ENTROPIC PHYSICS

Probability, Entropy,
and the Foundations of Physics

ARIEL CATICHA

Draft last modified on 07/26/2022. Sections marked $*$ require revision.

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. Thus, it is 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 N. Caticha and C. R. Rodríguez, whose views on these matters have over the years profoundly in-

fluenced my own, but I have also learned much from discussions with many colleagues and friends: D. Bartolomeo, C. Cafaro, N. Carrara, S. DiFranzo, V. Dose, K. Earle, A. Giffin, A. Golan, S. Ipek, D. T. Johnson, K. Knuth, O. Lunin, S. Nawaz, P. Pessoa, R. Preuss, M. Reginatto, J. Skilling, J. Stern, C.-Y. Tseng, K. Vanslette, and A. Yousefi. 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 pointed the way to clarifying my own questions and doubts.

Albany, ...

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 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. It is clear that the interpretations offered by various schools are at least partially successful or else they would already have been discarded long ago. 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.

1.1.1 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, a process that is thought to be independent on any feature or characteristic of the observers. Probabilities interpreted in this way have been called *objective*. This view has 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 even then there is always the question of choosing the relevant ensemble: doesn't this choice reflect somebody's decisions, interests, and purposes? How objective is this choice of ensemble? Indeed, 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 exist 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.

1.1.2 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 degree of belief of an agent in the truth of a proposition.[1] These probabilities are said to be *Bayesian* because of the central role played

[1] 'Degree of belief' is only a quick and dirty way to describe what Bayesian probabilities are about but there are many shades of interpretation. As we shall later argue a more useful definition of probability is the degree to which an ideally rational agent ought to believe in the truth of a proposition. Other interpretations include, for example, the degree of personal belief as contrasted to a degree of rational belief or reasonable expectation, the degree of plausibility of a proposition, the degree of credibility, and also the degree of implication (the degree to which b implies a is the conditional probability of a given b).

by Bayes' theorem – a theorem which was first written by 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 allow us to reason in a consistent and rational manner about quantities, such as parameters, that rather than being random might be merely unknown. Also Bayesian probabilities are not restricted to repeatable events and therefore they allow us to reason about singular, unique 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 agents 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. At one end of the spectrum we find the so-called subjective Bayesian or *personalistic* view (see, *e.g.*, [Savage 1972; Howson Urbach 1993; Jeffrey 2004]), and the other end there 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 degrees of personal belief. Subjective Bayesians accept that individuals can change their 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 then said that a probability measures a degree of implication, it is the degree to which one proposition implies another. It is assumed that the objective Bayesian has thought so long and so 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.

1.1.3 Subjective or objective? Epistemic or ontic?

Whether Bayesian probabilities are subjective or objective is still a matter of controversy. The confusion arises in large part because there are two different senses in which the subjective/objective distinction can be drawn. One is ontological (related to existence, to being) and the other epistemological (related to knowledge and opinion).

In the ontological sense entities such as pains and emotions are said to subjective because they exist only as experienced by some agent. Other entities, such as atoms and chairs, are called objective because they presumably exist out there in the real world quite independently of any agent. On the other hand there is the epistemological sense. A proposition that states a fact is said to be (epistemically) objective in that its truth can in principle be established independently of anyone's attitudes or thoughts while a proposition that reflects a value judgment is said to be (epistemically) subjective. Unfortunately the boundaries between these distinctions are not nearly as sharply defined as one might wish.

Bayesian probabilities are ontologically subjective because they are tools for reasoning. They exist only in the minds of those agents who use them. This is in stark contrast to those interpretations in which probabilities are conceived as something physically real. Examples of the latter include Popper's interpretation of probability as a *physical propensity*, and Heisenberg's probability as an *objective potentiality*, and then there is also probability as the *objective chance* for an event to actually happen. Such ontologically objective "probabilities" are not tools for reasoning or for inference. Unlike other physical entities – such as particles and fields – postulating their existence has not led to successful theoretical models.

In the epistemological sense however the characterization of probabilities as either subjective or objective is not nearly so clear. Our position is that probabilities can lie anywhere in between. Probabilities will always retain a subjective element because translating information into probabilities involves judgments and different people will often 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 can be "objectively" better in that they provide better guidance about how to cope with the world.[2] The adoption of better beliefs has real consequences. Not all probability assignments are equally useful and it is plausible that what makes some assignments better than others is that they correspond to 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" — whatever that might be.

We shall find that while the epistemic subjectivity can only be eliminated in idealized situations, the rules for processing information, that is, the rules for updating probabilities, are considerably more objective. This means that as new information is obtained it can be *objectively* incorporated into the updated probabilities. Ultimately, it is the conviction that the updated or *posterior* probabilities are somehow objectively better than the original *prior* probabilities that provides the justification for going through the trouble of gathering information to update our beliefs.

To summarize: Referring to probabilities as subjective or objective can lead to confusion because it is not clear whether these terms are deployed in an ontological or an epistemological sense. As we shall see in the next chapter probabilities will be designed as tools for handling uncertainty, for dealing with incomplete information. Such probabilities will turn out to be ontologically subjective but epistemically they can lie anywhere from fully objective to fully subjective.

[2] The approach is decidedly pragmatic: the purpose of thinking is to acquire beliefs; the purpose of beliefs is to guide our actions.

A suitable terminology A central concern is to maintain a clear distinction between ontologically objective and ontologically subjective entities. Ontologically objective entities will be succinctly described as being *real* or *ontic*. Such things *exist* in the sense that, at least within our theories, they constitute the furniture of the world; examples include familiar macroscopic objects such as tables and chairs.[3] The only ontologically subjective entities that will concern us here are probabilities and those other concepts closely associated with them such as entropies or quantum wave functions. Such entities will be referred to as *epistemic*.[4] Thus, we shall avoid the terms objective/subjective in the ontological sense and replace them by the terms ontic/epistemic.

On the other hand, in their epistemological sense the terms objective/subjective are useful and must be retained. There is much to be gained by rejecting a sharp objective/subjective dichotomy and replacing it with a continuous spectrum of intermediate possibilities.[5] Probabilities are epistemically hybrid; they incorporate both subjective and objective elements. We process information to suppress the former and enhance the latter because this is what leads to probabilities that are useful in practice.

1.2 Designing a framework for inductive inference

A common hope in science, mathematics and philosophy has been to find a secure foundation for knowledge. 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 belief — this is handled through the introduction of probabilities. Second, it must allow us to update from one state of belief to another in the fortunate circumstance that some new information becomes available — this is handled

[3] The ontic status of microscopic things depends on the particular model or theory. As we shall see in Chapter 11 in non-relativistic quantum mechanics particles and atoms will be described as ontic. However, in relativistic quantum field theory it is the fields that are ontic and those excited states (probabilistic distributions of fields) that are called particles are epistemic. Our position reflects a *pragmatic realism* that is close to the *internal realism* advocated by H. Putnam [Putnam 1979, 1981, 1987].

[4] The term 'epistemic' is not appropriate for emotions or values. However, such ontologically subjective entities will not enter our discussions and there is no pressing need to accommodate them in our terminology.

[5] Here again, our position bears some resemblance to that of H. Putnam who has forcefully argued for the rejection of the fact/value dichotomy [Putnam 1991, 2003].

through the introduction of relative entropy as the tool for updating. *The theory of probability would be severely handicapped – indeed it would be quite useless – without a companion theory for updating probabilities.*

The framework for inference will be constructed by a process of *eliminative induction*.[6] 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. Then, in the *eliminative* part of the process one proceeds to systematically rule out all those inference schemes 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.

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

[6] Eliminative induction is a method to select one alternative from within a set of possible ones. For example, to select the right answer to a question one considers a set of possible candidate answers and proceeds to systematically eliminate those that are found wrong or unacceptable in one way or another. The answer that survives after all others have been ruled out is the best choice. There is, of course, no guarantee that the last standing alternative is the correct answer – the only certainty is that all other answers were definitely wrong.

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 updating 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 E.T. 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[7] while thermodynamics was meant to be "true" by virtue of reflecting "laws of nature". The gap was enormous and 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.

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.

The insight derived from recognizing that one physical theory is an example of inference leads to the obvious question: is statistical mechanics the only one or are there other examples? The answer is yes. Starting in chapter 10 we explore new territory 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, 7, and 8 is needed here — the old entropic methods developed by Shannon and Jaynes are no longer sufficient.

The payoff is considerable. The mathematical framework of quantum mechanics 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

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

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.

