial x_n^i} = \frac{\partial S_{HJ}}{\partial X^i} \ . \tag{9.88}$$

The limit $\mu \to 0$ for fixed η, S_{HJ}, and m is also interesting. This situation is also ruled by the classical Hamilton-Jacobi equation (9.78), but the osmotic velocity does not vanish,

$$m v_a = \partial_a S_{HJ} \quad \text{and} \quad m u_a = \eta \partial_a \log \rho^{1/2} \ . \tag{9.89}$$

The expected trajectory also lies along a classical path but now, however, it does not coincide with the entropy gradient. More important perhaps is the fact that the fluctuations Δw^a about the classical trajectory do not vanish. *The limit $\mu \to 0$ is a different "classical" limit* whether it corresponds to an actual physical situation remains to be seen. We will briefly return to this topic it in the next chapter.

9.6.3 The Schrödinger equation

Next we show that, with one very interesting twist, the dynamical equations (9.75) and (9.74) turn out to be equivalent to the Schrödinger equation. We can always combine the functions ρ and ϕ into a complex function

$$\Psi = \rho^{1/2} \exp(i\phi) \ . \tag{9.90}$$

Computing its time derivative,

$$\dot{\Psi} = \left(\frac{\dot{\rho}}{2\rho} + i\dot{\phi} \right) \Psi \ , \tag{9.91}$$

and using eqs. (9.75) and (9.74) leads (after some considerable but straightforward algebra) to

$$i\eta\dot{\Psi} = -\frac{\eta^2}{2m}\nabla^2\Psi + V\Psi + \frac{\eta^2}{2m}\left(1 - \frac{\mu}{m}\right)\frac{\nabla^2(\Psi\Psi^*)^{1/2}}{(\Psi\Psi^*)^{1/2}}\Psi \ . \tag{9.92}$$

This reproduces the Schrödinger equation,

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\Psi + V\Psi \ , \tag{9.93}$$

provided the current and osmotic masses are equal, $m = \mu$, and η is identified with Planck's constant, $\eta = \hbar$.

But why should the osmotic mass be precisely equal to the inertial mass? Why can't we say that entropic dynamics predicts a non-linear generalization of quantum theory? This question is so central to quantum theory that we devote the next section to it. But before that we note that the non-linearity is undesirable both for experimental and theoretical reasons. On one hand, various types of non-linearities have been ruled out experimentally to an extreme degree through precision experiments on the Lamb shift [Smolin 1986a] and even more so in hyperfine transitions [Bollinger 1989]. On the other hand, from the theory side it is the fact that time evolution preserves linear superpositions that leads to the superposition principle and makes Hilbert spaces useful. In addition, there is a consistency argument that links the linearity of the Hilbert space and the linearity of time evolution [Caticha 1998]. Retaining one and not the other leads to inconsistently assigned amplitudes showing that the very concept of quantum amplitudes is a reflection of linearity. And, as if that were not enough, it has also been shown that in the presence of non-linear terms entangled particles could be used to achieve superluminal communication [Gisin 1990]. Therefore it is extremely probable that the identity of inertial and osmotic mass is exact.

There is another mystery in quantum theory — the central role played by complex numbers — that turns out to be related to these issues. The dynamical equations (9.75) and (9.74) contain no complex numbers. It is true that they contain two degrees of freedom ρ and ϕ and that these two quantities can be combined into a single complex number $\Psi = \rho^{1/2}e^{i\phi}$ but this is a triviality,

not a mystery: the dynamical equations can always be reformulated into an equation for Ψ *and its conjugate* Ψ^*. The statement that complex numbers play a fundamental role in quantum theory is the non-trivial assertion that the equation of evolution contains *only* Ψ and not Ψ and its conjugate Ψ^*. In the entropic approach both the linear time evolution and the special role of complex numbers are linked through the equality $m = \mu$.

9.7 A quantum equivalence principle

The generalization to N particles is straightforward. As indicated at the end of section 9.1, the configuration space has $3N$ dimensions and the system is represented by a point $x = x^A = (x^{a_1}, x^{a_2} \ldots)$. The corresponding Fokker-Planck equation [see eqs.(9.3), (9.51) and (9.56)] is

$$\partial_t \rho = -\frac{1}{\tau}\gamma^{AB}\partial_A(\rho\partial_B\phi) = -\sum_{n=1}^{N}\partial_{a_n}(\rho v_n^{a_n}) \tag{9.94}$$

where $\phi(x,t)$ is given by eq.(9.58). The current and osmotic velocities are

$$v_n^{a_n} = \frac{\sigma_n^2}{\tau}\,\partial^{a_n}\phi \quad \text{and} \quad \mu_n^{a_n} = -\frac{\sigma_n^2}{\tau}\,\partial^{a_n}\log\rho^{1/2}\ , \tag{9.95}$$

and the conserved energy is

$$E = \int d^{3N}x\,\rho(x,t)\left(A\gamma_{AB}v^Av^B + B\gamma_{AB}u^Au^B + V(x)\right)\ . \tag{9.96}$$

Introducing the inertial (or current) and osmotic masses,

$$m_n = \frac{2A}{\sigma_n^2} \quad \text{and} \quad \mu_n = \frac{2B}{\sigma_n^2}\ , \tag{9.97}$$

and the constant $\eta = 2A/\tau$, eqs.(9.94) and (9.96) become

$$\partial_t\rho = -\sum_n\frac{\eta}{m_n}\partial_{a_n}(\rho\partial^{a_n}\phi)\ , \tag{9.98}$$

$$E[\rho,S] = \int d^{3N}x\,\rho\left(\sum_n[\frac{\eta^2}{2m_n}(\partial_{a_n}\phi)^2 + \frac{\mu_n\eta^2}{8m_n^2}(\partial_{a_n}\log\rho)^2] + V(x)\right)\ . \tag{9.99}$$

Imposing, as before, that $\dot{E} - \int\rho\dot{V} = 0$ for arbitrary choices of $\dot{\rho}$ leads to the modified Hamilton-Jacobi equation,

$$\eta\dot{\phi} + \sum_n[\frac{\eta^2}{2m_n}(\partial_{a_n}\phi)^2 - \frac{\mu_n\eta^2}{2m_n^2}\frac{\nabla_n^2\rho^{1/2}}{\rho^{1/2}}] + V = 0\ . \tag{9.100}$$

Finally, the two eqs.(9.98) and (9.100) can be combined into a single equation for the complex wave function, $\Psi = \rho^{1/2}e^{i\phi}$,

$$i\eta\dot{\Psi} = \sum_n\frac{-\eta^2}{2m_n}[\nabla_n^2 - \left(1 - \frac{\mu_n}{m_n}\right)\frac{\nabla_n^2(\Psi\Psi^*)^{1/2}}{(\Psi\Psi^*)^{1/2}}]\Psi + V\Psi\ . \tag{9.101}$$

Eq.(9.97) shows that the ratio of osmotic to inertial mass turns out to be a universal constant, the same for all particles: $\mu_n/m_n = B/A$. This can be traced to a choice of energy, eq.(9.96), that reflects the translational and rotational symmetries of the configuration space. But why should $\mu_n = m_n$ *exactly*? To see this we go back to eq.(9.99). We can always change units and rescale η and τ by some constant κ into $\eta = \kappa\eta'$, $\tau = \tau'/\kappa$. If we also rescale ϕ into $\phi = \phi'/\kappa$, then eqs.(9.98) and (9.99) become

$$\partial_t \rho = -\sum_n \frac{\eta'}{m_n} \partial_{a_n} (\rho \partial^{a_n} \phi') \ , \tag{9.102}$$

$$E[\rho, S] = \int d^{3N}x \, \rho \left(\sum_n [\frac{\eta'^2}{2m_n} (\partial_{a_n} \phi')^2 + \frac{\mu_n \kappa^2 \eta'^2}{8m_n^2} (\partial_{a_n} \log \rho)^2] + V \right) \ . \tag{9.103}$$

Following the same steps that led to eq.(9.101), we can introduce a *different* wave function $\Psi' = \rho^{1/2} \exp(i\phi')$ which satisfies

$$i\eta' \dot{\Psi}' = \sum_n \frac{-\eta'^2}{2m_n} [\nabla_n^2 - \left(1 - \frac{\mu_n \kappa^2}{m_n}\right) \frac{\nabla_n^2 (\Psi' \Psi'^*)^{1/2}}{(\Psi' \Psi'^*)^{1/2}}] \Psi' + V\Psi' \ . \tag{9.104}$$

Since the mere rescaling by κ can have no physical implications the different "regraduated" theories are all equivalent and it is only natural to use the simplest one: we choose $\kappa = (A/B)^{1/2}$ so that $\mu_n \kappa^2 = m_n$ and we can rescale the old μ_n to a new osmotic mass $\mu'_n = \mu_n \kappa^2 = m_n$.

The net result is that the non-linear terms drop out. Dropping the prime on Ψ' and identifying the rescaled value η' with Planck's constant $\hbar$, leads to the linear Schrödinger equation,

$$i\hbar \dot{\Psi} = \sum_n \frac{-\hbar^2}{2m_n} \nabla_n^2 \Psi + V\Psi \ . \tag{9.105}$$

We conclude that *for any positive value of the original coefficients μ_n it is always possible to regraduate η, ϕ and μ_n to a physically equivalent but more convenient description where the Schrödinger equation is linear and complex numbers attain a special significance.* From this entropic perspective the linear superposition principle and the complex Hilbert spaces are important because they are convenient, but not because they are fundamental — a theme that was also explored in [Caticha 1998].

These considerations remind us of Einstein's original argument for the equivalence principle: He proposed the complete physical equivalence of a gravitational field with the corresponding acceleration of the reference frame because this offers a natural explanation of the equality of inertial and gravitational masses and opens the door to an explanation of gravity in purely geometrical terms.

Similarly, in the quantum case *we propose the complete equivalence of quantum and statistical fluctuations because this offers a natural explanation of the Schrödinger equation — its linearity, its unitarity, the role of complex numbers, the equality of inertial and osmotic masses. Furthermore, it opens the door to an explanation of quantum theory as an example of statistical inference.*

9.8 Entropic time *vs.* physical time

Now that the dynamics has been more fully developed we should revisit the question of time. Entropic time has turned out to be useful in ordering the inferential sequence of small changes but it is not at all clear that this order has anything to do with the order relative to a presumably more fundamental "physical" time. If so, why does entropic time deserve to be called 'time' at all?

The answer is that the systems we are typically concerned with include, in addition to the particles of interest, also another system that one might call the "clock". The goal is to make inferences about correlations among the particles themselves and with the various states of the clock. Whether the inferred sequence of states of the particle-clock composite agrees with the order in "physical" time or not turns out to be quite irrelevant. It is only the correlations among the particles and the clock that are observable and not their "absolute" order.

This is an idea that demands a more explicit discussion. Here we show how it gives rise to the notion of simultaneity that turned out to be central to our definition of an instant in section 9.3.1.

Consider a single particle. From the probability of a single step, eq.(9.10) or (9.17), we can calculate the probability of any given sequence of (short) steps $\{x, x_1, \ldots, x_n, \ldots\}$. Since the path is an ordered sequence of events when two events lie on the same path it is meaningful to assert that one is earlier (in the entropic time sense) than the other: x_n is earlier than x_{n+1}. The actual path, however, is uncertain: how do we compare possible events along different paths? We need a criterion that will allow us to decide whether an event x' reached along one path is earlier or later than another event x'' reached along a different path. This is where the clock comes in. The role of the clock can be played, for example, by a sufficiently massive particle. This guarantees that the clock follows a deterministic classical trajectory $x_C = \bar{x}_C(t)$ given by eqs.(9.78) and (9.80) and that it remains largely unaffected by the motion of the particle.

The idea is that when we compute the probability that, say, after n steps the particle is found at the point x_n we implicitly *assume* that its three coordinates x_n^1, x_n^2, and x_n^3 are attained *simultaneously*. This is part of our *definition* of an instant. We adopt the same *definition* for composite systems. In particular, for the particle-clock system, $x_n^A = (x_n^a, x_{Cn}^\alpha)$, the coordinates of the particle x_n^a ($a = 1, 2, 3$) are taken to be simultaneous with the remaining coordinates that describe the clock x_{Cn}^α ($\alpha = 4, 5, \ldots$). Thus, when we say that at the nth step the particle is at x_n^a while the clock is at x_{Cn}^α it is implicit that these positions are attained *at the same time*.

By "the time is t" we will just mean that "the clock is in its state $x_C = \bar{x}_C(t)$." We say that the possible event that the particle reached x' along one path is simultaneous with another possible event x'' reached along a different path when both are simultaneous with the same state $\bar{x}_C(t)$ of the clock: then we say that x' and x'' happen "at the same time t." This justifies using the distribution $\rho(x, t)$ as the definition of an instant of time.

In the end the justification for the assumptions underlying entropic dynamics

lies in experiment. The ordering scheme provided by entropic time allows one to predict correlations. Since these predictions, which are given by the Schrödinger equation, turn out to be empirically successful one concludes that nothing deeper or more "physical" than entropic time is needed. A similar claim has been made by J. Barbour in his relational approach to time in the context of classical dynamics [Barbour 1994].

9.9 Dynamics in an external electromagnetic field

Entropic dynamics is derived from the minimal assumptions that the y variables are intrinsically uncertain and that motion consists of a succession of short steps. These two pieces of information are taken into account through the two constraints (9.7) and (9.8). Special circumstances may however require additional constraints.

9.9.1 An additional constraint

Consider a single particle placed in an external field the action of which is to constrain the expected component of displacements along a certain direction represented by the unit covector $n_a(x)$. This effect is represented by the constraint

$$\langle \Delta x^a n_a(x) \rangle = C(x) \ , \tag{9.106}$$

where the spatial dependence of $C(x)$ reflects the non-uniform intensity of the external field. It is convenient to define the magnitude of the external field in terms of the effect it induces. Thus we introduce the external field

$$A_a(x) \propto \frac{n_a(x)}{C(x)} \tag{9.107}$$

and the constraint is

$$\langle \Delta x^a A_a(x) \rangle = C \ , \tag{9.108}$$

where C is some constant that reflects the strength of the coupling to A_a.

9.9.2 Entropic dynamics

The transition probability $P(x'|x)$ is that which maximizes the entropy $\mathcal{S}[P,Q]$ in (9.9) subject to the old constraints plus the new constraint (9.108). The result is

$$P(x'|x) = \frac{1}{\zeta(x,\alpha,\beta)} e^{S(x') - \frac{1}{2}\alpha\Delta\ell^2(x',x) - \beta\Delta x^a A_a(x)} \ , \tag{9.109}$$

where

$$\zeta(x,\alpha,\beta) = \int dx' \, e^{S(x') - \frac{1}{2}\alpha\Delta\ell^2(x',x) - \beta\Delta x^a A_a(x)} \ , \tag{9.110}$$

and the Lagrange multiplier β is determined from the constraint eq.(9.108),

$$\frac{\partial}{\partial\beta} \log\zeta(x,\alpha,\beta) = -C \ . \tag{9.111}$$

From here on the argument follows closely the previous sections. For large α the transition probability (9.109) can be written as

$$P(x'|x) \propto \exp\left[-\frac{m}{2\hbar\Delta t}\delta_{ab}\left(\Delta x^a - \Delta\bar{x}^a\right)\left(\Delta x^b - \Delta\bar{x}^b\right)\right] , \tag{9.112}$$

where we used (9.23), (9.65), and units have been regraduated to set $\eta = \hbar$. Therefore, the displacement Δx^a can be expressed in terms of a expected drift plus a fluctuation, $\Delta x^a = \Delta\bar{x}^a + \Delta w^a$, where

$$\langle \Delta x^a \rangle = \Delta\bar{x}^a = b^a \Delta t \quad \text{where} \quad b^a = \frac{\hbar}{m}\delta^{ab}[\partial_b S - \beta A_b] , \tag{9.113}$$

$$\langle \Delta w^a \rangle = 0 \quad \text{and} \quad \left\langle \Delta w^a \Delta w^b \right\rangle = \frac{\hbar}{m}\Delta t\, \delta^{ab} . \tag{9.114}$$

Once again, for short steps the dynamics is dominated by the fluctuations. The only difference is the replacement of ∂S by the gauge invariant combination $\partial S - \beta A$. Small changes accumulate according to the FP equation (9.51) but now the current velocity is no longer given by eq.(9.56) but rather by

$$v^a = \frac{\hbar}{m}(\,\partial^a\phi - \beta A^a) , \tag{9.115}$$

and the FP equation is

$$\dot{\rho} = -\partial_a\left(\rho v^a\right) = -\frac{\hbar}{m}\partial^a[\rho(\partial_a\phi - \beta A_a)] , \tag{9.116}$$

ϕ is still given by (9.58) and the osmotic velocity (9.53) remains unchanged.