In this chapter we will also address a special case of the second problem — that of updating from one consistent web of beliefs to another when new information in the form of data becomes available. 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.[2] As an illustration of the enormous power of

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

[2] The presentation in this chapter includes material published in [Caticha Giffin 2006, Caticha 2007, Caticha 2009, Caticha 2014a].

Bayes' rule we will briefly explore its application to data analysis. As we shall see in later chapters the minimal updating principle is not restricted to data but can also be deployed to process more general kinds of information. This will require the design of a more sophisticated updating tool – relative entropy.

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 a 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 turns out to be 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 probabilities, 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 pragmatic 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 — this important task is best left to psychology and to cognitive science. 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 an idealized standard of rationality that we humans ought to strive for at least when discussing scientific matters.

The concept of rationality is notoriously difficult to pin down. We adopt a pragmatic approach: 'rational' is a compliment we pay to a particular mode of argument that appears to lead to reliable conclusions. One thing we can safely assert 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 particular 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 is not 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 offering a precise criterion of rationality 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 requirement is that if a conclusion can be reached by arguments that follow two different paths then the two arguments must agree. Otherwise our framework is not performing the function for which it is being designed, namely, to provide guidance as to what we are supposed to believe. Thus, the web of rational beliefs must avoid inconsistencies. 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. The notion of 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 and therefore 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 – Boolean Algebra

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 we say they have the same truth value and write $a = b$, when a is true if and only if b is true. The conjunction of two propositions "a AND b" is denoted ab or $a \wedge b$. The conjunction is true if and only if both a and b are true. The disjunction of two propositions "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.

With these symbols one defines an algebra of logic – a Boolean algebra. Important properties of AND and OR include,

$$aa = a \ , \quad a \vee a = a \ , \tag{2.1}$$

commutivity,

$$ab = ba \ , \quad a \vee b = b \vee a \tag{2.2}$$

associativity,

$$a(bc) = (ab)c \ , \quad a \vee (b \vee c) = (a \vee b) \vee c \ , \tag{2.3}$$

and distributivity,

$$a(b \vee c) = (ab) \vee (ac) \ , \tag{2.4}$$

$$a \vee (bc) = (a \vee b)(a \vee c) \ . \tag{2.5}$$

Typically we want to quantify the degree of belief in a in the context of some background information expressed in terms of some other proposition b. Such propositions will be writen as $a|b$ and read "a given b", or "a assuming b". In cases such as $b = c\tilde{c}$ where b is guaranteed to be false, the conditional proposition $a|b$ is meaningless. To simplify the notation we will write $(a \vee b)|c = a \vee b|c$ and $(ab)|c = ab|c$.

In any inference problem the set of all propositions that can be constructed using OR and AND — the "universe of discourse" — forms an ordered lattice structure. Figure 2.1 shows the universe of discourse for a single toss of a three-sided die. The possible outcomes of the toss — the atomic propositions — correspond to each one of the three faces of the die a, b, and c. Using OR we can build the lattice up until we reach top proposition which is necessarily true. Using AND we build the lattice down; the bottom proposition is necessarily false.

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. Assigning a degree of belief to each proposition in the universe of discourse — or to each node in the lattice of propositions — produces what one might call a "web of belief".

Degrees of rational belief will range from the extreme value v_F that represents complete certainty that the proposition is false (for example, for any a, $[\tilde{a}|a] = v_F$), to the opposite extreme v_T that represents certainty that the proposition is true (for example, for any a, $[a|a] = v_T$). The transitivity of the ranking scheme implies that there is a single value v_F and a single v_T.

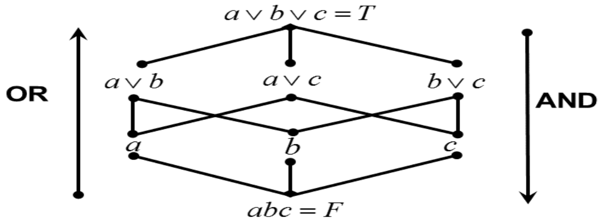

Figure 2.1: The universe of discourse — the set of all propositions — for a three-sided die with faces labelled a, b, and c forms an ordered lattice. The assignment of degrees of belief to each proposition $a \to [a]$ leads to a "web of belief". The web of belief is highly constrained because it must reflect the structure of the underlying lattice.

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,

$$[a \vee b|c] = F([a|c],[b|c],[a|bc],[b|ac]) \tag{2.6}$$

and

$$[ab|c] = G([a|c],[b|c],[a|bc],[b|ac]) \ . \tag{2.7}$$

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 OR and AND. 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 only one representation — all allowed representations are equivalent to each other — and that this unique representation is equivalent to probability theory.

In section 2.2 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 2.3 we focus on the distributive property of AND over OR and the corresponding constraint leads to the product rule for probabilities.[3]

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,

$$[\tilde{a}|b] = f\left([a|b]\right) \ . \tag{2.8}$$

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 $[b|ac]$,

$$[ab|c] = g\left([a|c], [b|ac]\right) \ . \tag{2.9}$$

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 $[b|ac]$. However, one could object that $[ab|c]$ could in principle depend on all four quantities $[a|c]$, $[b|c]$, $[a|bc]$ and $[b|ac]$. This objection, which we address below, has a long history. It was partially addressed in [Tribus 1969; Smith Erickson 1990; Garrett 1996] and finally resolved in [Caticha 2009b].

[3] 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.

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,

$$[a \vee b|d] = F([a|d], [b|d], v_F, v_F) \ , \tag{2.10}$$

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

$$[a \vee b|d] = F([a|d], [b|d]) \ . \tag{2.11}$$

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,

$$(a \vee b) \vee c = a \vee (b \vee c) \ , \tag{2.12}$$

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

$$[a \vee b \vee c|d] = F\left([a \vee b|d], [c|d]\right) = F\left([a|d], [b \vee c|d]\right) \ . \tag{2.13}$$

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

$$F\{F\left([a|d], [b|d]\right), [c|d]\} = F\{[a|d], F\left([b|d], [c|d]\right)\} \ . \tag{2.14}$$

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

$$F\left(F(x, y), z\right) = F\left(x, F(y, z)\right) \ . \tag{2.15}$$

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 (2.15) 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 (2.15) 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 (2.15), 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.[4]

2.2.2 The general solution and its regraduation

Equation (2.15) 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.(2.15) is satisfied by any function of the form

$$F(x, y) = \phi^{-1}(\phi(x) + \phi(y) + \beta) \ , \tag{2.16}$$

where ϕ is an arbitrary invertible function and β is an arbitrary constant. 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. A proof of this result – first given by Cox – is given in section 2.2.4 [Cox 1946; Jaynes 1957a; Aczel 1966].

The significance of eq.(2.16) becomes apparent once it is rewritten as

$$\phi(F(x, y)) = \phi(x) + \phi(y) + \beta \quad \text{or} \quad \phi([a \vee b|d]) = \phi([a|d]) + \phi([b|d]) + \beta \ . \tag{2.17}$$

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,

$$\xi(a|d) \overset{\text{def}}{=} \phi([a|d]) + \beta, \quad \xi(b|d) \overset{\text{def}}{=} \phi([b|d]) + \beta, ... \tag{2.18}$$

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. See figure 2-2. However, the regraduated scale is much more convenient because, instead of the complicated rule (2.16), for mutually exclusive $a|d$ and $b|d$ the OR operation is now represented by a much simpler rule: for mutually exclusive propositions a and b we have the sum rule

$$\xi(a \vee b|d) = \xi(a|d) + \xi(b|d) \ . \tag{2.19}$$

[4] 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].

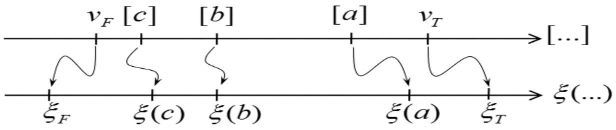

Figure 2.2: The degrees of belief $[a]$ can be regraduated, $[a] \to \xi(a)$, to another scale that is equivalent — it preserves transitivity of degrees of belief and the associativity of OR. The regraduated scale is more convenient in that it provides a more convenient representation of OR — a simple sum rule.

Thus, the new ξ 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, $[\tilde{a}|a] = \upsilon_F$, to the other extreme of total certainty, $[a|a] = \upsilon_T$. The regraduated value $\xi_F = \phi(\nu_F) + \beta$ is easy to find. Setting $d = \tilde{a}\tilde{b}$ in eq.(2.19) gives

$$\xi(a \vee b|\tilde{a}\tilde{b}) = \xi(a|\tilde{a}\tilde{b}) + \xi(b|\tilde{a}\tilde{b}) \Longrightarrow \xi_F = 2\xi_F \ , \tag{2.20}$$

and therefore

$$\xi_F = 0 \ . \tag{2.21}$$

At the opposite end, the regraduated $\xi_T = \phi(\nu_T) + \beta$ remains undetermined but if we set $b = \tilde{a}$ eq.(2.19) leads to the following normalization condition

$$\xi_T = \xi(a \vee \tilde{a}|d) = \xi(a|d) + \xi(\tilde{a}|d) \ . \tag{2.22}$$

2.2.3 The general sum rule

The restriction to mutually exclusive propositions in the sum rule eq.(2.19) 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

$$a \vee b = (ab) \vee (a\tilde{b}) \vee (\tilde{a}b) = a \vee (\tilde{a}b) \tag{2.23}$$

Since the two propositions on the right are mutually exclusive the sum rule (2.19) applies,

$$\begin{aligned} \xi(a \vee b|d) &= \xi(a|d) + \xi(\tilde{a}b|d) + [\xi(ab|d) - \xi(ab|d)] & (2.24) \\ &= \xi(a|d) + \xi(ab \vee \tilde{a}b|d) - \xi(ab|d) \ , & (2.25) \end{aligned}$$

which leads to the general sum rule,

$$\xi(a \vee b|d) = \xi(a|d) + \xi(b|d) - \xi(ab|d) \ . \tag{2.26}$$

2.2.4 Cox's proof

Understanding the proof that eq.(2.16) is the general solution of the associativity constraint, eq.(2.15), 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 (2.15) 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,

$$\frac{\partial F(x, y)}{\partial x} \geq 0 \quad \text{and} \quad \frac{\partial F(x, y)}{\partial y} \geq 0 \ . \tag{2.27}$$

Let

$$r \overset{\text{def}}{=} F(x,y) \quad \text{and} \quad s \overset{\text{def}}{=} F\left(y,z\right) \ , \tag{2.28}$$

and let partial derivatives be denoted by subscripts,

$$F_1(x,y) \overset{\text{def}}{=} \frac{\partial F\left(x,y\right)}{\partial x} \geq 0 \quad \text{and} \quad F_2(x,y) \overset{\text{def}}{=} \frac{\partial F(x,y)}{\partial y} \geq 0 \tag{2.29}$$

(F_1 denotes a derivative with respect to the first argument). Then eq.(2.15) and its derivatives with respect to x and y are

$$F\left(r,z\right) = F\left(x,s\right) \ , \tag{2.30}$$

$$F_1(r,z)F_1(x,y) = F_1(x,s) \ , \tag{2.31}$$

and

$$F_1(r,z)F_2(x,y) = F_2(x,s)F_1(y,z) \ . \tag{2.32}$$

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

$$K(x,y) = K(x,s)F_1(y,z) \ . \tag{2.33}$$

where

$$K(x,y) = \frac{F_2(x,y)}{F_1(x,y)} \ . \tag{2.34}$$

Multiplying eq.(2.33) by $K(y,z)$ and using (2.34) we get

$$K(x,y)K(y,z) = K(x,s)F_2(y,z) \ . \tag{2.35}$$

Differentiating the right hand side of eq.(2.35) with respect to y and comparing with the derivative of eq.(2.33) with respect to z, we have

$$\frac{\partial}{\partial y}\left(K\left(x,s\right)F_2\left(y,z\right)\right) = \frac{\partial}{\partial z}\left(K\left(x,s\right)F_1\left(y,z\right)\right) = \frac{\partial}{\partial z}\left(K\left(x,y\right)\right) = 0. \tag{2.36}$$

Therefore, the derivative of the left hand side of eq.(2.35) with respect to y is

$$\frac{\partial}{\partial y}\left(K\left(x,y\right)K\left(y,z\right)\right) = 0, \tag{2.37}$$

or,

$$\frac{1}{K\left(x,y\right)}\frac{\partial K\left(x,y\right)}{\partial y} = -\frac{1}{K\left(y,z\right)}\frac{\partial K\left(y,z\right)}{\partial y} \ . \tag{2.38}$$

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

$$\frac{1}{K\left(x,y\right)}\frac{\partial K\left(x,y\right)}{\partial y} \overset{\text{def}}{=} h\left(y\right) \ . \tag{2.39}$$

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

$$K(x, y) = K(x, 0)\ \exp \int_0^y h(y')dy'. \tag{2.40}$$

Similarly,

$$K\,(y, z) = K\,(0, z)\ \ \exp - \int_0^y h(y')dy'\ , \tag{2.41}$$

so that

$$K(x, y) = \frac{K(x, 0)}{H(y)} = K(0, y)H(x)\ , \tag{2.42}$$

where

$$H(x) \overset{\text{def}}{=} \exp \left[- \int_0^x h(x')dx' \right] \geq 0\ . \tag{2.43}$$

Therefore,

$$\frac{K(x, 0)}{H(x)} = K(0, y)H(y) \overset{\text{def}}{=} \alpha \tag{2.44}$$

where $\alpha = K(0, 0)$ is a constant and (2.40) becomes

$$K\,(x, y) = \alpha\, \frac{H\,(x)}{H\,(y)}\ . \tag{2.45}$$

On substituting back into eqs.(2.33) and (2.35) we get

$$F_1(y, z) = \frac{H(s)}{H(y)} \qquad \text{and} \qquad F_2(y, z) = \alpha\, \frac{H(s)}{H(z)}. \tag{2.46}$$

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

$$ds = F_1(y, z)dy + F_2(y, z)dz\ . \tag{2.47}$$

Substituting (2.46) we get

$$\frac{ds}{H(s)} = \frac{dy}{H(y)} + \alpha \frac{dz}{H(z)}\ . \tag{2.48}$$

This is easily integrated. Let

$$\phi\,(x) = \int_0^x \frac{dx'}{H(x')}\ , \tag{2.49}$$

so that $dx/H(x) = d\phi(x)$. Then

$$\phi(s) = \phi\,(F\,(y, z)) = \phi\,(y) + \alpha\phi\,(z) + \beta\ , \tag{2.50}$$

where β is an arbitrary integration constant. Therefore

$$F\left(y,z\right)=\phi^{-1}\left(\phi\left(y\right)+\alpha\phi\left(z\right)+\beta\right) \tag{2.51}$$

Substituting back into eq.(2.15) leads to $\alpha=1$. (The second possibility $\alpha=0$ is discarded because it leads to $F(y,z)=y$ which is not useful for inference.)

This completes the proof that eq.(2.16) is the general solution of eq.(2.15): Given any $F(x,y)$ that satisfies eq.(2.15) one can construct the corresponding $\phi(x)$ using eqs.(2.34), (2.39), (2.43), and (2.49). Finally, since $H(x)\geq 0$, eq. (2.43), the regraduating function $\phi(x)$, eq.(2.49), is a monotonic function of its variable x.

2.3 The product rule

Next we consider the function G in eq.(2.7) that represents AND. Once the original plausibilities are regraduated by ϕ according to eq.(2.18), the new function G for the plausibility of a conjunction reads

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

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:

$$\xi(ab|c)=G^{(1)}[\xi(a|c),\xi(b|c)] \tag{2.53}$$

$$\xi(ab|c)=G^{(2)}[\xi(a|c),\xi(a|bc)] \tag{2.54}$$

$$\xi(ab|c)=G^{(3)}[\xi(a|c),\xi(b|ac)] \tag{2.55}$$

$$\xi(ab|c)=G^{(4)}[\xi(a|bc),\xi(b|ac)] \tag{2.56}$$

$$\xi(ab|c)=G^{(5)}[\xi(a|c),\xi(b|c),\xi(a|bc)] \tag{2.57}$$

$$\xi(ab|c)=G^{(6)}[\xi(a|c),\xi(a|bc),\xi(b|ac)] \tag{2.58}$$

$$\xi(ab|c)=G^{(7)}[\xi(a|c),\xi(b|c),\xi(a|bc),\xi(b|ac)] \tag{2.59}$$

We want a function G that is of general applicability. This means that the arguments of $G^{(1)}\ldots G^{(7)}$ 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.(2.21). Derivatives are denoted with a subscript: the derivative of $G^{(3)}(x,y)$ with respect to its second argument y is $G_2^{(3)}(x,y)$.