The energy functional is the same as (9.62), but now v is given by eq.(9.115),

$$E = \int dx\, \rho\left(\frac{\hbar^2}{2m}(\partial_a\phi - \beta A_a)^2 + \frac{\hbar^2}{8m}(\partial_a \log\rho)^2 + V\right) , \tag{9.117}$$

where we set $\mu = m$ and $\eta = \hbar$.

It is simplest to start with static external potentials, $\dot{V} = 0$ and $\dot{A} = 0$, so that the energy is conserved, $\dot{E} = 0$. Just as before after taking the time derivative, integrating by parts, and imposing that $\dot{E} = 0$ for arbitrary choices of $\dot{\rho}$, we get

$$\hbar\dot{\phi} + \frac{\hbar^2}{2m}(\partial_a\phi - \beta A_a)^2 + V - \frac{\hbar^2}{2m}\frac{\nabla^2\rho^{1/2}}{\rho^{1/2}} = 0 . \tag{9.118}$$

Equations (9.116) and (9.118) are the coupled equations for ρ and ϕ that describe entropic dynamics in the external potential A_a.

Setting $S_{HJ} = \eta\phi$ and taking the classical limit $\hbar \to 0$ leads to the classical Hamilton-Jacobi equation in an external electromagnetic field showing that the Lagrange multiplier β plays the role of electric charge. More precisely,

$$\beta = \frac{e}{\hbar c} , \tag{9.119}$$

where e is the electric charge and c is the speed of light. Thus,

In entropic dynamics electric charge is a Lagrange multiplier that regulates the response to the external electromagnetic potential A_a.

(If desired we can further separate V into electric and non-electric components, $V = eA_0 + V'$, but this is not needed for our present purposes.)

As before, the Schrödinger equation results from combining the functions ρ and ϕ into the wave function, $\Psi = \rho^{1/2}\exp(i\phi)$. Computing the time derivative $\dot{\Psi}$ using eqs.(9.116) and (9.118) leads to the Schrödinger equation,

$$i\hbar\frac{\partial\Psi}{\partial t} = \frac{\hbar^2}{2m}(i\partial_a - \frac{e}{\hbar c}A_a)^2\Psi + V\Psi \ , \tag{9.120}$$

The derivation above assumed that energy is conserved, $\dot{E} = 0$, which is true when the external potentials are static, $\dot{V} = 0$ and $\dot{A} = 0$, but this limitation is easily lifted. For time-dependent potentials the relevant energy condition must take into account the work done by external sources: we require that the energy increase at the appropriate rate,

$$\dot{E} = \int dx\, \rho(\dot{V} + \frac{e}{c}\rho v^a \dot{A}_a) \ . \tag{9.121}$$

The net result is that equations (9.118) and (9.120) remain valid for time-dependent external potentials.

9.9.3 Gauge invariance

We have seen that in entropic dynamics the phase of the wave function receives a statistical interpretation, $\phi = S - \log\rho^{1/2}$. On the other hand, without any physical consequences, the phase can be shifted by an arbitrary amount,

$$\phi(x,t) \to \phi'(x,t) = \phi(x,t) + \beta\chi(x,t) \ , \tag{9.122}$$

provided the potential is transformed appropriately, $A_a \to A'_a = A_a + \partial_a\chi$. This raises several questions.

First, how is the statistical interpretation of ϕ affected by the possibility of gauge transformations? The straightforward answer is that ϕ reflects a combination of several effects — the y variables (through their entropy S), the osmotic effect of diffusion (through the density ρ), and the choice of potential (through the function χ) — but these separate contributions are not necessarily easy to disentangle. Indeed, eq.(9.113) for the drift velocity shows that the dynamics depends on S and on A only through the combination $\partial S - \beta A$. Therefore we can envision two situations that are informationally inequivalent: one agent assigns an entropy S and imposes a constraint $\langle\Delta x^a A_a\rangle = C$, while another agent assigns a different entropy S' and imposes a different constraint $\langle\Delta x^a A'_a\rangle = C$. Remarkably both reach exactly the same physical predictions provided the entropies and potentials are related by $S' = S + \beta\chi$ and $A' = A + \partial\chi$ where $\chi(x,t)$ is some arbitrary function. Thus local phase invariance can be interpreted as *local entropy invariance*.

There is another set of questions that were first raised by T. Wallstrom in the context of stochastic mechanics [Wallstrom 1989, 1994]. They are concerned with the single- or multi-valuedness of phases and wave functions. Wallstrom noted that when stochastic mechanics is formulated *à la* Nelson [Nelson 1966] the current velocity $\vec{v}$ is postulated to be the gradient of some locally defined function ϕ. Now, being a local gradient does not imply that $\vec{v}$ will also be a global gradient and therefore both the phases ϕ and their corresponding wave functions Ψ will, in general, be multi-valued — which is unsatisfactory. A possible way out is to formulate stochastic mechanics in terms of an action principle as in [Guerra Morato 1983]. Then the current velocity is indeed a global gradient and both phases and wave functions are single-valued. But this is a problem too: single-valued phases can be too restrictive and exclude physically relevant states. For example, the usual way to describe states with non-zero angular momentum is to use multi-valued phases (the azimuthal angle) while requiring that the corresponding wave functions remain single-valued. The conclusion is that stochastic mechanics does not lead to the same set of solutions as the Schrödinger equation; it either produces too many [Nelson 1966] or too few [Guerra Morato 1983].

Similar objections can be raised in entropic dynamics. What is most interesting — and this appears to have been forgotten — is that in the early days of quantum mechanics the founders faced exactly the same kind of question: Why should wave functions be single-valued? The answer we favor is essentially the same offered by Pauli in the context of standard quantum mechanics [Pauli 1939]. He suggested that the criterion for admissibility for wave functions is that they must form a basis for a representation of the transformation group (for example, the rotation group) that happens to be pertinent to the problem at hand. Pauli's criterion is extremely natural from the perspective of a theory of inference: in any physical situation symmetries constitute the most common and most obviously relevant pieces of information.

Let us be explicit. In entropic dynamics the entropy $S(x,t)$ and the probability density $\rho(x,t)$ are single-valued functions. Therefore, a natural choice is that the phase, $\phi = S - \log \rho^{1/2}$, be single-valued too. A situation with non-vanishing angular momentum can be handled by imposing an additional constraint. For example, one can use a single-valued phase and an appropriately chosen vector potential — which might perhaps be a pure gauge, $A_a = -\partial_a \chi$, where χ might possibly be multivalued. Alternatively, we can gauge the potential away to $A'_a = 0$ and use a multi-valued phase, $\phi' = S - \log \rho^{1/2} + \beta\chi$. Which of these two equivalent options is to be preferred depends on whether the goal is clarity of interpretation or simpler mathematics. As for the appropriate choice of potential, $A_a = -\partial_a \chi$, we adopt Pauli's criterion: the admissible wave functions — that is, the various functions (ρ, S, χ) that appear in the formalism — must form a basis for a representation of the pertinent symmetry group.

9.10 Is ED a hidden-variable model?

As we have seen a considerable part of quantum theory — perhaps all of it — can be derived using entropic methods provided we introduce these mysterious extra variables y. Should we think of entropic dynamics as a hidden-variable model?

There is a trivial sense in which the y variables are "hidden": they are not directly observable.[12] But being unobservable is not sufficient to qualify as a hidden variable. The original motivation behind attempts to construct hidden variable models was to explain or at least ameliorate certain aspects of quantum mechanics that clash with our classical preconceptions. For example,
(1) Indeterminism: Is ultimate reality random? Do the gods play dice?
(2) Non-classical mechanics: The secret wish is that a sub-quantum world will eventually be discovered where nature obeys essentially classical laws.
(3) Non-classical probabilities: It is often argued, for example, that classical probability fails in the double slit experiment.
(4) Non-locality: Realistic interpretations of the wave function often lead to the paradoxes as in wave function collapse and EPR correlations.

But the y variables address none of these problems. In the standard view quantum theory is considered an extension of classical mechanics — indeed, the subject is called quantum *mechanics* — and therefore deviations from causality demand an explanation. In the entropic view, on the other hand, *quantum theory is not mechanics; it is inference* — entropic inference is a framework designed to handle insufficient information. From the entropic perspective indeterminism requires no explanation. Uncertainty and probabilities are the norm; it is certainty and determinism that demand explanations.

In ED there is no underlying classical dynamics — as we saw both quantum and classical mechanics are derived. The peculiar non-classical effects associated with the wave-particle duality arise not so much from the y variables themselves but rather from the specific non-dissipative diffusion which leads to a Schrödinger equation. The important breakthrough here was Nelson's realization that diffusion phenomena could be much richer than previously expected — it can account for wave and interference effects.

It is the whole entropic framework — and not just the y variables — that is incompatible with the notion of quantum probabilities. From this perspective it makes as little sense to distinguish quantum from classical probabilities as it is would be to talk about economic or medical probabilities.

Finally, non-locality is not explained; it is rather accepted as the relevant information that is necessary for predictions and this information is incorporated from the start in much the same way that Schrödinger originally did it: by formulating the theory in configuration space.

Thus, in none of the problems above do the y variables play the role that hidden variables were meant to play. But the term 'hidden variable' has by now

[12]The y variables are not observable *at the current stage of development of the theory.* It may very well happen that once we learn where to look we will find that they have been staring us in the face all along.

also acquired a very technical meaning (see *e.g.* [Harrigan Spekkens 2010]) and we can ask whether the y variables are hidden in this more technical sense. The answer, as we argue below, is still no.

Quantum mechanics, in its usual formulation, stipulates rules (the Born rule and its variations involving density operators) that allow us to calculate the probability $p(k|\mu,\pi)$ that a system prepared according to a procedure π will yield an outcome k when subjected to a measurement of type μ. In hidden-variable models it is assumed that the measurement process reveals pre-existing properties of the system. A complete specification of all those properties amounts to specifying the actual "true" state of the system, which we will call the *ontic* state and will denote by λ. For example, in classical mechanics the ontic state λ is a point in phase space. In a hidden variable model λ may include some variables that can be observed and others that remain hidden.

It is possible that a precisely defined preparation procedure π is not sufficient to uniquely determine the ontic state λ — perhaps π is noisy. In such cases an agent will describe its uncertainty about λ through a distribution $p(\lambda|\pi)$. It is also possible that the ontic state λ might not uniquely determine the outcome of a measurement μ but only its probability, $p(k|\lambda,\mu)$. A straightforward application of the rules of probability theory implies that

$$\begin{aligned} p(k|\mu,\pi) &= \int d\lambda p(k,\lambda|\mu,\pi) \\ &= \int d\lambda p(k|\lambda,\mu,\pi)p(\lambda|\mu,\pi)\ . \end{aligned} \tag{9.123}$$

The basic assumptions in a hidden-variable model are two. The first is that the actual outcome of a measurement depends on the measurement μ that is being performed and on the actual state λ of the system and that, once we are given λ, the outcome does not depend on the previous preparation π: $p(k|\lambda,\mu,\pi) = p(k|\lambda,\mu)$. Thus, λ represents everything that we need to know for the purpose of future predictions. This is the ontological assumption: Conditional on λ the future k is independent of the past π. The second assumption is that distribution of λ depends on the past preparation π and not on what might later be or not be measured: $p(\lambda|\mu,\pi) = p(\lambda|\pi)$. Such variables λ would be non-contextual — their distribution is independent of the particular context of measurement μ. Then

$$p(k|\mu,\pi) = \int d\lambda p(k|\lambda,\mu)p(\lambda|\pi)\ . \tag{9.124}$$

In summary: In order for a model of QM to be a hidden-variable model it must prescribe (1) how to calculate the distribution of ontic states $p(\lambda|\pi)$ for every preparation π; (2) it must prescribe the probability of the outcomes k once the ontic state is given, $p(k|\lambda,\mu)$; and finally (3) the probability of the measurement outcomes $p(k|\mu,\pi)$ must agree with the prediction of quantum mechanics (the Born rule).

Examples: The "orthodox" interpretation is that the wave function ψ itself provides a complete description of reality. This is the absolute-ψ or the

complete-ψ model. Here $\lambda = \psi$ and ψ is something objective and "real" (by which we do not mean that it is any kind of material or even ethereal substance). There are no hidden variables except perhaps for the wave function itself — ψ is not directly observable, but it can be inferred from multiple measurements. Another example is the de Broglie-Bohm pilot wave theory. Here the ontic state is given by the particle positions x and by the pilot wave ψ that guides them, $\lambda = (x, \psi)$; the wave function ψ is the hidden variable. Of course, all these models have problems of their own that remain largely unsolved.

Now we can see once again that ED is not a hidden-variable model. We claim that ED is in fact a model for quantum mechanics: the probabilities $p(k|\mu, \pi)$ of outcomes k do indeed coincide with the predictions of quantum mechanics. (Earlier in this chapter we saw that this is true for position measurements, $|\psi(x)|^2 = \rho(x)$; in the next chapter the result will be generalized for quantities other than position.) In ED the ontic state consists of the positions x of the particles and the values of the y variables, $\lambda = (x, y)$. But entropic dynamics provides no prescription to calculate the distribution of the ontic state $p(\lambda|\pi) = p(x, y|\pi)$. Indeed, we can write

$$p(x, y|\pi) = p(x|\pi)p(y|x, \pi) = \rho(x|\pi)p(y|x) \ . \tag{9.125}$$

The distribution $\rho(x|\pi)$ following a preparation π can be calculated using $|\psi(x)|^2 = \rho(x)$ but $p(y|x)$ remains always unknown — what we can calculate is the entropy $S(x)$ of the y variables and not their actual distribution, $p(y|x)$. Furthermore, unlike a true hidden-variable model where if we only knew the ontic state we would know the distribution of experimental outcomes, ED provides no such prescription. What we need to know is the pair $\rho(x)$ and $S(x)$ and not the pair (x, y). In fact actual knowledge of intermediate values of (x, y) would lead to results in disagreement with QM. This is quite analogous to the disruption in the interference effects in a double-slit experiment when one knows which slit the particle goes through. Therefore the y variables are not hidden variables.

We conclude with a brief remark on the formal similarity between ED and both Nelson's stochastic mechanics [Nelson 1985] and the de Broglie-Bohm pilot wave theory [Bohm HiIley 1993, Holland 1993]. Setting $\hbar\phi = S_{HJ}$ all these three theories are described by the same continuity equation, eq.(9.75),

$$\dot{\rho} = -\partial_a \left(\rho v^a\right) \quad \text{where} \quad v^a = -\frac{1}{m}\partial^a S_{HJ} \ , \tag{9.126}$$

and a quantum Hamilton-Jacobi equation,

$$\dot{S}_{HJ} + \frac{1}{2m}(\partial_a S_{HJ})^2 + V + V_Q = 0 \quad \text{where} \quad V_Q = -\frac{\hbar^2}{2m}\frac{\nabla^2\rho^{1/2}}{\rho^{1/2}} \ , \tag{9.127}$$

which is the classical Hamilton-Jacobi equation supplemented by a "quantum' potential. Since this is the only term that contains $\hbar$ it is tempting to say that it is the quantum potential that is responsible for quantum behavior.

The fact that these theories share such close formal similarity is not at all surprising: If they are to reproduce the same Schrödinger equation it is

inevitable that at some point they have to start agreeing somewhere. The similarity, however, ends there.

Both Nelson's and Bohm's mechanics are meant to reflect "reality" and this raises the standard for what constitutes a satisfactory explanation. In stochastic mechanics particles follow Brownian trajectories. Their irregular motion is due to some underlying and as yet unidentified background random field. A critical difficulty is to construct a believable classical model that reproduces the non-local correlations induced by the quantum potential. Another difficulty is the requirement that the current velocity is the gradient of a scalar field. This assumption is introduced in a totally ad hoc manner and it turns out, by the way, that is not even universally true — as seen in eq.(9.115) it fails in the presence of electromagnetic fields.

In Bohmian mechanics the particles are supposed to follow smooth causal trajectories along the gradient of the phase. This is explained by accepting that the wave function is an objectively real entity that can actually push the particles around — it is not clear however why the particles do not react back.

Purely epistemic theories, however, carry a much lighter explanatory burden. One is not required to explain what it is that pushes the particles around; it is merely sufficient to admit ignorance and accept that the particles could go anywhere subject to a few very natural constraints — motion has to be continuous; there exists other stuff in the world (represented by the y variables) which eventually explains that the current velocity is the gradient of a scalar field; and the statistical manifold $\mathcal{M}$ does not have to remain frozen.[13] In ED the probability density ρ through its gradient — the osmotic "force" — has no causal power over the motion of the particles. Its role is purely epistemic: what it does influence is our beliefs about where the particles are most likely to be found.

9.11 Summary and Conclusions

Our goal has been to derive quantum theory as an example of entropic inference. The challenge is to develop a framework that clarifies the conceptual difficulties that have plagued quantum theory since its inception while still reproducing its undeniable experimental successes. This means that to the extent that what has been derived is quantum mechanics and not some other theory we should not expect predictions that deviate from those of the standard quantum theory — at least not in the non-relativistic regime discussed in this work. On the other hand, the motivation behind this whole program lies in the conviction that it is the clarification and removal of conceptual difficulties that will eventually allow us to extend physics to other realms — gravity, cosmology — where the status of quantum theory is more questionable.

[13]The particular form of the energy functional begs, admittedly, for a more satisfactory justification. We suspect, however, that a proper justification will require us to dig deeper — at the very least, relativity should be included in the picture.

The framework of entropic inference is of general applicability. Its application to any particular problem requires assumptions that specify the intended subject matter and those pieces of information that are considered relevant. The main assumptions can be summarized as follows:

(a) The goal is to predict the positions x of some point particles. Since the information available is limited we can at best obtain a probability distribution $\rho(x)$ in the configuration space $\mathcal{X}$. We assume that $\mathcal{X}$ is flat, and that it is isotropic or anisotropic depending on whether the particles are identical or not.

(b) We assume that the world includes other things in addition to the particles: these extra things are described by the y variables that can influence and in their turn can be influenced by the particles. The uncertainty in the values of y is described by distributions $p(y|x)$ in a statistical manifold $\mathcal{M}$. The theory is robust in the sense that its predictions are insensitive to most details about the y variables.

(c) We assume that large changes result from the accumulation of many successive short steps. The transition probability for a short step $P(x'|x)$ is found using the method of maximum entropy. This requires assumptions about the prior (which we take to be uniform) and constraints (that changes happen continuously and that after each short step the new $p(y'|x')$ remains within the same statistical manifold $\mathcal{M}$). The result is that the dynamics of the particles is driven by the entropy $S(x)$ of the extra variables.

(d) A notion of time is introduced in order to keep track of the accumulation of small changes. This requires assumptions about what constitutes an instant and about how time is constructed as a succession of such instants. The choice of interval between instants is a matter of convenience — we choose a notion of duration that reflects the translational symmetry of the configuration space. The result is that the distribution ρ evolves according to a Fokker-Planck equation.

(e) We assume that the particles react back and affect the entropy $S(x)$ of the extra variables in such a way that there is a conserved "energy" $E[\rho, S] = \text{const}$. The specifics of this interaction are described through the functional form of $E[\rho, S]$.

(f) Electromagnetic interactions are described by including an additional constraint on the expected displacement along a certain field $A_a(x)$.

No further assumptions are made. The statistical model is specified by several parameters, $\{\sigma_n^2, \tau, A, B, \beta_n\}$. The anisotropy of configuration space for non-identical particles is parametrized by σ_n^2 with $n = 1 \ldots N$; τ defines units of time; A and B parametrize the relative strengths of the current and osmotic

terms in the energy functional; and, finally, β_n are Lagrange multipliers associated to the constraints for motion in electromagnetic fields. These parameters can be suitably regraduated and combined with each other into the familiar set which includes the masses and charges of the particles and Planck's constant.

We conclude with a summary of our conclusions.

On epistemology *vs.* ontology: Quantum theory has been derived as an example of entropic dynamics. What we have is a model. Within this model the positions of the particles and the values of the y variables are meant to be real. Our "limited information about reality" is represented in the probabilities as they are updated to reflect the physically relevant constraints. The wave function ψ is fully epistemic — which means neither fully subjective nor fully objective.

Quantum non-locality: Entropic dynamics may appear classical because no "quantum" probabilities were introduced. But this is deceptive. Probabilities, in this approach, are neither classical nor quantum; they are merely tools for inference. Phenomena that would normally be considered non-classical, such as non-local correlations, emerge naturally from constraints in configuration space which include the osmotic or quantum potential terms in the energy functional.

On interpretation: Ever since Born the magnitude of the wave function $|\Psi|^2$ has received a statistical interpretation. Within the entropic dynamics approach the phase of the wave function is also recognized as a feature of purely statistical origin. When electromagnetic interactions are introduced the gauge invariance is interpreted as an invariance under local entropy transformations.

On dynamical laws: The principles of entropic inference form the backbone of this approach to dynamics. The requirement that an energy be conserved is an important piece of information (*i.e.*, a constraint) which will probably receive its full justification once a completely relativistic version of entropic dynamics is developed.

On time: The derivation of laws of physics as examples of inference requires an account of the concept of time. Entropic time is modelled as an ordered sequence of instants with the natural measure of duration chosen to simplify the description of motion. We argued that whether the entropic order agrees with an objective order in an external physical time turns out to be an empirically inaccessible question, and in this sense, the notion of a physical time is not needed. Most interestingly, the entropic model of time explains the arrow of time.

Equivalence principle: The derivation of the Schrödinger equation from entropic inference led to an interesting analogy with general relativity. The statistical manifold $\mathcal{M}$ is not a fixed background but actively participates in the dynamics.

Chapter 10

Topics in Quantum Theory

In the Entropic Dynamics (ED) framework quantum theory is derived as an application of the method of maximum entropy. In this chapter the immediate goal is to demonstrate that the entropic approach to quantum theory can prove its worth through the clarification and removal of conceptual difficulties. We will tackle three topics that are central to quantum theory: the quantum measurement problem, the introduction and interpretation of observables other than position, including momentum, and the corresponding uncertainty relations. The presentation follows closely the work presented in [Johnson Caticha 2011; Nawaz Caticha 2011]. More details can be found in [Johnson 2011; Nawaz 2012].

10.1 The quantum measurement problem

Quantum mechanics introduced several new elements into physical theory. One is indeterminism, another is the superposition principle embodied in both the linearity of the Hilbert space and the linearity of the Schrödinger equation. The founders faced the double challenge of locating the source of indeterminism and of explaining why straightforward consequences of the superposition principle are not observed in the macroscopic world. The quantum measurement problem embodies most of these questions.[1] One is the problem of macroscopic entanglement; another is the problem of definite outcomes. How does a measurement yield a definite outcome or how do events ever get to happen? Are the values of observables created during the act of measurement?

To illustrate the nature of the problem consider the following idealization, due to von Neumann, of the process of measurement.[2] A generic state $|\Psi\rangle$ of a quantum system $\mathcal{S}$ can be represented in a basis $\{|s_n\rangle\}$ that spans the Hilbert

[1] A clear formulation of the problem is [Wigner 1963]; see also [Ballentine 1998]. Modern reviews with references to the literature appear in [Schlosshauer 2004] and [Jaeger 2009].

[2] This brief reminder is definitely too brief for those who do not already have some familiarity with quantum mechanics. See [Ballentine 1998].

space $\mathcal{H}_S$. The system interacts with an apparatus $\mathcal{A}$ that is meant to measure a particular observable. According to von Neumann the apparatus $\mathcal{A}$ is also a quantum system and its states can be represented in a basis $\{|a_n\rangle\}$ that spans the Hilbert space $\mathcal{H}_A$. In order to be a good measuring device the states $|a_n\rangle$ are assumed to represent states of $\mathcal{A}$ that are macroscopically distinguishable so they can play the role of the positions of a "pointer".[3]

The measurement consists of allowing systems $\mathcal{S}$ and $\mathcal{A}$ to interact. Their joint time evolution, which is described by the appropriate Schrödinger equation, can be represented by a unitary evolution operator $\hat{U}_A$. The apparatus $\mathcal{A}$ is designed so that when system $\mathcal{S}$ is, for example, in state $|s_n\rangle$ and the apparatus is in its initial "ready to measure" reference state $|a_{\rm ref}\rangle$ then their joint evolution makes the reference state $|a_{\rm ref}\rangle$ evolve to the appropriate pointer position $|a_n\rangle$,

$$\hat{U}_A|s_n\rangle|a_{\rm ref}\rangle = |s_n\rangle|a_n\rangle \ . \tag{10.1}$$

Since the apparatus $\mathcal{A}$ is macroscopic we can read $|a_n\rangle$ and we therefore infer that the original state of $\mathcal{S}$ was $|s_n\rangle$. Plenty of generalizations are possible — for example, it is not necessary that the final state of $\mathcal{S}$ coincide with the initial state $|s_n\rangle$ — but this toy model is already sufficient for our purposes.

The problem arises when the system $\mathcal{S}$ is in a generic superposition state,

$$|\Psi\rangle = \sum_n c_n|s_n\rangle \, . \tag{10.2}$$

Then, since the evolution operator $\hat{U}_A$ is a linear operator the coupling to the measuring device $\mathcal{A}$ leads to the state

$$\hat{U}_A|\Psi\rangle|a_{\rm ref}\rangle = \hat{U}_A\sum_n c_n\hat{U}_A|s_n\rangle|a_{\rm ref}\rangle = \sum_n c_n|s_n\rangle|a_n\rangle \ , \tag{10.3}$$

which is a linear superposition of macroscopically distinct quantum states: the pointer can't make up its mind about which direction to point. Note that this is not saying that the pointer fluctuates as if there was some noise present. It is not that the pointer jumps from one position to another: according to the orthodox interpretation of quantum mechanics the pointer is both in none and in all positions at the same time. Nobody has ever seen such a monstrosity. The pointer has not recorded a definite outcome. Since superpositions evolve to superpositions a linear quantum evolution will never allow definite outcomes to occur.

Remark: The fact that linear time evolution would lead to such *grotesque* states[4] was noticed by both Schrödinger and Einstein very early in the history of quantum mechanics. It is one of the main reasons why Einstein advocated an epistemic or statistical interpretation of quantum theory instead of an ontic interpretation such as the orthodox or Copenhagen interpretations. In the

[3] The center of mass of an N-particle body can, for sufficiently large N, play the role of the pointer variable. (See section 9.6.2.)

[4] The term 'grotesque' to denote such macroscopic superpositions strikes me as particularly apt. It was suggested by L. Schulman.

famous example of Schrödinger's cat the microscopic system $\mathcal{S}$ is a radioactive atom and the macroscopic apparatus $\mathcal{A}$ is a cat. The time evolution is arranged so that as long as the atom remains undecayed (state $|s_U\rangle$) the cat remains alive (state $|a_{\rm alive}\rangle$), but if the atom decays (state $|s_D\rangle$) the cat dies (state $|a_{\rm dead}\rangle$). The normal time evolution of an atom will lead to situations where the atom is in a superposition of decayed and not decayed. Therefore,

$$|s_U\rangle|a_{\rm alive}\rangle \longrightarrow (c_U|s_U\rangle + c_D|s_D\rangle)|a_{\rm alive}\rangle \longrightarrow c_U|s_U\rangle|a_{\rm alive}\rangle + c_D|s_D\rangle|a_{\rm dead}\rangle \ , \tag{10.4}$$

which describes a grotesque state — the cat is "undead".

An early "solution" due to von Neumann [Ballentine 1998] was to postulate a dual mode of wave function evolution. Quantum systems would normally follow a continuous and deterministic evolution according to the Schrödinger equation except during the process of measurement process when the wave function would suffer a discontinuous and stochastic jump into one of the states in the superposition in eq.(10.3). It is in the latter process — the wave function collapse or projection postulate — where probabilities are introduced. (For an excellent criticism of the projection postulate see [Ballentine 1990].)

Other proposed solutions involve denying that collapse ever occurs which has led to the many worlds, the many minds, and the modal interpretations. These issues and others (such as the preferred basis problem) can nowadays be tackled within the decoherence program [Zurek 2003; Schlosshauer 2004] but with one strong caveat. Decoherence works but only at the observational level — it saves the appearances. In this view quantum mechanics is merely empirically adequate and it fails to provide an objective picture of reality. In ontic interpretations of quantum theory this is not acceptable.

Our goal here is to revisit the problem of measurement from the fresh perspective of Entropic Dynamics (ED) which introduces some new elements of its own. The general attitude is pragmatic: physical theories are mere models for inference. They do not attempt to mirror reality and, therefore, all we *want* is that they be empirically adequate, that is, good "for all practical purposes". This is not just the best one can do; since ultimate reality is inaccessible, it is the best that one can ever hope to do. Therefore in the entropic framework the program of decoherence is completely unobjectionable. But this is not the direction that we will pursue here.

Once one accepts quantum theory as a theory of inference the dichotomy between two distinct modes of wave function evolution is erased. As we shall see below the continuous unitary evolution and discontinuous collapse correspond to two modes of processing information, namely the entropic updating in infinitesimal steps (discussed in the previous chapter) and Bayesian updating in discrete finite steps. Indeed, as shown in section 6.6 these two updating rules are not qualitatively different; they are special cases within a broader scheme of entropic inference.

The other element that is significant for our present purpose is that in entropic dynamics particles have only one attribute — position. Particles have neither momentum nor energy. Unlike the standard interpretation of quantum

mechanics, in ED the positions of particles have definite values and just as in classical physics these values are not created by the act of measurement.[5] Therefore the problem of definite outcomes does not arise. Below we will introduce other so-called "observables" but only as a convenient way to describe more complex position measurements. As we shall see these observables will turn out to be attributes not of the particles but of the probability distributions and their values are effectively created by the act of measurement. This opens the opportunity of interpreting all other "observables" in purely informational terms.