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 $b|c$. For example, let $a = d \vee \tilde{d} = T$, then $ab = b$ and $\xi(b|c) = G^{(2)}[\xi_T, \xi_T]$ which is clearly unsatisfactory since the right hand side is a constant while b on the left hand side is quite arbitrary.

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

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

and, with the appropriate identifications,

$$G^{(5)}[G^{(5)}(x,y,z), u, G^{(5)}(v,w,s)] = G^{(5)}[x, G^{(5)}(y,u,w), s] \ . \tag{2.60}$$

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

$$G_1^{(5)}[G^{(5)}(x,y,z), u, G^{(5)}(v,w,s)] G_3^{(5)}(x,y,z) = 0 \ , \tag{2.61}$$

where $G_1^{(5)}$ and $G_3^{(5)}$ denote derivatives with respect to the first and third arguments respectively. Therefore, either

$$G_3^{(5)}(x,y,z)=0 \quad \text{or} \quad G_1^{(5)}[G^{(5)}(x,y,z),u,G^{(5)}(v,w,s)]=0\ . \tag{2.62}$$

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 $G^{(3)}$,

$$\xi(ab|c)=G[\xi(a|c),\xi(b|ac)]\ . \tag{2.63}$$

2.3.2 The distributivity constraint

The AND function G will be determined by requiring that it be compatible with the regraduated OR 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,

$$a\,(b\vee c)=ab\vee ac\ , \tag{2.64}$$

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

$$\xi\left(a\,(b\vee c)\,|d\right)=\xi\left((ab|d)\vee(ac|d)\right)\ . \tag{2.65}$$

Using eq.(2.19) and (2.63) leads to

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

which we rewrite as

$$G\left(u,v+w\right)=G\left(u,v\right)+G\left(u,w\right)\ , \tag{2.66}$$

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

To solve the functional equation (2.66) we first transform it into a differential equation. Differentiate with respect to v and w,

$$\frac{\partial^2 G\,(u, v + w)}{\partial v \partial w} = 0 \ , \tag{2.67}$$

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

$$\frac{\partial^2 G\,(u, z)}{\partial z^2} = 0 \ , \tag{2.68}$$

which shows that G is linear in its second argument,

$$G(u, v) = A(u)v + B(u) \ . \tag{2.69}$$

Substituting back into eq.(2.66) gives $B(u) = 0$. To determine the function $A(u)$ we note that $ad|d = a|d$ and therefore,

$$\xi(a|d) = \xi(ad|d) = G[\xi(a|d), \xi(d|ad)] = G[\xi(a|d), \xi_T] \ , \tag{2.70}$$

or,

$$u = A(u)\xi_T \Rightarrow A(u) = \frac{u}{\xi_T} \ . \tag{2.71}$$

Therefore

$$G\,(u, v) = \frac{uv}{\xi_T} \quad \text{or} \quad \frac{\xi\,(ab|d)}{\xi_T} = \frac{\xi\,(a|d)}{\xi_T}\frac{\xi\,(b|ad)}{\xi_T} \ . \tag{2.72}$$

The constant ξ_T is easily regraduated away: just normalize ξ to $p = \xi/\xi_T$. The corresponding regraduation of the sum rule, eq.(2.26) 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,

$$p\,(ab|d) = p\,(a|d)\ p\,(b|ad) \ , \tag{2.73}$$

and the OR operation is represented by the sum rule,

$$p\,(a \vee b|d) = p\,(a|d) + p\,(b|d) - p(ab|d) \ . \tag{2.74}$$

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 representations — that is, other regraduations — of AND and OR are possible. They would be equivalent in that they lead to the same inferences but they would also be considerably 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.(2.73) and (2.74). 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

$$p\left(a|bc\right)=p\left(a|c\right)\ . \tag{2.75}$$

For independent events the product rule simplifies to

$$p(ab|c)=p(a|c)p(b|c) \quad \text{or} \quad p(ab)=p(a)p(b)\ . \tag{2.76}$$

The symmetry of these expressions implies that $p\left(b|ac\right)=p\left(b|c\right)$ 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

$$p(a_1\vee a_2|b)=p(a_1|b)+p(a_2|b). \tag{2.77}$$

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

$$p(a_1\vee a_2\vee\cdots\vee a_n|b)=\sum_{i=1}^{n}p(a_i|b)\ . \tag{2.78}$$

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, $a_1\vee a_2\vee\cdots\vee a_n=T$, so that for any b,

$$p(T|b)=p(a_1\vee a_2\vee\cdots\vee a_n|b)=1. \tag{2.79}$$

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

$$p(T)=\sum_{i=1}^{n}p(a_i)=1\ . \tag{2.80}$$

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

$$\sum_{i=1}^{n}p(a_i|b)=1\ . \tag{2.81}$$

The proof is straightforward:

$$p(b)=p(bT)=\sum_{i=1}^{n}p(ba_i)=p(b)\sum_{i=1}^{n}p(a_i|b)\ . \tag{2.82}$$

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.(2.81) to get

$$\sum_j p(ba_j) = p(b) \sum_j p(a_j|b) = p(b) \ . \tag{2.83}$$

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 $p(ab)$.

Here is a second example. 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

$$p(b) = \sum_j p(ba_j) = \sum_j p(b|a_j)p(a_j) \ . \tag{2.84}$$

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 How about "quantum" probabilities?

Despite the enormous effort spent in understanding the peculiar behavior of quantum particles it is widely believed that our understanding of quantum mechanics has been and remains unsatisfactory. From the very beginning it was recognized that some deeply cherished principles would have to be abandoned. Foremost among these proposals is the notion that quantum effects are evidence that reasoning in the quantum domain lies beyond the reach of classical probability theory, that a new theory of "quantum" probabilities or perhaps even a quantum logic is required.

If this proposal turns out to be true then probability theory cannot, as we have claimed in previous sections, be of universal applicability. It is therefore

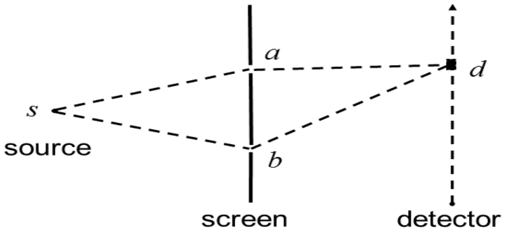

Figure 2.3: In the double slit experiment particles are generated at s, pass through a screen with two slits A and B, and are detected at the detector screen. The observed interference pattern is evidence of wave-like behavior.

necessary for us to show that quantum effects are not a counterexample to the universality of probability theory.[5]

The argument below is also valuable in other ways. First, it provides an example of the systematic use of the sum and product rules. Second, it underscores the importance of remembering that probabilities are always conditional on something and that it is often useful to be very explicit about what those conditioning statements might be. Finally, we will learn something important about quantum mechancis.

The paradigmatic example for interference effects is the double slit experiment. It was first discussed in 1807 by Thomas Young who sought a demonstration of the wave nature of light that would be as clear and definitive as the interference effects of water waves. It has also been used to demonstrate the peculiarly counter-intuitive behavior of quantum particles which seem to propagate as waves but are only detected as particles — the so-called wave-particle duality.

The quantum version of the double slit problem can be briefly stated as follows. The experimental setup is illustrated in Figure 2.3. A single particle is emitted at a source s, it goes through a screen where two slits a and b have been cut, and the particle is later detected farther downstream at some location d.

The standard treatment goes something like this: according to the rules of

[5] The flaw in the argument that quantum theory is incompatible with the standard rules for manipulating probabilities was pointed out long ago by B. O. Koopman in a paper that went largely unnoticed [Koopman 1955]. See also [Ballentine 1986].

quantum mechanics the probability that the particle is detected at d is proportional to the square of the magnitude of a complex number, the "amplitude" ψ. The quantum rules further stipulate that the amplitude for the experimental setup in which both slits a and b are open is given by the sum of the amplitude for the setup with slit a is open and b closed plus the amplitude for the setup with slit a closed and b open. This is written as

$$\psi_{ab} = \psi_a + \psi_b \ , \tag{2.85}$$

and the probability of detection at d is

$$\begin{aligned} p_{ab}(d) &\propto |\psi_{ab}|^2 = |\psi_a + \psi_b|^2 \\ &\propto |\psi_a|^2 + |\psi_b|^2 \ + \psi_a^* \psi_b + \psi_a \psi_b^* \ . \end{aligned} \tag{2.86}$$

The first term on the right $|\psi_a|^2 \propto p_a(d)$ reflects the probability of detection when only slit a is open and, similarly, $|\psi_b|^2 \propto p_b(d)$ is the probability when only b is open. The presence of the interference terms $\psi_a^* \psi_b + \psi_a \psi_b^*$ is taken as evidence that in quantum mechanics

$$p_{ab}(d) \neq p_a(d) + p_b(d) \ . \tag{2.87}$$

So far so good.

One might go further and vaguely interpret (2.87) as "the probability of paths a OR b is *not* the sum of the probability of path a plus the probability of path b". And here the trouble starts because it is then almost natural to reach the troubling conclusions that "quantum mechanics violates the sum rule", that "quantum mechanics lies outside the reach of classical probability theory", and that "quantum mechanics requires quantum probabilities." As we shall see, all these conclusions are unwarranted but in the attempt to "explain" them extreme ideas have been proposed. For example, according to the standard and still dominant explanation — the Copenhagen interpretation — quantum particles are not characterized as having definite positions or trajectories. It is only at the moment of detection that a particle acquires a definite position. Thus the Copenhagen interpretation evades the impasse about the alleged violation of the sum rule by claiming that it makes no sense to even raise the question of whether the particle went through one slit, through the other slit, through both slits or through neither.

The notion that physics is an example of inference and that probability theory is the universal framework for reasoning with incomplete information leads us along a different track. To construct our model of quantum mechanics we must first establish the subject matter:

The model: We shall assume that a "point particle" is a system characterized by its position, that the position of a particle has definite values at all times, and that the particle moves along trajectories that are continuous. Since the positions and the trajectories are in general unknown we are justified in invoking the use of probabilities.

Our goal is to show that these physical assumptions about the nature of particles can be combined with the rules of probability theory in a way that is compatible with the predictions of quantum mechanics. It is useful to introduce a notation that is explicit. We deal with the following propositions:

$s =$ “particle is generated at the source s”

$a =$ “slit a is open” and $\tilde{a} =$ “slit a is closed”

$b =$ “slit b is open” and $\tilde{b} =$ “slit b is closed”

$\alpha =$ “particle goes through slit a”

$\beta =$ “particle goes through slit b”

$d =$ “particle is detected at d”

In this model we have, for example, $p(\alpha|\tilde{a}) = 0$. The probability of interest is $p(d|sab)$. It refers to a situation in which the particle is generated at s, both slits are open, and we are uncertain about whether it is detected at d. The rules of probability theory give

$$p[(\alpha \vee \beta)d|sab] = p(\alpha d|sab) + p(\beta d|sab) - p(\alpha\beta d|sab) \tag{2.88}$$

Since particles with definite positions cannot go through both a *and* b the last term vanishes, $p(\alpha\beta d|ab) = 0$. Therefore,

$$p[(\alpha \vee \beta)d|sab] = p(\alpha d|sab) + p(\beta d|sab)\ . \tag{2.89}$$

Using the product rule the left hand side of (2.89) can be written as

$$p[(\alpha \vee \beta)d|sab] = p(d|sab)p(\alpha \vee \beta|sabd)\ . \tag{2.90}$$

Furthermore, since we assume the trajectories to be continuous, a particle that leaves s and reaches d must with certainty have passed either through a or through b, therefore $p(\alpha \vee \beta|sabd) = 1$. Therefore,

$$p[(\alpha \vee \beta)d|sab] = p(d|sab)\ , \tag{2.91}$$

so that

$$p(d|sab) = p(\alpha d|sab) + p(\beta d|sab)\ . \tag{2.92}$$

In order to compare this result with quantum mechanics, we rewrite eq.(2.87) in the new more explicit notation,

$$p(d|sab) \neq p(\alpha d|sa\tilde{b}) + p(\beta d|s\tilde{a}b)\ . \tag{2.93}$$

We can now see that probability theory, eq.(2.92), and quantum mechanics, eq.(2.93), are not in contradiction; they differ because they refer to the probabilities of different statements.

It is important to appreciate what we have shown and also what we have not shown. What we have just shown is that eqs.(2.87) or (2.93) are not in conflict with the sum rule of probability theory. What we have not (yet) shown is that the rules of quantum mechanics such as eq.(2.85) and (2.86) can be derived as an example of inference; that is a much lengthier matter that will be tackled later in Chapter 11.

We pursue this matter further to find how one might be easily misled into a paradox. Use the product rule to rewrite eq.(2.92) as

$$p(d|sab) = p(\alpha|sab)p(d|sab\alpha) + p(\beta|sab)p(d|sab\beta) \ , \tag{2.94}$$

and consider the first term on the right. To a classically trained mind (or perhaps a classically brainwashed mind) it would appear reasonable to believe that the passage of the particle through slit a is completely unaffected by whether the distant slit b is open or not. We are therefore tempted to make the substitutions

$$p(\alpha|sab) \overset{?}{=} p(\alpha|sa\tilde{b}) \quad \text{and} \quad p(d|sab\alpha) \overset{?}{=} p(d|sa\tilde{b}\alpha) \ . \tag{C1}$$

Then, making similar substitutions in the second term in (2.94), we get

$$p(d|sab) \overset{?}{=} p(\alpha|sa\tilde{b})p(d|sa\tilde{b}\alpha) + p(\beta|s\tilde{a}b)p(d|s\tilde{a}b\beta) \ , \tag{C2}$$

or

$$p(d|sab) \overset{?}{=} p(\alpha d|sa\tilde{b}) + p(\beta|s\tilde{a}b) \ . \tag{C3}$$

This equation does, indeed, contradict quantum mechanics, eq.(2.93). What is particularly insidious about the "classical" eqs.(C1-C3) is that, beyond being intuitive, there are situations in which these substitutions are actually correct — but not always.

We might ask, what is wrong with (C1-C3)? How could it possibly be otherwise? Well, it is otherwise. Equations (C1-C3) represent an assumption that happens to be wrong. The assumption does not reflect a wrong probability theory; it reflects wrong physics. It represents a piece of physical information that does not apply to quantum particles. Quantum mechanics looks so strange to classically trained minds because opening a slit at some distant location can have important effects even when the particle does not go through it. Quantum mechanics is indeed strange but this is not because it violates probability theory; it is strange because it is not local.

We conclude that there is no need to construct a theory of "quantum" probabilities. Conversely, there is no need to refer to probabilities as being "classical". There is only one kind of probability and quantum mechanics does not refute the claim that probability theory is of universal applicability.

2.6 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 the estimate of 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

$$\langle x\rangle \overset{\text{def}}{=} \sum_i p_i\, x_i \ . \tag{2.95}$$

The term 'expected' value is not always an appropriate one because it can happen that $\langle x\rangle$ is not one of the values x_i that is actually allowed; in such cases the "expected" value $\langle x\rangle$ is not a value we would expect. For example, 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 of x may be reasonable, but it is also somewhat arbitrary. Alternative estimates are possible; one could, for example, 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

$$\Delta x \overset{\text{def}}{=} \langle (x-\langle x\rangle)^2\rangle^{1/2}. \tag{2.96}$$

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.[6]

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.

[6] The interpretation of probability matters. Among the many infinities that afflict quantum field theories the variance of fields and of the corresponding energies at a point are badly divergent quantities. If these variances reflect actual physical fluctuations one should also expect those fluctuations of the spacetime geometry that are sometimes described as a spacetime foam. On the other hand, if one adopts a view of probability as a tool for inference then the situation changes significantly. One can argue that the information codified into quantum field theories is sufficient to provide successful estimates of some quantities — which accounts for the tremendous success of these theories — but is completely inadequate for the estimations of other quantities. Thus divergent variances may be more descriptive of our complete ignorance rather than of large physical fluctuations.

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 an alternative estimate x'. The alternative x' is "good" if the deviations from it are small, i.e., $\langle (x-x')^2\rangle$ is small. The condition for the "best" x' is that its variance be a minimum

$$\frac{d}{dx'}\langle (x-x')^2\rangle\bigg|_{x'_{\text{best}}}=0\ , \tag{2.97}$$

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,

$$\langle x-\langle x\rangle\rangle=0, \tag{2.98}$$

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

$$\left\langle (x-\langle x\rangle)^2\right\rangle=\langle x^2\rangle-\langle x\rangle^2. \tag{2.99}$$

The proofs are trivial — just use the definition (2.95).