In the standard approach to quantum theory it is postulated that observables are represented by self-adjoint operators acting on a suitable Hilbert space. In fact, it is often asserted that all self-adjoint operators are in principle observable. In the entropic framework Hilbert spaces are no longer fundamental which means that self-adjoint operators lose their fundamental status too — they are only useful to the extent that they aid in the analysis of complex position measurements. And there is more: along with Hilbert spaces and most operators, Bohr's doctrine of complementarity must also be abandoned. The point is that once momentum is not a description of reality it is no longer clear in what sense the pair position/momentum could possibly be said to complement each other.

10.2 Observables other than position

In practice the measurement of position can be technically challenging because it requires the amplification of microscopic details to a macroscopically observable scale. However, no intrinsically quantum effects need be involved: the position of a particle has a definite, albeit unknown, value x and its probability distribution is, by construction, given by the Born rule, $\rho(x) = |\Psi(x)|^2$. We can therefore assume that suitable position detectors are available. This is not in any way different from the way information in the form of data is handled in any other Bayesian inference problem. The goal there is to make an inference on the basis of given data; the issue of how the data was collected or itself inferred is not under discussion. If we want we can, of course, address the issue of where the data came from but this is a separate inference problem that requires an independent analysis. In the next section we offer some additional remarks of the amplification problem from a Bayesian perspective.

Our main concern here is with observables other than position: how they are defined and how they are measured.[6] For notational convenience we initially consider the case of a particle that lives on a lattice; the measurement of position leads to a discrete set of possible outcomes. The probabilities of the previously

[5] In this work ED has been developed as a model for the quantum mechanics of particles. The same framework can be deployed to construct models for the quantum mechanics of fields, in which case it is the fields that are "real" and have well defined (but possibly unknown) values.

[6] See [Caticha 2000; Johnson 2011; Johnson Caticha 2011].

continuous positions

$$\rho(x)\,dx = |\langle x|\Psi\rangle|^2\,dx \quad \text{become} \quad p_i = |\langle x_i|\Psi\rangle|^2 \ . \tag{10.5}$$

If the state is

$$|\Psi\rangle = \sum_i c_i|x_i\rangle \quad \text{then} \quad p_i = |\langle x_i|\Psi\rangle|^2 = |c_i|^2 \ . \tag{10.6}$$

Since position is the only objectively real quantity there is no reason to define other observables except that they may turn out to be convenient when considering more complex experiments in which the particle is subjected to additional interactions, say magnetic fields or diffraction gratings, before it reaches the position detectors. Suppose the interactions within the complex measurement device $\mathcal{A}$ are described by the Schrödinger eq.(9.93), that is, by a particular unitary evolution $\hat{U}_A$. The particle will be detected with certainty at position $|x_i\rangle$ provided it was initially in a state $|s_i\rangle$ such that

$$\hat{U}_A|s_i\rangle = |x_i\rangle \ . \tag{10.7}$$

Since the set $\{|x_i\rangle\}$ is orthonormal and complete, the corresponding set $\{|s_i\rangle\}$ is also orthonormal and complete,

$$\langle s_i|s_j\rangle = \delta_{ij} \quad \text{and} \quad \textstyle\sum_i |s_i\rangle\langle s_i| = \hat{I} \ . \tag{10.8}$$

Now consider the effect of this complex detector $\mathcal{A}$ on some generic initial state vector $|\Psi\rangle$ which can always be expanded as

$$|\Psi\rangle = \textstyle\sum_i c_i|s_i\rangle \ , \tag{10.9}$$

where $c_i = \langle s_i|\Psi\rangle$ are complex coefficients. The state $|\Psi\rangle$ will evolve according to $\hat{U}_A$ so that as it approaches the position detectors the new state is

$$\hat{U}_A|\Psi\rangle = \textstyle\sum_i c_i\hat{U}_A|s_i\rangle = \sum_i c_i|x_i\rangle \ . \tag{10.10}$$

which, invoking the Born rule for position measurements, implies that the probability of finding the particle at the position x_i is

$$p_i = |c_i|^2 \ . \tag{10.11}$$

Thus, the probability that the particle in state $\hat{U}_A|\Psi\rangle$ is found at position x_i is $|c_i|^2$.

But we can describe the same outcome from a point of view in which the inner workings of the complex detector are not emphasized; the complex detector is a black box. *The particle is detected in state $|x_i\rangle$ as if it had earlier been in the state $|s_i\rangle$.* We adopt a new language and say, perhaps inappropriately, that the particle has effectively been "detected" in the state $|s_i\rangle$, and therefore, the probability that the particle in state $|\Psi\rangle$ is "detected" in state $|s_i\rangle$ is $|c_i|^2 = |\langle s_i|\Psi\rangle|^2$ — which reproduces Born's rule for a generic measurement device.

The shift in language is not particularly fundamental — it is merely a matter of convenience but we can pursue it further and assert that this complex detector "measures" all operators of the form $\hat{A} = \sum_i \lambda_i |s_i\rangle\langle s_i|$ where the eigenvalues λ_i are arbitrary scalars.

Remark: Note that when we say we have detected the particle at x_i *as if* it had earlier been in state $|s_i\rangle$ we are absolutely not implying that the particle was in the particular state $|s_i\rangle$ — this is just a figure of speech. The actual state is $|\Psi\rangle$ — and this is not the actual physical state of the particle; it is an epistemic state represented by probabilities and entropies.

Remark: Note that it is not necessary that the eigenvalues of the operator $\hat{A}$ be real — they could be complex numbers. What is necessary is that its eigenvectors $|s_i\rangle$ be orthogonal. This means that the Hermitian and anti-Hermitian parts of $\hat{A}$ will be simultaneously diagonalizable. Thus, while $\hat{A}$ does not have to be Hermitian ($\hat{A} = \hat{A}^\dagger$) it must certainly be *normal*, that $\hat{A}$ must commute with its Hermitian adjoint $\hat{A}$, that is, $\hat{A}\hat{A}^\dagger = \hat{A}^\dagger\hat{A}$.

Note also that if a sentence such as "a particle has momentum $\vec{p}$" is used only as a linguistic shortcut that conveys information about the wave function before the particle enters the complex detector then, strictly speaking, there is no such thing as the momentum of the particle: the momentum is not an attribute of the particle but rather it is a statistical attribute of the probability distribution $\rho(x)$ and entropy $S(x)$, a point that is more fully explored later in this chapter.

The generalization to the continuous configuration space is straightforward. For simplicity we consider a discrete one-dimensional lattice and take the limit as the lattice spacing $\Delta x = x_{i+1} - x_i \to 0$. If we call s the eigenvalue corresponding to $|s\rangle$, that is $\hat{A}|s\rangle = s|s\rangle$, then the corresponding limit is $\Delta s_i = s_{i+1} - s_i \to 0$. The discrete completeness relation, eq. (10.8),

$$\sum_i \Delta s_i \; \frac{|s_i\rangle}{(\Delta s_i)^{1/2}} \frac{\langle s_i|}{(\Delta s_i)^{1/2}} = \hat{I} \quad \text{becomes} \quad \int ds \; |s\rangle\langle s| \; = \hat{I} \; , \tag{10.12}$$

where we defined

$$\frac{|s_i\rangle}{(\Delta s_i)^{1/2}} \;\; \to \;\; |s\rangle \; . \tag{10.13}$$

We again consider a measurement device that evolves eigenstates $|s\rangle$ of the operator $\hat{A}$ into unique position eigenstates $|x\rangle$, $\hat{U}_A|s\rangle = |x\rangle$. The mapping from x to s can be represented by an appropriately smooth function $s = g(x)$. In the limit $\Delta x \to 0$, the orthogonality of position states is expressed by a Dirac delta distribution,

$$\frac{\langle x_i|}{\Delta x^{1/2}} \frac{|x_j\rangle}{\Delta x^{1/2}} = \frac{\delta_{ij}}{\Delta x} \quad \to \quad \langle x|x'\rangle = \delta(x - x') \; . \tag{10.14}$$

An arbitrary wave function can be expanded as

$$|\Psi\rangle = \sum_i \Delta s_i \; \frac{|s_i\rangle}{\Delta s_i^{1/2}} \frac{\langle s_i|\Psi\rangle}{\Delta s_i^{1/2}} \quad \text{or} \quad |\Psi\rangle = \int ds \; |s\rangle \, \langle s|\Psi\rangle \; . \tag{10.15}$$

The unitary evolution $\hat{U}_A$ of the wave function leads to

$$\begin{aligned}\hat{U}_A|\Psi\rangle &= \textstyle\sum_i \hat{U}_A|s_i\rangle\langle s_i|\Psi\rangle = \sum_i |x_i\rangle\langle s_i|\Psi\rangle \\ &= \textstyle\sum_i \Delta x \, \dfrac{|x_i\rangle}{\Delta x^{1/2}} \dfrac{\langle s_i|\Psi\rangle}{\Delta s_i^{1/2}} \left(\dfrac{\Delta s_i}{\Delta x}\right)^{1/2} \\ &\to \quad \int dx \; |x\rangle \, \langle s|\Psi\rangle| \frac{ds}{dx}|^{1/2}, \end{aligned} \tag{10.16}$$

so that

$$p_i = |\langle x_i|\hat{U}_A|\Psi\rangle|^2 = |\langle s_i|\Psi\rangle|^2 \quad \to \quad \rho(x)dx = |\langle s|\Psi\rangle|^2 |\frac{ds}{dx}| \, dx = \rho_A(s)ds \; . \tag{10.17}$$

Thus, "the probability that the particle in state $\hat{U}_A|\Psi\rangle$ is found within the range dx is $\rho(x)dx$" can be rephrased as "the probability that the particle in state $|\Psi\rangle$ is found within the range ds is $\rho_A(s)ds$ " where

$$\rho_A(s)ds = |\langle s|\Psi\rangle|^2 \, ds \; , \tag{10.18}$$

which is the continuum version of the Born rule for the observable $\hat{A}$.

In the standard interpretation of quantum mechanics Born's rule is a postulate; within ED it is the natural consequence of unitary time evolution and the hypothesis that *all measurements are ultimately position measurements*. This raises the question of whether our scheme is sufficiently general to encompass all measurements of interest. While there is no general answer that will address all cases — who can, after all, even list all the measurements that future physicists might perform? — we can, nevertheless, ask whether our scheme includes a sufficiently large class of interesting measurements. How, for example, does one measure those observables for which there is no unitary transformation that maps its eigenstates to position eigenstates? Every case demands its own specific analysis. For example, how does one measure the energy of a free particle? Earlier we pointed out that a particular measurement device characterized by eigenvectors $\{|s\rangle\}$ measures all operators of the form $\hat{A} = \int ds \, \lambda(s)|s\rangle \, \langle s|$. Therefore the same device that measures the momentum $\hat{p}$ of a particle (*e.g.*, using a magnetic field or a diffraction grating followed by a position detector such as a photographic plate or a photoelectric cell) can also be used to infer the energy $\hat{H} = \hat{p}^2/2m$ of a free particle.

Here is a trickier example: It is not so easy to place a probe inside the atom, so how does one measure the energy of an electron that is bound to an atom? We take a hint from the way such experiments are typically done in practice: What is measured is the energy of photons (which, being free particles, is not problematic) emitted in transitions between the bound states. The energy of the bound particle is never measured directly; it is inferred. The whole process is a special case of the standard scheme in which the system of interest and the pointer variable of an apparatus become correlated in such a way that

observation of the pointer allows one to infer a quantity of interest. For example, in a Stern-Gerlach experiment the particle's position is the pointer variable which allows one to infer its spin.

The difficulty with the standard von Neumann interpretation is that it is not clear at what stage the pointer variable "collapses" and attains a definite value. This is precisely the difficulty of principle that is resolved in the entropic approach: the pointer variable is a position variable too and therefore always has a definite value.

10.3 Amplification

The technical problem of amplifying microscopic details so they can become macroscopically observable is usually handled with a detection device set up in an initial unstable equilibrium. The particle of interest activates the amplifying system by inducing a cascade reaction that leaves the amplifier in a definite macroscopic final state described by some pointer variable a.

A state $|s_i\rangle$ of the system $\mathcal{S}$ evolves to a position x_i and the goal of the amplification process is to infer the value x_i from the observed value a_j of the pointer variable. The design of the device is deemed successful when x_i and a_j are suitably correlated and this information is conveyed through a likelihood function $P(a_j|x_i)$. An ideal amplification device would be described by $P(a_j|x_i) = \delta_{ji}$. Then the value x_i can be inferred following a standard application of Bayes rule,

$$P(x_i|a_j) = P(x_i)\frac{P(a_j|x_i)}{P(a_j)} \ . \tag{10.19}$$

The point of these considerations is to emphasize that there is nothing intrinsically quantum mechanical about the amplification process. The issue is one of appropriate selection of the information (in this case a_j) that happens to be relevant to a certain inference (in this case x_i). A successful inference is, of course, a matter of clever design: a skilled experimentalist will design the device so that no spurious correlations — whether quantum or otherwise — nor any other kind of interfering noise will stand in the way of inferring x_i.

10.4 But isn't the measuring device a quantum system too?

von Neumann famously drew a boundary line between the quantum and the classical. One side of the boundary is governed by quantum mechanics with a superposition principle and a unitary and linear time evolution given by the Schrödinger equation. The other side is governed by classical physics, possibly by classical statistical mechanics — it is the instability of the macroscopic device that introduces the stochastic element. Our treatment of the amplifying system appears to be drawing a von Neumann boundary too; and in a sense, it is. However, the boundary drawn here is not between a classical reality on one side

and a quantum reality on the other — it is between the microscopic particle with a definite but unknown position and an amplifying system skillfully designed so its pointer has a definite position at all times while the remaining microscopic degrees of freedom turn out to be of no interest. In fact, the dividing line can be drawn anywhere: the amplifier itself can be treated as a fully quantum system too and, as we argue below, this makes absolutely no difference. (See [Johnson 2011].)

The state of the apparatus can be expressed in the position basis $\{|i\rangle|\mu\rangle\}$ where $|i\rangle$ represents the pointer position and $|\mu\rangle$ represents the *positions* of all the other microscopic degrees of freedom. (It is possible to make the model more realistic and consider pointer variables with positions defined over a range of values instead of a sharply defined $|i\rangle$, but this simple model is already sufficient to illustrate our point.)