2.7 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

$$\frac{N!}{m!(N-m)!}=\binom{N}{m} \tag{2.100}$$

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

$$P(m|N,\theta)=\binom{N}{m}\theta^m(1-\theta)^{N-m}. \tag{2.101}$$

This is called the binomial distribution. The range of applicability of this distribution is enormous. Whenever trials are *identical* (same probability θ in every trial) and *independent* (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.

Next we briefly review some properties of the binomial distribution. The parameter θ plays two separate roles. On one hand θ is a parameter that labels the distributions $P(m|N,\theta)$; on the other hand, we have $P(1|1,\theta)=\theta$ so that the parameter θ also happens to be the probability of α in a single trial.

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

$$\sum_{m=0}^{N} P(m|N,\theta) = \sum_{m=0}^{N} \binom{N}{m} \theta^m (1-\theta)^{N-m} = (\theta + (1-\theta))^N = 1. \tag{2.102}$$

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

$$\langle m\rangle = N\theta\,. \tag{2.103}$$

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

$$(\Delta m)^2 = \left\langle (m - \langle m\rangle)^2 \right\rangle = \langle m^2\rangle - \langle m\rangle^2\ .$$

To find $\langle m^2\rangle$,

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

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,

$$\langle m^2\rangle=(N\theta)^2+N\theta(1-\theta), \tag{2.104}$$

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

$$\Delta m=\sqrt{N\theta\,(1-\theta)}. \tag{2.105}$$

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

$$\frac{\Delta m}{\langle m\rangle}=\sqrt{\frac{(1-\theta)}{N\theta}}\propto\frac{1}{N^{1/2}}, \tag{2.106}$$

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

2.8 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 observed frequency f can take any one of the allowed values $0/N,1/N,2/N,\ldots N/N$. What is equal to θ is not the frequency itself but its expected value. Indeed, using eq.(2.103), we have

$$\langle f\rangle=\langle\frac{m}{N}\rangle=\theta\ . \tag{2.107}$$

Is this a good estimate of f? To find out use eq.(2.105) to get

$$\Delta f=\Delta\left(\frac{m}{N}\right)=\frac{\Delta m}{N}=\sqrt{\frac{\theta\,(1-\theta)}{N}}\propto\frac{1}{N^{1/2}}. \tag{2.108}$$

Therefore, for large N the distribution of frequencies is quite narrow and the probability that the observed frequency of αs differs from θ tends to zero as $N\to\infty$.

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 which we derive next. Let $\rho(x)\,dx$ be the probability that a variable X lies in the range between x and $x+dx$,

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

The variance of X satisfies

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

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

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

which is Tchebyshev's inequality,

$$P(|x-\langle x\rangle|\geq\varepsilon)\leq\left(\frac{\Delta x}{\varepsilon}\right)^2\,. \tag{2.109}$$

Next we prove Bernoulli's theorem. Consider first a special case. Let θ be the probability of outcome α in a single experiment, $P(\alpha|N=1)=\theta$. In a sequence of N independent repetitions of the experiment the probability of m outcomes α is binomial. Substituting

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

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

$$P(|f-\theta|\geq\varepsilon\,|N)\leq\frac{\theta(1-\theta)}{N\varepsilon^2}\,. \tag{2.110}$$

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 $\theta-\varepsilon$ and $\theta+\varepsilon$ tends to unity as $N\to\infty$.,

$$\lim_{N\to\infty}P(|f-\theta|\leq\varepsilon\,|N)=1\,. \tag{2.111}$$

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

$$f\text{ tends to }\theta\textit{ in probability.} \tag{2.112}$$

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 any fixed amount ε.

Thus, probabilities and frequencies are related to each other but they are not the same thing. Since $\langle f\rangle = \theta$, one might have been 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.[7]

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 $f \neq m(N)/N$ to take a limit; the limit makes no sense.

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

$$x = \frac{1}{N}\sum_{r=1}^{N} x_r \ , \tag{2.113}$$

where $x_1, \ldots, x_N$ are N independent variables with the same mean $\langle x_r\rangle = \mu$ and variance $\operatorname{var}(x_r) = (\Delta x_r)^2 = \sigma^2$. (In the previous discussion leading to eq.(2.110) 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.(2.109), we need the mean and the variance of x. Clearly,

$$\langle x\rangle = \frac{1}{N}\sum_{r=1}^{N}\langle x_r\rangle = \frac{1}{N}N\mu = \mu \ . \tag{2.114}$$

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

$$\operatorname{var}(x_1 + x_2) = \operatorname{var}(x_1) + \operatorname{var}(x_2) \ . \tag{2.115}$$

(Prove it.) Therefore,

$$\operatorname{var}(x) = \sum_{r=1}^{N}\operatorname{var}(\frac{x_r}{N}) = N\left(\frac{\sigma}{N}\right)^2 = \frac{\sigma^2}{N} \ . \tag{2.116}$$

Tchebyshev's inequality now gives,

$$P\left(|x-\mu| \geq \varepsilon|N\right) \leq \frac{\sigma^2}{N\varepsilon^2} \tag{2.117}$$

so that for any $\varepsilon > 0$

$$\lim_{N\to\infty} P\left(|x-\mu| \geq \varepsilon|N\right) = 0 \quad \text{or} \quad \lim_{N\to\infty} P\left(|x-\mu| \leq \varepsilon|N\right) = 1\,, \tag{2.118}$$

or

$$x \longrightarrow \mu \quad \textit{in probability.} \tag{2.119}$$

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.

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

Example: the simplest form of data analysis

We want to estimate a certain quantity x so we proceed to measure it. The problem is that the result of the measurement x_1 is afflicted by an error that is essentially unknown. We could just set $x \approx x_1$ but we can do much better. The procedure is to perform the measurement several times to collect data $(x_1 x_2 \ldots x_N)$. Then instead of using the result of any single measurement x_r as an estimator for x one uses the *sample average*,

$$\bar{x} = \frac{1}{N} \sum_{r=1}^{N} x_r \ . \tag{2.120}$$

The intuition behind this idea is that the errors of the individual measurements will probably be positive just as often as they are negative so that in the sum the errors will tend to cancel out. Thus one expects that the error of $\bar{x}$ will be smaller than that of any of the individual x_r.

This intuition can be put on a firmer ground as follows. Let us assume that the measurements are performed under identical conditions and are independent of each other. We also assume that although there is some unknown error the experiments are unbiased — that is, they are at least *expected* to yield the right answer. This is expressed by

$$\langle x_r \rangle = x \quad \text{and} \quad \Delta x_r = \sigma \ . \tag{2.121}$$

The sample average $\bar{x}$ is also afflicted by some unknown error so that strictly $\bar{x}$ is not the same as x but its expected value is. Indeed,

$$\langle \bar{x} \rangle = \frac{1}{N} \sum_{r=1}^{N} \langle x_r \rangle = x \ . \tag{2.122}$$

Since the measurements are independent the variances are additive

$$(\Delta \bar{x})^2 = \sum_{r=1}^{N} (\Delta \frac{x_r}{N})^2 = N(\frac{\sigma}{N})^2 = \frac{\sigma^2}{N} \quad \text{or} \quad \Delta \bar{x} = \frac{\sigma}{N^{1/2}} \ . \tag{2.123}$$

We conclude that estimating $x \approx \bar{x}$ is much better than just setting $x \approx x_1$. And the estimator $\bar{x}$ becomes better and better as $N \to \infty$. Indeed, Tchebyshev's inequality gives

$$P\left(|\bar{x} - x| \geq \varepsilon | N\right) \leq \frac{\sigma^2}{N\varepsilon^2} \quad \text{or} \quad \lim_{N\to\infty} P\left(|\bar{x} - x| \leq \varepsilon | N\right) = 1 \,, \tag{2.124}$$

so that as the number of measurements increases $\bar{x} \to x$ (in probability).

2.9 The Gaussian distribution

The Gaussian distribution is quite remarkable. It appears in an enormously 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, in diffusion phenomena, and as we shall later see, even at the very foundations of quantum mechanics. One suspects that a deeply fundamental reason must exist for its wide applicability. The Central Limit Theorem discussed below provides an explanation.