The initial full state of the apparatus $\mathcal{A}$ in its reference position and ready to take the next measurement is described by

$$|a_{\rm ref}\rangle = \sum_\mu C_\mu\,|r\rangle|\mu\rangle \tag{10.20}$$

where the state $|r\rangle$ represents the reference position of the pointer. (We write the superposition as a discrete sum merely to simplify the notation.) The probabilities $|C_\mu|^2$ are, of course, normalized,

$$\sum_\mu |C_\mu|^2 = 1\ . \tag{10.21}$$

Coupling this apparatus $\mathcal{A}$ in a particular state $|r\rangle|\mu\rangle$ to the system $\mathcal{S}$ in the state $|s_i\rangle$ leads to some final state

$$\hat{U}_A|s_i\rangle|r\rangle|\mu\rangle = \sum_{j\nu} C_{j\nu}^{(i\mu)}|x_j\rangle|i\rangle|\nu\rangle\ , \tag{10.22}$$

where the particle and the apparatus are correlated in some very complicated way. Note that the apparatus has been cleverly designed so that the macroscopically observable position of the pointer $|i\rangle$ allows us to infer that the initial state of the system $\mathcal{S}$ was the state $|s_i\rangle$. Since $\hat{U}_A$ is unitary the probabilities remain normalized,

$$\sum_{j\nu} |C_{j\nu}^{(i\mu)}|^2 = 1\ . \tag{10.23}$$

When the system $\mathcal{S}$ is in a generic superposition $|\Psi\rangle = \sum_i c_i|s_i\rangle$ and the apparatus is in the superposition $|a_{\rm ref}\rangle$ their coupling leads to the state

$$\begin{aligned}
\hat{U}_A|\Psi\rangle|a_{\rm ref}\rangle &= \hat{U}_A\sum_i c_i\,|s_i\rangle\sum_\mu C_\mu|r\rangle|\mu\rangle \\
&= \sum_{i\mu} c_i C_\mu\,\hat{U}_A|s_i\rangle|r\rangle|\mu\rangle \\
&= \sum_{i\mu} c_i C_\mu \sum_{j\nu} C_{j\nu}^{(i\mu)}|x_j\rangle|i\rangle|\nu\rangle \\
&= \sum_{i\mu j\nu} c_i C_\mu\ C_{j\nu}^{(i\mu)}|x_j\rangle|i\rangle|\nu\rangle
\end{aligned} \tag{10.24}$$

The states on the right hand side are all position eigenstates, therefore the probability that the pointer variable is at position i while the other (microscopic and therefore unobservable) degrees of freedom take values x_j (for the particle) and ν (for the apparatus) is

$$P(x_j, i, \nu) = |\sum_\mu c_i C_\mu \; C_{j\nu}^{(i\mu)}|^2 \tag{10.25}$$

$$= |c_i|^2 \, |\sum_\mu C_\mu \; C_{j\nu}^{(i\mu)}|^2 \; . \tag{10.26}$$

Since $\hat{U}_A$ is unitary these probabilities are normalized so that

$$\sum_{ji\nu} P(x_j, i, \nu) = \sum_i |c_i|^2 \sum_{j\nu} |\sum_\mu C_\mu \; C_{j\nu}^{(i\mu)}|^2 = 1 \; , \tag{10.27}$$

which, using

$$\sum_i |c_i|^2 = 1 \; , \tag{10.28}$$

implies that

$$\sum_{j\nu} |\sum_\mu C_\mu \; C_{j\nu}^{(i\mu)}|^2 = 1 \; . \tag{10.29}$$

But we are only interested in the probability of i. Therefore, marginalizing over x_j and ν,

$$P(i) = \sum_{j\nu} P(x_j, i, \nu) \tag{10.30}$$

$$= |c_i|^2 \sum_{j\nu} |\sum_\mu C_\mu \; C_{j\nu}^{(i\mu)}|^2 \; , \tag{10.31}$$

which, using eq.(10.29), gives

$$P(i) = |c_i|^2 \; . \tag{10.32}$$

This coincides with the previous result, eq.(10.11) and concludes our proof: If we want we can treat the apparatus $\mathcal{A}$ in full quantum detail, but since the microscopic degrees of freedom are not relevant they make no difference.

Let us emphasize the main point once again: in entropic dynamics measurements yield definite outcomes because positions, whether macroscopic or otherwise, always have definite values.

10.5 Momentum in Entropic Dynamics

When quantum mechanics was invented a central problem was to identify the concept that would in the appropriate limit correspond to the classical momentum. We face an analogous (but easier) problem: our goal is to identify what concept may reasonably be called momentum within the entropic framework.

Since the particle follows a Brownian non-differentiable trajectory it is clear that the classical momentum $md\vec{x}/dt$ tangent to the trajectory cannot be defined. Nevertheless, four different notions of momentum can be usefully introduced.

First, there is the usual notion of momentum, already familiar from the standard quantum formalism, which is represented as a differential operator:
The quantum momentum is the generator of infinitesimal translations,

$$\vec{p}_q = -i\hbar\vec{\nabla} \ . \tag{10.33}$$

The other obvious momentum candidates correspond to each of the various velocities available to us:
The drift momentum is associated to the velocity with which probability flows due to the entropy gradient,

$$\vec{p}_d = m\vec{b} = \hbar\vec{\nabla}S \, , \tag{10.34}$$

where $\vec{b}$ is the drift velocity given in eq.(9.26) and (9.65).
The osmotic momentum is associated to the velocity with which probability flows due to diffusion,

$$\vec{p}_o = m\vec{u} = -\hbar\vec{\nabla}\log\rho^{1/2} \ . \tag{10.35}$$

(See eq.(9.53) or (9.66).)
The current momentum is associated to the velocity of total probability flow,

$$\vec{p}_c = m\vec{v} = \hbar\vec{\nabla}\phi \quad \text{where} \quad \phi = S - \log\rho^{1/2} \ . \tag{10.36}$$

(See eqs.(9.54) and (9.56).)
What are these mathematical objects? Why should we care about them? Should any of them be called 'momentum'?

Perhaps most important feature of all three of these notions of momentum is that they are expressed in terms of probability ρ and entropy S. This makes it explicit that *they are not attributes associated to the particles* but rather *they are statistical concepts associated to the state of incomplete knowledge of the rational agent engaged in doing inference.* This is precisely in the spirit of the previous sections where we argued that the only actual observables are the positions of the particles, that *all measurements are ultimately position measurements.*

Notice also that the three momenta $\vec{p}_d(\vec{x})$, $\vec{p}_o(\vec{x})$, and $\vec{p}_c(\vec{x})$ are local functions of $\vec{x}$ and this makes them conceptually very different from the differential operator $\vec{p}_q$. The usual language adopted in quantum mechanics is that in a generic state $\Psi(\vec{x})$ the momentum does not have a definite value. It is only in the eigenstates of $\vec{p}_q$,

$$\vec{p}_q \, e^{i\vec{k}\cdot\vec{r}} = \hbar\vec{k} \ e^{i\vec{k}\cdot\vec{r}} \tag{10.37}$$

that the momentum has a definite value, namely, the eigenvalue $\hbar\vec{k}$. Even here note that the definite value $\hbar\vec{k}$ is not localized: it is associated to the wave function $e^{i\vec{k}\cdot\vec{r}}$ as a whole and not to any specific location $\vec{x}$.

In summary, these momenta are neither the classical $md\vec{x}/dt$ nor the quantum momentum, $-i\hbar\vec{\nabla}$. To explore their differences and similarities we find relations among these four momenta and the corresponding uncertainty relations. The results below show a close formal similarity to analogous relations derived in the context of Nelson's stochastic mechanics.[7]

10.5.1 Expected values

The three momenta are not independent. They are related by eq.(9.54)

$$v^a = b^a + u^a \implies \vec{p}_c = \vec{p}_d + \vec{p}_o \ . \tag{10.38}$$

The first important theorem is rather trivial: the expectation of the osmotic momentum vanishes. Indeed, using (10.35) and the fact that ρ vanishes at infinity,

$$\langle p^a_o \rangle = -\hbar \int d^3x \, \rho \, \partial^a \log \rho^{1/2} = -\frac{\hbar}{2} \int d^3x \, \partial^a \rho = 0 \ . \tag{10.39}$$

The immediate consequence is that $\langle p^a_c \rangle = \langle p^a_d \rangle$.

To study the connection to the quantum mechanical momentum we calculate

$$\langle p^a_q \rangle = \int d^3x \Psi^* \frac{\hbar}{i} \partial^a \, \Psi \ . \tag{10.40}$$

Using $\Psi = \rho^{1/2} e^{i\phi}$, (10.39) and (10.36) one gets

$$\begin{aligned} \langle p^a_q \rangle &= -i\hbar \int d^3x \rho \left(\partial^a \log \rho^{1/2} + i \partial^a S - i \partial^a \log \rho^{1/2} \right) \\ &= \hbar \, \langle \partial^a S \rangle = \langle p^a_c \rangle . \end{aligned} \tag{10.41}$$

Therefore

$$\langle \vec{p}_q \rangle = \langle \vec{p}_c \rangle = \langle \vec{p}_d \rangle \ , \tag{10.42}$$

the expectations of quantum momentum, current momentum and drift momentum coincide.

10.5.2 Uncertainty relations

We start by recalling a couple of definitions and an inequality. The variance of a quantity A is

$$\text{var} \, A = \langle (A - \langle A \rangle)^2 \rangle = \left\langle A^2 \right\rangle - \langle A \rangle^2 \, , \tag{10.43}$$

and its covariance with B is

$$\text{cov} \, (A, B) = \langle (A - \langle A \rangle) \, (B - \langle B \rangle) \rangle = \langle AB \rangle - \langle A \rangle \, \langle B \rangle \ . \tag{10.44}$$

The general form of uncertainty relation to be used bellow follows from the Schwarz inequality,

$$\langle a^2 \rangle \langle b^2 \rangle \geq |\langle ab \rangle|^2 \tag{10.45}$$

[7] See [Nelson 1985; de Falco et al 1982; De Martino et al 1984; Golin 1985, 1986] and also the Hall-Reginatto "exact uncertainty" formalism [Hall Reginatto 2002].

or,

$$\left\langle (A-\langle A\rangle)^2 \right\rangle \left\langle (B-\langle B\rangle)^2 \right\rangle \geq |\langle (A-\langle A\rangle)(B-\langle B\rangle)\rangle|^2 \ , \tag{10.46}$$

so that,

$$(\mathrm{var}\, A)(\mathrm{var}\, B) \geq \mathrm{cov}^2(A,B) \ . \tag{10.47}$$

Next we apply these notions to the various momenta. In the context of stochastic mechanics an analogous calculation was given in [de Falco et al 1982; Golin 1985].

Uncertainty relation for osmotic momentum

For simplicity we consider the one-dimensional case. The generalization to many dimensions is immediate. Eq. (10.47) gives

$$(\mathrm{var}\, x)(\mathrm{var}\, p_o) \geq \mathrm{cov}^2(x,p_o) \ . \tag{10.48}$$

Using (10.35) and (10.39) we have

$$\mathrm{cov}\,(x,p_o) = \langle x p_o\rangle - \langle x\rangle\langle p_o\rangle = -\hbar \int dx\, \rho x \partial \log \rho^{1/2} = \frac{\hbar}{2} \, . \tag{10.49}$$

Therefore,

$$(\mathrm{var}\, x)(\mathrm{var}\, p_o) \geq (\frac{\hbar}{2})^2 \quad \text{or} \quad \Delta x\, \Delta p_o \geq \frac{\hbar}{2} \ , \tag{10.50}$$

which resembles the Heisenberg uncertainty relation.

Uncertainty relation for drift momentum

The uncertainty relation is

$$(\mathrm{var}\, x)(\mathrm{var}\, p_d) \geq \mathrm{cov}^2(x,p_d) \ . \tag{10.51}$$

Using (10.36) and (10.44) we have

$$\mathrm{cov}\,(x,p_d) = \hbar \int dx\, \rho x \partial S - (\int dx\, \rho x)(\hbar \int dx\, \rho \partial S) \ . \tag{10.52}$$

The integrands involve two functions ρ and ∂S that can be chosen independently. In particular, we can choose as narrow a probability distribution ρ as we like. For example, the choice $\rho \to \delta(x-x_0)$ leads to $\mathrm{cov}\,(x,p_d) \to 0$. Therefore, the uncertainty relation for drift momentum is

$$(\mathrm{var}\, x)(\mathrm{var}\, p_d) \geq 0 \quad \text{or} \quad \Delta x\, \Delta p_d \geq 0 \ . \tag{10.53}$$

The Schrödinger and the Heisenberg Uncertainty Relations

Next we derive the uncertainty relation for the quantum momentum, p_q. In the standard derivation, which applies to non-commuting operators such as x and p_q, the inequality (10.47) is replaced by

$$(\mathrm{var}\, x)\,(\mathrm{var}\, p_q) \geq \mathrm{cov}^2\,(x, p_q) + \frac{1}{4}|\langle [x, p_q]\rangle|^2 \ , \tag{10.54}$$

which, using $[x, p_q] = i\hbar$ leads to

$$(\mathrm{var}\, x)\,(\mathrm{var}\, p_q) \geq \mathrm{cov}^2\,(x, p_q) + (\frac{\hbar}{2})^2 \ . \tag{10.55}$$

This uncertainty relation was originally proposed in [Schrödinger 1930]. The better known uncertainty relation due to Heisenberg is somewhat weaker: since $\mathrm{cov}^2\,(x, p_q) \geq 0$ it follows that

$$(\mathrm{var}\, x)\,(\mathrm{var}\, p_q) \geq (\frac{\hbar}{2})^2 \quad \text{or} \quad \Delta x\, \Delta p_q \geq \frac{\hbar}{2} \ . \tag{10.56}$$

Our goal is to see how these results arise within the entropic approach and to explore whether any further insights are to be found. Using $\Psi = \rho^{1/2} e^{i\phi}$, (10.35) and (10.36) we have,

$$\langle p_q^2\rangle = \int dx \Psi^* (\frac{\hbar}{i}\partial)^2\, \Psi = \langle p_c^2\rangle + \langle p_o^2\rangle \ . \tag{10.57}$$

Together with $\langle p_0\rangle = 0$ and $\langle p_q\rangle = \langle p_c\rangle$ (see eqs.(10.39) and (10.42)) this leads to

$$\mathrm{var}\, p_q = \langle p_q^2\rangle - \langle p_q\rangle^2 = \mathrm{var}\, p_c + \mathrm{var}\, p_o \ , \tag{10.58}$$

so that

$$(\mathrm{var}\, x)\,(\mathrm{var}\, p_q) = (\mathrm{var}\, x)\,(\mathrm{var}\, p_c) + (\mathrm{var}\, x)\,(\mathrm{var}\, p_o) \ . \tag{10.59}$$

Using the uncertainty relation for the current momentum,

$$(\mathrm{var}\, x)\,(\mathrm{var}\, p_c) \geq \mathrm{cov}^2\,(x, p_c) \ , \tag{10.60}$$

(to which we will return below) and the uncertainty relation for the osmotic momentum, eq.(10.50), gives

$$(\mathrm{var}\, x)\,(\mathrm{var}\, p_q) \geq \mathrm{cov}^2\,(x, p_c) + (\frac{\hbar}{2})^2 \ . \tag{10.61}$$

Next, a straightforward calculation gives

$$\mathrm{cov}\,(x, p_q) = \frac{1}{2}\langle x p_q + p_q x\rangle - \langle x\rangle\langle p_q\rangle = \mathrm{cov}\,(x, p_c) \ . \tag{10.62}$$

Substituting into (10.61) shows that ED reproduces Schrödinger uncertainty relation (10.55) and therefore also Heisenberg's (10.56) — as desired. But, as we shall see below, we can get a bit more insight from the uncertainty relation written in the form (10.59).

Uncertainty relation for the current momentum

Finally, we turn to the uncertainty relation for the current momentum. The challenge is to place a bound on the right hand side of eq.(10.60). We use the fact that certain quantum states are known to exist such the Heisenberg inequality, eq.(10.56), is saturated: $\Delta x\,\Delta p_q = \hbar/2$. In other words, there exist minimum uncertainty states. For example, the ground state of a harmonic oscillator or, more generally the so-called coherent states (see *e.g.* [Ballentine 1998]) are minimum uncertainty states. Together with eq.(10.55) and eq.(10.62) this implies

$$\mathrm{cov}^2\,(x, p_q) = \mathrm{cov}^2\,(x, p_c) \geq 0\ . \tag{10.63}$$

Which leads to the uncertainty relation for the current momentum,

$$(\mathrm{var}\,x)\,(\mathrm{var}\,p_c) \geq 0 \quad \text{or} \quad \Delta x\,\Delta p_c \geq 0\ . \tag{10.64}$$

10.5.3 Discussion

To learn more about these momenta it is useful to recall the classical limit defined by $\hbar \to 0$ with $S_{HJ} = \hbar\phi$, m, and μ fixed. According to eq.(9.77),

$$\vec{p}_d = \vec{p}_c = \vec{\nabla} S_{HJ} \quad \text{and} \quad \vec{p}_o = m\vec{u} = 0\ , \tag{10.65}$$

where S_{HJ} satisfies the classical Hamilton-Jacobi equation, eq.(9.78). Furthermore, according to eq.(9.80) the fluctuations about the expected trajectory vanish.