2.9.1 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 at $\langle m\rangle = N\theta$. 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

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

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. This leads to a very useful approximation — called the Stirling approximation — for the logarithm of a large factorial

$$\log m! = \sum_{n=1}^{m}\log n \approx \int_1^{m+1}\log x\,dx = \left.(x\log x - x)\right|_1^{m+1} \approx m\log m - m\ .$$

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

$$\log m! \approx m \log m - m + \frac{1}{2} \log 2\pi m + \ldots \tag{2.125}$$

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

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

The remaining unknown constant $P(\langle m \rangle)$ 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

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

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

$$P(m) = \frac{1}{\sqrt{2\pi (\Delta m)^2}} \exp\left[-\frac{(m - \langle m \rangle)^2}{2 (\Delta m)^2}\right]. \tag{2.126}$$

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

$$p(f)df = \frac{1}{\sqrt{2\pi\sigma_N^2}}\exp\left[-\frac{(f-\theta)^2}{2\,\sigma_N^2}\right]df \ , \tag{2.127}$$

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 $m\varepsilon \pm dx/2$ is

$$p(x)dx = \frac{1}{\sqrt{2\pi\,(\Delta x)^2}}\exp\left[-\frac{(x-\langle x\rangle)^2}{2\,(\Delta x)^2}\right]dx \ , \tag{2.128}$$

where $\langle x\rangle = N\theta\varepsilon$ and $(\Delta x)^2 = N\theta(1-\theta)\varepsilon^2$. 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.

2.9.2 The Central Limit Theorem

The result of the previous section can be strengthened considerably. Consider the sum

$$X = \sum_{r=1}^{N} x_r \ , \tag{2.129}$$

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

$$\langle x_r\rangle = \mu_r \quad \text{and} \quad (\Delta x_r)^2 = \sigma_r^2 \ . \tag{2.130}$$

Note that now we no longer assume that the variables x_r be identically distributed nor that the distributions $p_r(x_r)$ be binomial, but we still assume independence.

The probability density for X_N is given by the integral

$$P_N\,(X) = \int dx_1\ldots dx_N\ p_1(x_1)\ldots p_N(x_N)\,\delta\left(X - \sum_{r=1}^{N} x_r\right) \ . \tag{2.131}$$

(The expression on the right is the expected value of an indicator function. The derivation of (2.131) is left as an exercise.)

A minor annoyance is that as $N \to \infty$ the limits such as

$$\lim_{N\to\infty} \langle X\rangle_N = \lim_{N\to\infty}\sum_{r=1}^{N}\mu_r \ , \tag{2.132}$$

$$\lim_{N\to\infty} \left\langle (X-\langle X\rangle)^2\right\rangle_N = \lim_{N\to\infty}\sum_{r=1}^{N}\sigma_r^2 \ , \tag{2.133}$$

diverge in cases that are physically interesting such as when the variables x_r are identically distributed (μ_r and σ_r are independent of r). To resolve this difficulty instead of the variable X_N we will consider a different suitably shifted and normalized variable,

$$Y = \frac{X-m_N}{s_N} \ , \tag{2.134}$$

where

$$m_N \overset{\text{def}}{=} \sum_{r=1}^{N}\mu_r \quad \text{and} \quad s_N^2 \overset{\text{def}}{=} \sum_{r=1}^{N}\sigma_r^2 \ . \tag{2.135}$$

The probability distribution of Y_N is given by

$$P_N\left(Y\right) = \int dx_1\ldots dx_N\, p_1(x_1)\ldots p_N(x_N)\,\delta\left(Y-\frac{1}{s_N}\sum_{r=1}^{N}(x_r-\mu_r)\right) \ , \tag{2.136}$$

and its limit as $N\to\infty$ is given by the following theorem.

The Central Limit Theorem:

If the independent variables x_r, $r = 1\ldots N$, with means μ_r and variances σ_r^2 satisfy the *Lyapunov condition*,

$$\lim_{N\to\infty}\frac{1}{s_N^3}\sum_{r=1}^{N}\left\langle \left|x_r-\mu_r\right|^3\right\rangle = 0 \ , \tag{2.137}$$

then

$$\lim_{N\to\infty} P_N(Y) = P(Y) = \frac{1}{\sqrt{2\pi}}e^{-Y^2/2} \ , \tag{2.138}$$

which is Gaussian with zero mean and unit variance, $\langle Y\rangle = 0$ and $\Delta Y = 1$.

Proof:

Consider the Fourier transform,

$$\begin{aligned} F_N(k) &= \int_{-\infty}^{+\infty} dY\, P_N(Y)e^{ikY} \\ &= \int dx_1\ldots dx_N\ p_1(x_1)\ldots p_N(x_N)\ \exp\left[\frac{ik}{s_N}\sum_{r=1}^{N}\left(x_r-\mu_r\right)\right] \end{aligned}$$

which can be rearranged into a product of the individual Fourier transforms

$$F_N(k) = \prod_{r=1}^{N} \int dx_r\, p_r(x_r) \exp[i\frac{k}{s_N}(x_r - \mu_r)] \ . \tag{2.139}$$

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

$$f(k) = \int dx\, p(x) e^{ikx} = \left\langle e^{ikx} \right\rangle \ , \tag{2.140}$$

while the series expansion of the exponential gives

$$f(k) = \left\langle \sum_{\ell=0}^{\infty} \frac{(ikx)^\ell}{\ell!} \right\rangle = \sum_{\ell=0}^{\infty} \frac{(ik)^\ell}{\ell!} \left\langle x^\ell \right\rangle \ . \tag{2.141}$$

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

Going back to the calculation of $P_n(Y)$, eq.(2.136), its Fourier transform, eq.(2.139) is,

$$F_N(k) = \prod_{r=1}^{N} f_r(k) \ , \tag{2.142}$$

where

$$f_r(k) = \int dx_r\, p_r(x_r) \exp\left[\frac{ik}{s_N}(x_r - \mu_r)\right] \ .$$

Since s_N diverges as $N \to \infty$ we can expand

$$\begin{aligned} f_r(k) &= 1 + i\frac{k}{s_N}\langle x_r - \mu_r\rangle - \frac{k^2}{2s_N^2}\left\langle (x_r - \mu_r)^2 \right\rangle + R_r(\frac{k}{s_N}) \\ &= 1 - \frac{k^2\sigma_r^2}{2s_N^2} + R_r(\frac{k}{s_N}) \ , \end{aligned} \tag{2.143}$$

with a remainder, $R_r(k/s_N)$, bounded by

$$\left| R_r(\frac{k}{s_N}) \right| \leq C|\frac{k}{s_N}|^3 \left\langle |x_r - \mu_r|^3 \right\rangle \tag{2.144}$$

for some constant C. Therefore, for sufficiently large N, using $\log(1+x) \sim x$,

$$\log F_n(k) = \sum_{r=1}^{N} \log\, f_r(k) \tag{2.145}$$

$$= \sum_{r=1}^{N} \left(-\frac{k^2\sigma_r^2}{2s_N^2} + R_r(\frac{k}{s_N}) \right) = -\frac{k^2}{2} + \sum_{r=1}^{N} R_r(\frac{k}{s_N}) \tag{2.146}$$

By the Lyapunov condition, eq.(2.137), we have

$$\left|\sum_{r=1}^{N} R_r(\frac{k}{s_N})\right| \leq C\frac{|k|^3}{s_N^3}\sum_{r=1}^{N}\left\langle |x_r-\mu_r|^3\right\rangle \to 0 \tag{2.147}$$

as $N\to\infty$ uniformly in every finite interval $k'<k<k''$. Therefore, as $N\to\infty$

$$\log F_N(k)\to -\frac{k^2}{2} \quad \text{or} \quad F_N(k)\to F(k)=e^{-k^2/2} \tag{2.148}$$

Taking the inverse Fourier transform leads to eq.(2.138) and concludes the proof.

It is easy to check that the Lyapunov condition is satisfied when the x_r variables are identically distributed. Indeed, if all $\mu_r=\mu$, $\sigma_r=\sigma$ and $\langle |x_r-\mu|^3\rangle=\tau$ then

$$s_N^2=\sum_{r=1}^{N}\sigma_r^2=N\sigma^2 \quad \text{and} \quad \lim_{N\to\infty}\frac{1}{s_N^3}\sum_{r=1}^{N}\left\langle |x_r-\mu_r|^3\right\rangle=\lim_{N\to\infty}\frac{N\tau}{N^{3/2}\sigma^{3/2}}=0\ . \tag{2.149}$$

We can now return to our original goal of calculating the probability distribution of X_N. For any given but sufficiently large N we have

$$P_N(X)dX=P_N(Y)dY\approx\frac{1}{\sqrt{2\pi}}e^{-Y^2/2}dY\ . \tag{2.150}$$

From eq.(2.134) we have

$$dY=\frac{dX}{s_N} \tag{2.151}$$

therefore

$$P_N(X)\approx\frac{1}{\sqrt{2\pi s_N^2}}\exp-\frac{(X-m_N)^2}{2s_N^2}\ . \tag{2.152}$$

And when the x_r variables are identically distributed we get

$$P_N(X)\approx\frac{1}{\sqrt{2\pi N\sigma^2}}\exp-\frac{(X-N\mu)^2}{2N\sigma^2}\ . \tag{2.153}$$

To conclude we comment on the significance of the central limit theorem. We have shown that almost independently of the form of the distributions $p_r\,(x_r)$ the distribution of the sum X is Gaussian centered at $\sum_r\mu_r$ with standard deviation $\sum_r\sigma_r^2$. 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 Gaussian distributions: they apply to almost any 'macro-variables' (such as X) that result from adding a large number of independent 'micro-variables' (such as x_r).