Let us collect all our results in one place:

Expected values:

$$\langle p_q\rangle = \langle p_c\rangle = \langle p_d\rangle\ , \quad \langle p_o\rangle = 0 \tag{10.66}$$

Uncertainty relations:

$$\Delta x\,\Delta p_q \geq \frac{\hbar}{2}\ , \quad \Delta x\,\Delta p_o \geq \frac{\hbar}{2} \tag{10.67}$$

$$\Delta x\,\Delta p_c \geq 0\ , \quad \Delta x\,\Delta p_d \geq 0 \tag{10.68}$$

Classical limit:

$$p_d = p_c = \nabla S_{HJ} \quad \text{and} \quad p_o = 0\ , \tag{10.69}$$

We find that both the current or the drift momentum can reasonably be called 'momentum' because their expected values agree with that of the quantum momentum operator, eq.(10.42), and in the classical limit they coincide with the classical momentum, eq.(10.65).

The derivation of the uncertainty relations within the entropic framework yields, of course, the standard result $\Delta x\,\Delta p_q = \hbar/2$, but it also leads to a new insight. As we can see from eq.(10.59),

$$(\mathrm{var}\,x)\,(\mathrm{var}\,p_q) = (\mathrm{var}\,x)\,(\mathrm{var}\,p_c) + (\mathrm{var}\,x)\,(\mathrm{var}\,p_o)\ , \tag{10.70}$$

together with eqs.(10.64) and (10.50),

$$(\mathrm{var}\, x)\,(\mathrm{var}\, p_c) \geq 0 \quad \text{and} \quad (\mathrm{var}\, x)\,(\mathrm{var}\, p_o) \geq (\frac{\hbar}{2})^2 \ , \tag{10.71}$$

the non-trivial contribution to the Heisenberg uncertainty relation arises from the osmotic momentum.[8] In other words,

> The Heisenberg uncertainty relation is a diffusion effect — it can be traced back to the original constraint $\left\langle \gamma_{ab}\Delta x^a \Delta x^b \right\rangle = \Delta\overline{\ell}^2$, eq.(9.8), which allowed the particle to move in any direction but only in short steps and led to the non-differentiability of the Brownian paths.

10.5.4 An aside: the hybrid $\mu = 0$ theory

Non-dissipative ED is defined by the Fokker-Planck equation (9.75) and the quantum Hamilton-Jacobi eq.(9.74). Here we focus on the special case with $\mu = 0$. Setting $\hbar = \eta$ and $S_{HJ} = \hbar\phi$ in eq.(9.74) gives

$$\dot{S}_{HJ} + \frac{1}{2m}(\vec{\nabla} S_{HJ})^2 + V = 0 \ , \tag{10.72}$$

which is the classical Hamilton-Jacobi equation. One might be tempted to dismiss this model as a classical stochastic dynamics but this is wrong. The limit $\mu \to 0$ with $\hbar$ and m fixed is very peculiar. The *expected* trajectory lies along a classical path but the osmotic momentum does not vanish,

$$\vec{p}_c = m\vec{v} = \vec{\nabla} S_{HJ} \quad \text{and} \quad \vec{p}_o = m\vec{u} = \hbar\vec{\nabla}\log\rho^{1/2} \ , \tag{10.73}$$

and, since $\hbar/m$ need not be small, the fluctuations,

$$\left\langle \Delta w^a \Delta w^b \right\rangle = \frac{\hbar}{m}\Delta t\, \delta^{ab} \ , \tag{10.74}$$

about the expected trajectory do not vanish either. In fact, they are as strong as regular quantum fluctuations.

All the considerations about momentum described in the previous section apply to the $\mu = 0$ case. In particular, just as in quantum theory, it makes sense to introduce the generator of translations as a momentum operator, $\vec{p}_q = -i\hbar\vec{\nabla}$. This implies that the $\mu = 0$ model obeys uncertainty relations identical to quantum theory,

$$\Delta x\, \Delta p_o \geq \frac{\hbar}{2} \quad \text{and} \quad \Delta x\, \Delta p_q \geq \frac{\hbar}{2} \ . \tag{10.75}$$

And yet, this is not quantum theory: the corresponding Schrödinger equation, obtained by setting $\mu = 0$ in eq.(9.92), leads to

$$i\hbar\dot{\Psi} = -\frac{\hbar^2}{2m}\nabla^2\Psi + V\Psi + \frac{\hbar^2}{2m}\frac{\nabla^2(\Psi\Psi^*)^{1/2}}{(\Psi\Psi^*)^{1/2}}\Psi \ , \tag{10.76}$$

[8] This result was first noted in [de Falco et al 1982] in the context of stochastic mechanics.

which is nonlinear. Therefore there is no superposition principle and the whole Hilbert space framework is not useful here.

We conclude that the $\mu = 0$ model is a hybrid theory, neither fully classical nor fully quantum. It appears classical in that it resembles Brownian motion and obeys the classical Hamilton-Jacobi equation — but no classical theory could possibly exhibit both infinite friction and no dissipation. On the other hand, it seems a quantum theory in that it applies to the usual regime where $\hbar$ is not negligible and it obeys the usual uncertainty principles. The $\mu = 0$ model clearly deserves further study.

10.6 Conclusions

The solution of the problem of measurement within the entropic dynamics framework hinges on two points: first, entropic quantum dynamics is a theory of inference not a law of nature. This erases the dichotomy of dual modes of evolution — continuous unitary evolution versus discrete wave function collapse. The two modes of evolution turn out to correspond to two modes of updating — continuous entropic and discrete Bayesian — which, within the entropic inference framework, are unified into a single updating rule.

The second point is the privileged role of position — particles (and also pointer variables) have definite positions and therefore their values are not created but merely ascertained during the act of measurement. Other "observables" are introduced as a matter of linguistic convenience to describe more complex experiments. These observables turn out to be attributes of the probability distributions and not of the particles; their "values" are indeed "created" during the act of measurement.

Entropic dynamics as a general framework for physics is still in its infancy. Many are the topics that remain to be explored, and some of them can be developed along lines suggested by stochastic mechanics. For example, for the stochastic mechanics of spin see [Dankel 1970; Faris 1982; Nelson 1985; Wallstrom 1990]; for quantum theory on curved manifolds see [Dohrn Guerra 1978, Nelson 1985], and for the quantum theory of fields see [Guerra 1981, Nelson 1986] and references therein. The corresponding ED approaches to spin and to quantum theory on curved spaces are both developed in [Nawaz 2012]. The entropic dynamics approach to quantum scalar fields is developed in [Caticha 2012] and at this point there seems to be no impediment to its generalization to other types of fields.

The overall conclusion is that quantum mechanics is not different from other inference theories. So it appears that Wheeler's conjecture *"...that every law of physics, pushed to the extreme, will be found statistical and approximate, not mathematically perfect and precise"* might be right.

References

[**Aczel 1966**] J. Aczél, *Lectures on Functional Equations and Their Applications* (Academic Press, New York, 1996).

[**Aczel 1975**] J. Aczél and Z. Daróczy, *On Measures of Information and their Characterizations* (Academic Press, New York 1975).

[**Adler 2004**] S. L. Adler, *Quantum Theory as an Emergent Phenomenon* (Cambridge U. Press, Cambridge 2004) (arXiv:hep-th/0206120).

[**Adriaans 2008**] P. W. Adriaans and J.F.A.K. van Benthem (eds.), *Handbook of Philosophy of Information* (Elsevier, 2008).

[**Adriaans 2012**] P. W. Adriaans, "Philosophy of Information", to appear in the *Stanford Encyclopedia of Philosophy* (2012).

[**Amari 1985**] S. Amari, *Differential-Geometrical Methods in Statistics* (Springer-Verlag, 1985).

[**Amari Nagaoka 2000**] S. Amari and H. Nagaoka, *Methods of Information Geometry* (Am. Math. Soc./Oxford U. Press, Providence 2000).

[**Atkinson Mitchell 1981**] C. Atkinson and A. F. S. Mitchell, "Rao's distance measure", Sankhyā **43**A, 345 (1981).

[**Balasubramanian 1997**] V. Balasubramanian, "Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions", Neural Computation **9**, 349 (1997).

[**Ballentine 1970**] L. Ballentine, "The statistical interpretation of quantum mechanics", Rev. Mod. Phys. **42**, 358 (1970).

[**Ballentine 1990**] L. Ballentine, "Limitations of the projection postulate", Found. Phys. **20**, 1329 (1990).

[**Ballentine 1998**] L. Ballentine, *Quantum Mechanics: A Modern Development* (World Scientific, Singapore 1998).

[**Barbour 1994a**] J. B. Barbour, "The timelessness of quantum gravity: I. The evidence from the classical theory", Class. Quant. Grav. **11**, 2853(1994).

[**Barbour 1994b**] J. B. Barbour, "The timelessness of quantum gravity: II. The appearance of dynamics in static configurations", Class. Quant. Grav. **11**, 2875 (1994).

[**Barbour 1994c**] J. B. Barbour, "The emergence of time and its arrow from timelessness" in *Physical Origins of Time Asymmetry*, eds. J. Halliwell et al, (Cambridge U. Press, Cambridge 1994).

[**Blanchard et al 1986**] P. Blanchard, S. Golin and M. Serva, "Repeated mesurements in stochastic mechanics", Phys. Rev. **D34**, 3732 (1986).

[**Bohm Hiley 1993**] D. Bohm and B. J. Hiley, *The Undivided Universe: an ontological interpretation on quantum theory* (Routledge, New York 1993).

[**Bollinger 1989**] J. J. Bollinger et al., "Test of the Linearity of Quantum Mechanics by *rf* Spectroscopy of the $^9Be^+$ Ground State," Phys. Rev. Lett. **63**, 1031 (1989).

[**Bretthorst 1988**] G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation* (Springer, Berlin 1988); available at http://bayes.wustl.edu.

[**Brillouin 1952**] L. Brillouin, *Science and Information Theory* (Academic Press, New York, 1952).

[**Brukner Zeilinger 2002**] C. Brukner and A. Zeilinger, "Information and Fundamental Elements of the Structure of Quantum Theory," in *Time, Quantum, Information*, ed. L. Castell and O. Ischebeck (Springer, 2003) (arXiv:quant-ph/0212084).

[**Callen 1985**] H. B. Callen, *Thermodynamics and an Introduction to Thermostatistics* (Wiley, New York, 1985).

[**Campbell 1986**] L. L. Campbell, "An extended Čencov characterization of the information metric", Proc. Am. Math. Soc. **98**, 135 (1986).

[**Caticha 1998a**] A. Caticha, "Consistency and Linearity in Quantum Theory", Phys. Lett. **A244**, 13 (1998) (arXiv.org/abs/quant-ph/9803086).

[**Caticha 1998b**] A. Caticha, "Consistency, Amplitudes and Probabilities in Quantum Theory", Phys. Rev. **A57**, 1572 (1998) (arXiv.org/abs/quant-ph/ 9804012).

[**Caticha 1998c**] A. Caticha, "Insufficient reason and entropy in quantum theory", Found. Phys. **30**, 227 (2000) (arXiv.org/abs/quant-ph/9810074).

[**Caticha 2000**] A. Caticha, "Maximum entropy, fluctuations and priors", *Bayesian Methods and Maximum Entropy in Science and Engineering*, ed. by A. Mohammad-Djafari, AIP Conf. Proc. **568**, 94 (2001) (arXiv.org/abs/math-ph/0008017).

[**Caticha 2001**] A. Caticha, "Entropic Dynamics", *Bayesian Methods and Maximum Entropy in Science and Engineering*, ed. by R. L. Fry, A.I.P. Conf. Proc. **617** (2002) (arXiv.org/abs/gr-qc/0109068).

[**Caticha 2003**] A. Caticha, "Relative Entropy and Inductive Inference", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by G. Erickson and Y. Zhai, AIP Conf. Proc. **707**, 75 (2004) (arXiv.org/abs/physics/0311093).

[**Caticha 2004**] A. Caticha "Questions, Relevance and Relative Entropy", in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, R. Fischer *et al.* A.I.P. Conf. Proc. Vol. **735**, (2004) (arXiv:cond-mat/0409175).

[**Caticha 2005**] A. Caticha, "The Information Geometry of Space and Time" in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by K. Knuth *et al.* AIP Conf. Proc. **803**, 355 (2006) (arXiv.org/abs/gr-qc/0508108).

[**Caticha 2006**] A. Caticha, "From Objective Amplitudes to Bayesian Probabilities," in *Foundations of Probability and Physics-4,* G. Adenier, C. Fuchs, and A. Khrennikov (eds.), AIP Conf. Proc. **889**, 62 (2007) (arXiv.org/abs/quant-ph/0610076).

[**Caticha 2007**] A. Caticha, "Information and Entropy", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by K. Knuth *et al.*, AIP Conf. Proc. **954**, 11 (2007) (arXiv.org/abs/0710.1068).

[**Caticha 2008**] A. Caticha, *Lectures on Probability, Entropy, and Statistical Physics* (MaxEnt 2008, São Paulo, Brazil) (arXiv.org/abs/0808.0012).

[**Caticha 2009**] A. Caticha, "Quantifying Rational Belief", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by P. Goggans *et al.*, AIP Conf. Proc. **1193**, 60 (2009) (arXiv.org/abs/0908.3212).

[**Caticha 2010a**] A. Caticha, "Entropic Dynamics, Time, and Quantum Theory", J. Phys. **A 44**, 225303 (2011); (arXiv.org/abs/1005.2357).

[**Caticha 2010b**] A. Caticha, "Entropic time", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by A. Mohammad-Djafari, AIP Conf. Proc. **1305** (2010) (arXiv:1011.0746).

[**Caticha 2010c**] A. Caticha, "Entropic Inference", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by A. Mohammad-Djafari *et al.*, AIP Conf. Proc. **1305** (2010) (arXiv.org: 1011.0723).

[**Caticha Cafaro 2007**] A. Caticha and C. Cafaro, "From Information Geometry to Newtonian Dynamics", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by K. Knuth *et al.*, AIP Conf. Proc. **954**, 165 (2007) (arXiv.org/abs/0710.1071).

[**Caticha Giffin 2006**] A. Caticha and A. Giffin, "Updating Probabilities", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by A. Mohammad-Djafari, AIP Conf. Proc. **872**, 31 (2006) (arXiv.org/ abs/physics/0608185).

[**CatichaN Kinouchi 1998**] N. Caticha and O. Kinouchi, "Time ordering in the evolution of information processing and modulation systems", Phil. Mag. **B 77**, 1565 (1998).

[**CatichaN Neirotti 2006**] N. Caticha and J. P. Neirotti, "The evolution of learning systems: to Bayes or not to be", in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by A. Mohammad-Djafari, AIP Conf. Proc. **872**, 203 (2006).

[**Caticha Preuss 2004**] A. Caticha and R. Preuss, "Maximum entropy and Bayesian data analysis: entropic prior distributions", Phys. Rev. **E70**, 046127 (2004) (arXiv.org/abs/physics/0307055).

[**Caves et al 2007**] C. Caves, C. Fuchs, and R. Schack, "Subjective probability and quantum certainty", Studies in History and Philosophy of Modrn Physics **38**, 244 (2007).

[**Cencov 1981**] N. N. Čencov: *Statistical Decision Rules and Optimal Inference*, Transl. Math. Monographs, vol. 53, Am. Math. Soc. (Providence, 1981).

[**Chandrasekhar 1943**] See *e.g.*, S. Chandrasekhar, "Stochastic Problems in Physics and Astronomy" Rev. Mod. Phys. **15**, 1 (1943).

[**Costa de Beauregard Tribus 1974**] O. Costa de Beauregard and M. Tribus, "Information Theory and Thermodynamics", Helv. Phys. Acta **47**, 238 (1974).

[**Cover Thomas 1991**] T. Cover and J. Thomas, *Elements of Information Theory* (Wiley, New York 1991).

[**Cox 1946**] R.T. Cox, "Probability, Frequency and Reasonable Expectation", Am. J. Phys. **14**, 1 (1946).

[**Cox 1961**] R.T. Cox, *The Algebra of Probable Inference* (Johns Hopkins, Baltimore 1961).

[**Cropper 1986**] W. H. Cropper, "Rudolf Clausius and the road to entropy", Am. J. Phys. **54**, 1068 (1986).

[**Csiszar 1984**] I. Csiszar, "Sanov property, generalized I-projection and a conditional limit theorem", Ann. Prob. **12**, 768 (1984).

[**Csiszar 1985**] I. Csiszár "An extended Maximum Entropy Principle and a Bayesian justification", *Bayesian Statistics 2*, p.83, ed. by J. M. Bernardo. M. H. de Groot, D. V. Lindley, and A. F. M. Smith (North Holland, 1985); "MaxEnt, mathematics and information theory", *Maximum Entropy and Bayesian Methods*, p. 35, ed. by K. M. Hanson and R. N.Silver (Kluwer, Dordrecht 1996).

[**Csiszar 1991**] I. Csiszár, "Why least squares and maximum entropy: an axiomatic approach to inference for linear inverse problems", Ann. Stat. **19**, 2032 (1991).

[**Dankel 1970**] T. G. Dankel, Jr., "Mechanics on Manifolds and the incorporation of spin into Nelson's stochastic mechanics", Arch. Rat. Mech. Anal. **37**, 192 (1970).

[**de Falco et al 1982**] D. de Falco, S. D. Martino and S. De Siena, "Position-Momentum Uncertainty Relations in Stochastic Mechanics", Phys. Rev. Lett. **49**, 181 (1982).

[**De Martino et al 1984**] S. De Martino and S. De Siena, "On Uncertainty Relations in Stochastic Mechanics", Il Nuovo Cimento **79B**, 175 (1984).

[**Dewar 2003**] R. Dewar, "Information theory explanation of the fluctuation theorem, maximum entropy production and self-organized criticality in non-equilibrium stationary states", J. Phys. A: Math. Gen. **36** 631 (2003).

[**Dewar 2005**] R. Dewar, "Maximum entropy production and the fluctuation theorem", J. Phys. A: Math. Gen. **38** L371 (2003).

[**Diaconis 1982**] P. Diaconis and S. L. Zabell, "Updating Subjective Probabilities", J. Am. Stat. Assoc. **77**, 822 (1982).

[**Dohrn Guerra 1978**] D. Dohrn and F. Guerra, "Nelson's stochastic mechanics on Riemannian manifolds", Lett. Nuovo Cimento **22**, 121 (1978).

[**Earman 1992**] J. Earman, *Bayes or Bust?: A Critical Examination of Bayesian Confirmation Theory* (MIT Press, Cambridge, 1992).

[**Elze 2002**] H. T. Elze and O. Schipper, "Time withour time: a stochastic clock model", Phys. Rev. **D66**, 044020 (2002).

[**Elze 2003**] H. T. Elze, "Emergent discrete time and quantization: relativistiv particle with extra dimensions", Phys. Lett. **A310**, 110 (2003).

[**Faris 1982a**] W. G. Faris, "A stochastic picture of spin", in *Stochastic Processes in Quantum Theory and Statistical Physics* ed. By S. Albeverio *et al.*, Lecture Notes in Physics **173** (Springer, 1982).

[**Faris 1982b**] W. G. Faris, "Spin correlation in stochastic mechanics", Found. Phys. **12**, 1 (1982).

[**Ferrero et al 2004**] M. Ferrero, D. Salgado, and J. L. Sánchez-Gómez, "Is the Epistemic View of Quantum Mechanics Incomplete?", Found. Phys. **34**, 1993 (2004).

[**Fisher 1925**] R. A. Fisher, "Theory of statistical estimation", Proc. Cambridge Philos. Soc. **122**, 700 (1925).

[**Floridi 2011**] L. Floridi, *The Philosophy of Information* (Oxford U. Press, Oxford 2011).

[**Friederich 2011**] S. Friederich, "How to spell out the epistemic conception of quantum states", Studies in History and Philosophy of Modern Physics **42**, 149 (2011).

[**Fritsche Haugk 2009**] L. Fritsche and M. Haugk, "Stochastic Foundation of Quantum Mechanics and the Origin of Particle Spin", arXiv:0912.3442.

[**Fuchs 2002**] C. Fuchs, "Quantum mechanics as quantum information (and only a little more)," in *Quantum Theory: Reconstruction of Foundations* ed. by A. Khrennikov (Vaxjo U. Press, 2002) (arXiv:quant-ph/0205039).

[**Garrett 1996**] A. Garrett, "Belief and Desire", *Maximum Entropy and Bayesian Methods* ed. by G. R. Heidbreder (Kluwer, Dordrecht 1996).

[**Gibbs 1875-78**] J. W. Gibbs, "On the Equilibrium of Heterogeneous Substances", Trans. Conn. Acad. III (1875-78), reprinted in *The Scientific Papers of J. W. Gibbs* (Dover, New York 1961).

[**Gibbs 1902**] J. W. Gibbs, *Elementary Principles in Statistical Mechanics* (Yale U. Press, New Haven 1902; reprinted by Ox Bow Press, Connecticut 1981).

[**Giffin Caticha 2007**] A. Giffin and A. Caticha, "Updating Probabilities with Data and Moments", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by K. Knuth *et al.*, AIP Conf. Proc. **954**, 74 (2007) (arXiv.org/abs/0708.1593).

[**Gissin 1990**] N. Gisin, Phys. Lett. A **143**, 1 (1990); J. Polchinski, Phys. Rev. Lett. **66**, 397 (1991).

[**Giulini et al 1996**] D. Giulini, E. Joos, C. Kiefer, J. Kupsch, I.-O. Stamatescu, and H.D. Zeh, *Decoherence and the Appearance of a Classical World in Quantum Theory* (Springer, Berlin, 1996).

[**Godfrey-Smith 2003**] P. Godfrey-Smith, *Theory and Reality* (U. Chicago Press, Chicago 2003).

[**Golan 2008**] A. Golan, "Information and Entropy in Econometrics – A Review and Synthesis", Foundations and Trends in Econometrics **2**, 1–145 (2008).

[**Golin 1985**] S. Golin, "Uncertainty relations in stochastic mechanics", J. Math. Phys. **26**, 2781 (1985).

[**Golin 1986**] S. Golin, "Comment on momentum in stochastic mechanics", J. Math. Phys. **27**, 1549 (1986).

[**Good 1950**] I. J. Good, *Probability and the Weighing of Evidence* (Griffin, London 1950).

[**Good 1983**] I. J. Good, *Good Thinking, The Foundations of Probability and its Applications* (University of Minnesota Press, 1983).

[**Goyal Knuth Skilling 2010**] P. Goyal, K. Knuth, J. Skilling, "Origin of complex quantum amplitudes and Feynman's rules", Phys. Rev. **A 81**, 022109 (2010).

[**Grad 1961**] H. Grad, "The Many Faces of Entropy", Comm. Pure and Appl. Math. **14**, 323 (1961), and "Levels of Description in Statistical Mechanics and Thermodynamics", *Delaware Seminar in the Foundations of Physics*, ed. by M. Bunge (Springer-Verlag, New York 1967).

[**Gregory 2005**] P. C. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences* (Cambridge UP, 2005).

[**Grendar 2003**] M. Grendar, Jr. and M. Grendar "Maximum Probability and Maximum Entropy Methods: Bayesian interpretation", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by G. Erickson and Y. Zhai, AIP Conf. Proc. **707**, p. 490 (2004) (arXiv.org/abs/physics/0308005).

[**Greven et al 2003**] A. Greven, G. Keller, and G. Warnecke (eds.), *Entropy* (Princeton U. Press, Princeton 2003).

[**Groessing 2008**] G. Groessing, "The vacuum fluctuation theorem: Exact Schrödinger equation via nonequilibrium thermodynamics", Phys. Lett. **A 372**, 4556 (2008).

[**Groessing 2009**] G. Groessing, "On the thermodynamic origin of the quantum potential", Physica **A 388**, 811 (2009).

[**Guerra 1981**] F. Guerra, "Structural aspects of stochastic mechanics and stochastic field theory", Phys. Rep. **77**, 263 (1981).

[**Guerra Morato 1983**] F. Guerra and L. Morato, "Quantization of dynamical systems and stochastic control theory", Phys. Rev. **D27**, 1774 (1983).

[**Hacking 2001**] I. Hacking, *An Introduction to Probability and Inductive Logic* (Cambridge U. Press, Cambridge 2001).

[**Heisenberg 1958**] W. Heisenberg, *Physics and Philosophy. The Revolution in Modern Science* (Harper, New York, 1958).

[**Hall Reginatto 2002a**] M. J. W. Hall and M. Reginatto, "Schrödinger equation from an exact uncertainty principle", J. Phys. **A 35**, 3289 (2002).

[**Hall Reginatto 2002b**] M. J. W. Hall and M. Reginatto, "Quantum mechanics from a Heisenberg-type equality", Fortschr. Phys. **50**, 646 (2002).

[**Halpern 1999**] J. Y. Halpern, "A Counterexample to Theorems of Cox and Fine", Journal of Artificial Intelligence Research **10**, 67 (1999).

[**Hardy 2001**] L. Hardy, "Quantum Theory From Five Reasonable Axioms" (arXiv.org/quant- ph/0101012).

[**Hardy 2011**] L. Hardy, "Reformulating and Reconstructing Quantum Theory" (arXiv.org:1104.2066).

[**Harrigan Spekkens 2010**] N. Harrigan and R. Spekkens, "Einstein, Incompleteness, and the Epistemic View of Quantum States", Found. Phys. **40**,125 (2010).

[**Hawthorne 1993**] J. Hawthorne, "Bayesian Induction is Eliminative Induction", Philosophical Topics, **21**, 99 (1993).

[**Hempel 1967**] C. G. Hempel, "The white shoe: No red herring", Brit. J. Phil. Sci. **18**, 239 (1967).

[**Holland 1993**] P. R. Holland, *The quantum Theory of Motion* (Cambridge U. Press, Cambridge 1993).

[**Howson Urbach 1993**] C. Howson and P. Urbach, *Scientific Reasoning, the Bayesian Approach* (Open Court, Chicago 1993).

[**Jaeger 2009**] G. Jaeger, *Entanglement, Information, and the Interpretation of Quantum Mechanics* (Springer, Berlin 2009).

[**James 1907**] W. James, *Pragmatism* (Dover, 1995) and *The Meaning of Truth* (Prometheus, 1997).

[**Jammer 1966**] M. Jammer, *The Conceptual Development of Quantum Mechanics* (McGraw-Hill, New York 1966).

[**Jammer 1974**] M. Jammer, *The Philosophy of Quantum Mechanics – The Interpretations of Quantum Mechanics in Historical Perspective* (Wiley, New York 1974).

[**Jeffrey 2004**] R. Jeffrey, *Subjective Probability, the Real Thing* (Cambridge U. Press, Cambridge 2004).

[**Jeffreys 1939**] H. Jeffreys, *Theory of Probability* (Oxford U. Press, Oxford 1939).

[**Jaynes 1957a**] E. T. Jaynes, "How does the Brain do Plausible Reasoning", Stanford Univ. Microwave Lab. report 421 (1957); also published in *Maximum Entropy and Bayesian Methods in Science and Engineering*, G. J. Erickson and C. R. Smith (eds.) (Kluwer, Dordrecht 1988) and at http://bayes.wustl.edu.

[**Jaynes 1957b**] E. T. Jaynes, "Information Theory and Statistical Mechanics", Phys. Rev. **106**, 620 and **108**, 171 (1957).

[**Jaynes 1965**] E. T. Jaynes, "Gibbs vs. Boltzmann Entropies", Am. J. Phys. **33**, 391 (1965).

[**Jaynes 1979**] E. T. Jaynes, "Where do we stand on maximum entropy?" *The Maximum Entropy Principle* ed. by R. D. Levine and M. Tribus (MIT Press 1979); reprinted in [Jaynes 1983] and at http://bayes.wustl.edu.

[**Jaynes 1983**] *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics* edited by R. D. Rosenkrantz (Reidel, Dordrecht, 1983), and papers online at http://bayes.wustl.edu.

[**Jaynes 1985**] E. T. Jaynes, "Bayesian Methods: General Background", in *Maximum Entropy and Bayesian Methods in Applied Statistics*, J. H. Justice (ed.) (Cambridge UP, 1985) and at http://bayes.wustl.edu.

[**Jaynes 1988**] E. T. Jaynes, "The Evolution of Carnot's Principle," pp. 267-281 in *Maximum Entropy and Bayesian Methods in Science and Engineering* ed. by G. J. Erickson and C. R. Smith (Kluwer, Dordrecht 1988) and at http://bayes.wustl.edu.

[**Jaynes 1989**] E. T. Jaynes, "Clearing up the mysteries—the original goal", in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling (Kluwer, Dordrecht 1989).

[**Jaynes 1992**] E. T. Jaynes, "The Gibbs Paradox", *Maximum Entropy and Bayesian Methods*, ed. by C. R. Smith, G. J. Erickson and P. O. Neudorfer (Kluwer, Dordrecht 1992) and at http://bayes.wustl.edu.

[**Jaynes 2003**] E. T. Jaynes, *Probability Theory: The Logic of Science* edited by G. L. Bretthorst (Cambridge UP, 2003).

[**Jeffreys 1946**] H. Jeffreys, "An invariant form for the prior probability in estimation problems", Proc. Roy. Soc. London Ser. A **196**, 453 (1946).

[**Johnson 2011**] D. T. Johnson, "Generalized Galilean Transformations and the Measurement Problem in the Entropic Dynamics Approach to Quantum Theory", Ph.D. thesis, University at Albany (2011) (arXiv:1105.1384).

[**Johnson Caticha 2010**] D. T. Johnson and A. Caticha, "Non-relativistic gravity in entropic quantum dynamics", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by A. Mohammad-Djafari, AIP Conf. Proc. **1305** (2010) (arXiv:1010.1467).

[**Johnson Caticha 2011**] D. T. Johnson and A. Caticha, "Entropic dynamics and the quantum measurement problem", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by K. Knuth *et al.*, AIP Conf. Proc. (2012) (arXiv:1108.2550).

[**Karbelkar 1986**] S. N. Karbelkar, "On the axiomatic approach to the maximum entropy principle of inference", Pramana – J. Phys. **26**, 301 (1986).

[**Kass Wasserman 1996**] R. E. Kass and L. Wasserman, "The Selection of Prior Distributions by Formal Rules", J. Am. Stat. Assoc. **91**, 1343 (1996).

[**Klein 1970**] M. J. Klein, "Maxwell, His Demon, and the Second Law of Thermodynamics", American Scientist **58**, 84 (1970).

[**Klein 1973**] M. J. Klein, "The Development of Boltzmann's Statistical Ideas", *The Boltzmann Equation* ed. by E. G. D. Cohen and W. Thirring, (Springer Verlag, 1973).

[**Knuth 2002**] K. H. Knuth, "What is a question?" in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by C. Williams, AIP Conf. Proc. **659**, 227 (2002).

[**Knuth 2003**] K. H. Knuth, "Deriving laws from ordering relations", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by G.J. Erickson and Y. Zhai, AIP Conf. Proc. **707**, 204 (2003).

[**Knuth 2005**] K. H. Knuth, "Lattice duality: The origin of probability and entropy", Neurocomputing **67C**, 245 (2005).

[**Kullback 1959**] S. Kullback, *Information Theory and Statistics* (Wiley, New York 1959).

[**Landau 1977**] L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Pergamon, New York 1977).

[**Landau 1993**] L. D. Landau and E. M. Lifshitz, *Mechanics* (Butterworth, Oxford 1993).

[**Lindley 1956**] D. V. Lindley, "On a measure of the information provided by an experiment", Ann. Math. Statist. **27**, 986 (1956).

[**Loredo 2003**] T. J. Loredo and D. F. Chernoff, "Bayesian adaptive exploration", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by G. Erickson and Y. Zhai, AIP Conf. Proc. **707**, 330 (2004)

[**Lucas 1970**] J. R. Lucas, *The Concept of Probability* (Clarendon Press, Oxford 1970).

[**Marchildon 2004**] L. Marchildon, "Why Should We Interpret Quantum Mechanics?", Found. Phys. **34**, 1453 (1998).

[**Mehra 1998**] J. Mehra, "Josiah Willard Gibbs and the Foundations of Statistical Mechanics", Found. Phys. **28**, 1785 (1998).

[**Merzbacher 1962**] E. Merzbacher, "Single Valuedness of Wave Functions", Am. J. Phys. **30**, 237 (1962).

[**Nawaz 2012**] S. Nawaz, "Momentum and Spin in Entropic Quantum Dynamics", Ph.D. thesis, University at Albany (2012) .

[**Nawaz Caticha 2011**] S. Nawaz and A. Caticha, "Momentum and uncertainty relations in the entropic approach to quantum theory", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by K. Knuth *et al.*, AIP Conf. Proc. (2012) (arXiv:1108.2629).

[**Neirotti CatichaN 2003**] J. P. Neirotti and N. Caticha, "Dynamics of the evolution of learning algorithms by selection" Phys. Rev. **E 67**, 041912 (2003).

[**Nelson 1966**] E. Nelson, "Derivation of the Schrödinger equation from Newtonian Mechanics", Phys. Rev. **150**, 1079 (1966).

[**Nelson 1967**] E. Nelson, *Dynamical theories of Brownian motion* (Princeton U. Press, Princeton 1967).

[**Nelson 1979**] E. Nelson, "Connection between Brownian motion and quantum mechanics", p.168 in *Einstein Symposium Berlin*, Lecture Notes in Physics 100 (Springer-Verlag, Berlin 1979).

[**Nelson 1985**] E. Nelson, *Quantum Fluctuations* (Princeton U. Press, Princeton 1985).

[**Nelson 1986**] E. Nelson, "Field theory and the future of stochastic mechanics", in in *Stochastic Processes in Classical and Quantum Systems*, ed. By S. Albeverio *et al.*, Lecture Notes in Physics **262** (Springer, Berlin 1986).

[**Papineau 1996**] D. Papineau (ed.), *The Philosophy of Science* (Oxford U. Press, Oxford 1996).

[**Pauli 1939**] W. Pauli, Helv. Phys. Acta **12**, 147 (1939) and W. Pauli, *General Principles of Quantum Mechanics* section 6 (Springer-Verlag, Berlin 1980).

[**Peres 1993**] A. Peres, *Quantum Theory: Concepts and Methods* (Kluwer, Dordrecht 1993).

[**Plastino 1994**] A. R. Plastino and A. Plastino, "From Gibbs microcanonical ensemble to Tsallis generalized canonical distribution", Phys. Lett. **A193**, 140 (1994).

[**Price 1996**] H. Price, *Time's Arrow and Archimides' Point* (Oxford U. Press, Oxford 1996).

[**Putnam 1981**] H. Putnam, *Reason, Truth and History* (Cambridge U. Press, Cambridge 1981).

[**Putnam 2003**] H. Putnam, *The Collapse of the Fact/Value Dichotomy and Other Essays* (Harvard U. Press, Cambridge 2003).

[**Rao 1945**] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters", Bull. Calcutta Math. Soc. **37**, 81 (1945).

[**Renyi 1961**] A. Renyi, "On measures of entropy and information", *Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability*, Vol 1, p. 547 (U. of California Press, Berkeley 1961).

[**Rodriguez 1988**] C. C. Rodríguez, "Understanding ignorance", *Maximum Entropy and Bayesian Methods*, G. J. Erickson and C. R. Smith (eds.) (Kluwer, Dordrecht 1988).

[**Rodriguez 1989**] C. C. Rodríguez, "The metrics generated by the Kullback number", *Maximum Entropy and Bayesian Methods*, J. Skilling (ed.) (Kluwer, Dordrecht 1989)

[**Rodriguez 1990**] C. C. Rodríguez, "Objective Bayesianism and geometry", *Maximum Entropy and Bayesian Methods*, P. F. Fougère (ed.) (Kluwer, Dordrecht 1990).

[**Rodriguez 1991**] C. C. Rodríguez, "Entropic priors", *Maximum Entropy and Bayesian Methods*, edited by W. T. Grandy Jr. and L. H. Schick (Kluwer, Dordrecht 1991).

[**Rodriguez 1998**] C. C. Rodríguez, "Are we cruising a hypothesis space?" (arxiv.org/abs/physics/9808009).

[**Rodriguez 2002**] C. C. Rodríguez: "Entropic Priors for Discrete Probabilistic Networks and for Mixtures of Gaussian Models", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by R. L. Fry, AIP Conf. Proc. **617**, 410 (2002) (arXiv.org/abs/physics/0201016).

[**Rodriguez 2003**] C. C. Rodríguez, "A Geometric Theory of Ignorance" (omega.albany.edu:8008/ignorance/ignorance03.pdf).

[**Rodriguez 2004**] C. C. Rodríguez, "The Volume of Bitnets" (omega.albany.edu:8008/bitnets/bitnets.pdf).

[**Rodriguez 2005**] C. C. Rodríguez, "The ABC of model selection: AIC, BIC and the new CIC", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by K. Knuth *et al.*, AIP Conf. Proc. Vol. **803**, 80 (2006) (omega.albany.edu:8008/CIC/me05.pdf).

[**Savage 1972**] L. J. Savage, *The Foundations of Statistics* (Dover, 1972).

[**Schlosshauer 2004**] M. Schlosshauer, "Decoherence, the measurement problem, and interpretations of quantum mechanics", Rev. Mod. Phys. **76**, 1267 (2004).

[**Schrodinger 1930**] E. Schrodinger, "About the Heisenberg Uncertainty Relation", Sitzungberichten der Preussischen Akademie der Wissenschaften (Phys. Math. Kasse) **19**, 296 (1930); the English translation by A. Agelow and M. Batoni is found at arxiv.org/abs/quant-ph/9903100.

[**Sebastiani Wynn 00**] P. Sebastiani and H. P. Wynn, "Maximum entropy sampling and optimal Bayesian experimental design", J. Roy. Stat. Soc. **B**, 145 (2000).

[**Seidenfeld 1986**] T. Seidenfeld, "Entropy and Uncertainty", Philosophy of Science **53**, 467 (1986); reprinted in *Foundations of Statistical Inference*, I. B. MacNeill and G. J. Umphrey (eds.) (Reidel, Dordrecht 1987).

[**Shannon 1948**] C. E. Shannon, "The Mathematical Theory of Communication", Bell Syst. Tech. J. **27**, 379 (1948).

[**Shannon Weaver 1949**] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, (U. Illinois Press, Urbana 1949).

[**Shimony 1985**] A. Shimony, "The status of the principle of maximum entropy", Synthese **63**, 35 (1985).

[**Shore Johnson 1980**] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy", IEEE Trans. Inf. Theory **IT-26**, 26 (1980); "Properties of Cross-Entropy Minimization", IEEE Trans. Inf. Theory **IT-27**, 26 (1981).

[**Sivia Skilling 2006**] D. S. Sivia and J. Skilling, *Data Analysis: a Bayesian tutorial* (Oxford U. Press, Oxford 2006).

[**Skilling 1988**] J. Skilling, "The Axioms of Maximum Entropy", *Maximum-Entropy and Bayesian Methods in Science and Engineering*, G. J. Erickson and C. R. Smith (eds.) (Kluwer, Dordrecht 1988).

[**Skilling 1989**] J. Skilling, "Classic Maximum Entropy", *Maximum Entropy and Bayesian Methods*, ed. by J. Skilling (Kluwer, Dordrecht 1989).

[**Skilling 1990**] J. Skilling, "Quantified Maximum Entropy", *Maximum Entropy and Bayesian Methods*, ed. by P. F. Fougère (Kluwer, Dordrecht 1990).

[**Smolin 1986a**] L. Smolin, Class. Quantum Grav. **3**, 347 (1986).

[**Smolin 1986b**] L. Smolin, Phys. Lett. **113A**, 408 (1986).

[**Smolin 2006**] L. Smolin, "Could quantum mechanics be an approximation to another theory?" (arXiv.org/abs/quant-ph/0609109).

[**Smith Erickson 1990**] C. R. Smith, G. J. Erickson, "Probability Theory and the Asociativity Equation", in *Maximum Entropy and Bayesian Methods* ed. by P. F. Fougère (Kluwer, Dordrecht 1990).

[**Spekkens 2007**] R. Spekkens, "Evidence for the epistemic view of quantum states: a toy theory", Phys. Rev. **A 75**, 032110 (2007).

[**Stapp 1972**] H. P. Stapp, "The Copenhagen Interpretation", Am. J. Phys. **40**, 1098 (1972).

[**'t Hooft 1999**] G. 't Hooft, "Quantum Gravity as a Dissipative Deterministic System", Class. Quant. Grav. **16**, 3263 (1999) (arXiv:gr-qc/9903084).

[**Tribus 1961**] M. Tribus, "Information Theory as the Basis for Thermostatics and Thermodynamics", J. Appl. Mech. (March 1961) p. 1-8.

[**Tribus 1969**] M. Tribus, *Rational Descriptions, Decisions and Designs* (Pergamon, New York 1969).

[**Tribus 1978**] M. Tribus, "Thirty Years of Information Theory", *The Maximum Entropy Formalism*, R.D. Levine and M. Tribus (eds.) (MIT Press, Cambridge 1978).

[**Tsallis 1988**] C. Tsallis, "Possible Generalization of Boltzmann-Gibbs Statistics", J. Stat. Phys. **52**, 479 (1988).

[**Tsallis 2011**] C. Tsallis, "The nonadditive entropy S_q and its applications in physics and elsewhere; some remarks", Entropy **13**, 1765 (2011).

[**Tseng Caticha 2001**] C.-Y. Tseng and A. Caticha, "Yet another resolution of the Gibbs paradox: an information theory approach", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by R. L. Fry, A.I.P. Conf. Proc. **617**, 331 (2002) (arXiv.org/abs/cond-mat/0109324).

[**Tseng Caticha 2008**] C. Y. Tseng and A. Caticha, "Using relative entropy to find optimal approximations: An application to simple fluids", Physica **A 387**, 6759 (2008) (arXiv:0808.4160).

[**van Fraasen 1980**] B. C. van Fraasen, *The Scientific Image* (Clarendon, Oxford 1980).

[**van Fraasen 1981**] B. C. van Fraasen, "A problem for relative information minimizers in probability kinematics", Brit. J. Phil. Sci. **32**, 375 (1981).

[**van Fraasen 1986**] B. C. van Fraasen, "A problem for relative information minimizers, continued", Brit. J. Phil. Sci. **37**, 453 (1986).

[**van Fraasen 1989**] B. C. van Fraasen, *Laws and Symmetry* (Clarendon, Oxford 1989).

[**Van Horn 2003**] K. Van Horn, "Constructing a Logic of Plausible Inference: a Guide to Cox's Theorem", Int. J. Approx. Reasoning **34**, 3 (2003).

[**von Mises 1957**] R. von Mises, *Probability, Statistics and Truth* (Dover, 1957).

[**Uffink 1995**] J. Uffink, "Can the Maximum Entropy Principle be explained as a consistency Requirement?" Studies in History and Philosophy of Modern Physics **26**, 223 (1995).

[**Uffink 1996**] J. Uffink, "The Constraint Rule of the Maximum Entropy Principle", Studies in History and Philosophy of Modern Physics **27**, 47 (1996).

[**Uffink 2003**] J. Uffink, "Irreversibility and the Second Law of Thermodynamics", in *Entropy*, ed. by A. Greven et al. (Princeton UP, 2003).

[**Uffink 2004**] J. Uffink, "Boltzmann's Work in Statistical Physics", *The Stanford Encyclopedia of Philosophy* (http://plato.stanford.edu).

[**Wallstrom 1989**] T. C. Wallstrom, "On the derivation of the Schrödinger equation from stochastic mechanics", Found. Phys. Lett. **2**, 113 (1989).

[**Wallstrom 1990**] T. C. Wallstrom, "The stochastic mechanics of the Pauli equation", Trans. Am. Math. Soc. **318**, 749 (1990).

[**Wallstrom 1994**] T. C. Wallstrom, "The inequivalence between the Schrödinger equation and the Madelung hydrodynamic equations", Phys. Rev. **A49**, 1613 (1994).

[**Wetterich 2010**] C. Wetterich, "Quantum particles from coarse grained classical probabilities in phase space", Ann. Phys. **325**, 1359 (2010) (arXiv:1003.3351).

[**Wheeler Zurek 1983**] J. A. Wheeler and W. H. Zurek, *Quantum Theory and Measurement* (Princeton U. Press, Princeton 1983).

[**Wigner 1963**] E. P. Wigner, "The problem of measurement", Am J. Phys. **31**, 6 (1963).

[**Williams 1980**] P. M. Williams, "Bayesian Conditionalization and the Principle of Minimum Relative Information", Brit. J. Phil. Sci. **31**, 131 (1980).

[**Wilson 1981**] S. S. Wilson, "Sadi Carnot", Scientific American, August 1981, p. 134.

[**Wootters 1981**] W. K. Wootters, "Statistical distance and Hilbert space", Phys. Rev. **D**, 357 (1981).

[**Zeh 2001**] H. D. Zeh, *The Physical Basis of the Direction of Time* (Springer, Berlin 2002).

[**Zeh 2002**] H. D. Zeh, "The Wave Function: It or Bit?" (arXiv.org/abs/quant-ph/0204088).

[**Zellner 1997**] A. Zellner, "The Bayesian Method of Moments", Advances in Econometrics **12**, 85 (1997).

[**Zurek 2003**] W. H. Zurek, "Decoherence, einselection, and the quantum origins of the classical", Rev. Mod. Phys. **75**, 715 (2003).