But there are restrictions. Although Gaussian distributions are very common, there are exceptions. The derivation shows that the Lyapunov condition played a critical role. 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 apply to processes where the third and higher moments are not relevant information.

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 also be derived as the distribution that codifies information about the mean and the variance while remaining maximally ignorant about everything else.

2.10 Updating probabilities: Bayes' rule

Now that we have solved the problem of how to represent a state of partial knowledge as a consistent web of interconnected beliefs we can start to 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.[8]

Specifically the problem is to update our beliefs about a quantity 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.[9]

2.10.1 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

[8] The discussion below may seem unnecessarily contrived to readers who are familiar with Bayesian inference. But our goal is not to merely introduce Bayes theorem as a tool for applications in data analysis. Our goal is to present Bayesian inference in a way that provides a first stepping stone towards a more general inference framework which will eventually result in a complete unification of Bayesian and entropic methods [Caticha Giffin 2006, Caticha 2007, Caticha 2014a].

[9] On notation: it is important to distinguish priors from posteriors. Here we will denote priors by q and posteriors by p. Later, however, we will often revert to the common practice of referring to all probabilities as p. Hopefuly no confusion should arise as the correct meaning should be clear from the context.

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 (but less appropriately) the *likelihood* function. 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 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,

$$q(\theta, x) = q(\theta)q(x|\theta) \ . \tag{2.154}$$

This is important. We must be clear about what we are talking about: the relevant universe of discourse is neither the space Θ of possible θs nor the space X of possible data x. It is rather the product space $\Theta \times X$ 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 $p(\theta, x)$? Since the new data tells us that the value of x is now known to be x' the new web of beliefs is constrained to satisfy

$$p(x) = \int d\theta \, p(\theta, x) = \delta(x - x') \ . \tag{2.155}$$

(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 and this is where the prior information comes in.

2.10.2 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 the 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 ought to be revised only to the minimal extent required by the new data.*[10]

[10] The 'ought' in the PMU indicates that the design of the inference framework – the decision of how we ought to choose our beliefs – incorporates an ethical component. Pursued to its

This seems so reasonable and natural that an explicit statement may appear 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.(2.155), does not fully determine the joint distribution

$$p(\theta,x)=p(x)p(\theta|x)=\delta(x-x')p(\theta|x)\,. \tag{2.156}$$

All distributions of the form

$$p(\theta,x)=\delta(x-x')p(\theta|x')\ , \tag{2.157}$$

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

$$p(\theta,x)=\delta(x-x')q(\theta|x')\ . \tag{2.158}$$

To obtain the posterior probability for θ marginalize over x,

$$p(\theta)=\int dx\,p(\theta,x)=\int dx\,\delta(x-x')q(\theta|x')\ , \tag{2.159}$$

to get

$$p(\theta)=q(\theta|x')\ . \tag{2.160}$$

In words, the *posterior probability equals the prior conditional probability* of θ given x'. This result, which we will call 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' and 'discard' are the key words: the former reflects the PMU in action, the latter is the updating.

Using the product rule

$$q(\theta,x')=q(\theta)q(x'|\theta)=q(x')q(\theta|x')\ , \tag{2.161}$$

Bayes' rule can be written as

$$p(\theta)=q(\theta)\frac{q(x'|\theta)}{q(x')}\ . \tag{2.162}$$

The interpretation of Bayes' rule is straightforward: according to eq.(2.162) 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

ultimate conclusion it suggests that the foundations of science are closely tied to ethics. This is an important topic that deserves further exploration.

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 often called the 'evidence', is the prior probability of the data. It is given by

$$q(x') = \int q(\theta)q(x'|\theta)\, d\theta \ , \tag{2.163}$$

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 the evidence turns out to be important in problems of model selection (see eq. 2.234).

Remark: Bayes' rule is often written in the form

$$q(\theta|x') = q(\theta)\frac{q(x'|\theta)}{q(x')} \ , \tag{2.164}$$

and called Bayes' theorem.[11] This formula is very simple; but perhaps it is too simple. It is true for any value of x' whether observed or not. Eq.(2.164) is just a restatement of the product rule, eq.(2.161), and therefore it is a simple consequence of the *internal* consistency of the *prior* web of beliefs. No posteriors are involved: the left hand side is not a *posterior* but rather a *prior probability* – the prior conditional on x'. To put it differently, in an actual update, $q(\theta) \to p(\theta)$, both probabilities refer to the same proposition θ. In (2.164), $q(\theta) \to q(\theta|x')$ cannot be an update because it refers to the probabilities of two different propositions, θ and $\theta|x'$. Of course these subtleties have not stood in the way of the many extremely successful applications of Bayes theorem. But by confusing priors with posteriors the formula (2.164) has contributed to obscure the fact that an additional principle – the PMU – was needed for updating. And this has stood in the way of a deeper understanding of the connection between the Bayesian and entropic methods of inference.

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 $\theta =$ '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

[11] Neither the rule, eq.(2.162), nor the theorem, eq.(2.164), were 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.

no set of trials. The name 'Bayesian probabilities' given to 'degrees of belief' originates in the fact that it is only under this type of 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 justifies setting $q(c) = 0.01$. 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

$$q(c|y) = \frac{q(c)A}{q(c)A + (1 - q(c))\,(1 - A)} \ . \qquad (2.165)$$

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 was 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, nuisance variables 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 the new $p(\theta|x')$ to be the old $q(\theta|x')$ and yields Bayes' rule.

It may happen, however, that there is a measurement error and 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,

$$p(\theta|x) = q(\theta|x) \ , \tag{2.166}$$

and therefore, the joint posterior is

$$p(\theta, x) = P_{x'}(x)q(\theta|x) \ . \tag{2.167}$$

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

$$p(\theta) = \int dx\, P_{x'}(x)q(\theta|x) \ . \tag{2.168}$$

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 also 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 conveyed as a constraint on the allowed posterior marginal distributions $p(x)$.[12] Later, in chapter 5, we shall see how the updating rules can be generalized still further to apply to even more general constraints.

There is an alternative way to interpret (or derive) Jeffrey's rule. Just as with Bayesian updating the goal is to make an inference about θ on the basis of observed data x' and a known relation between θ and x'. The difference in this case is that the relation between θ and x' is expressed indirectly in terms of some other auxiliary variables y. We are given the relation between θ and the y variables and also the relation between y and the data x'. These relations are expressed by $q(y|\theta)$ and $q(y|x') = P_{x'}(y)$. Although we have no particular interest in these intermediate y variables they must nevertheless be included in the analysis. Since they add an additional layer of complication, they are often called *nuisance* variables. As before, the posterior $p(\theta)$ is given by Bayes rule, eq.(2.160),

$$\begin{aligned} p(\theta) = q(\theta|x') = \frac{q(\theta, x')}{q(x')} &= \frac{1}{q(x')} \int dy\; q(\theta, x', y) \\ &= \int dy\; q(\theta|x', y) q(y|x') \end{aligned} \tag{2.169}$$

Assuming that $q(\theta|x', y) = q(\theta|y)$, that is, conditional on y, θ is independent of x, and $q(y|x') = P_{x'}(y)$ we recover Jeffrey's rule, eq.(2.168).

2.10.3 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 $\{x_1, \ldots, x_n\}$.

[12] The concept of information is central to our discussions but so far we have been vague about its meaning. So here is a preview of things to come: What is information? We will continue to use the term with its usual colloquial meaning, namely, roughly, information is what you get when your question receives a satisfactory answer. But we will also need a more precise and technical definition. Later we shall elaborate on the idea that information is a constraint on our beliefs, or better, on what our beliefs ought to be if only we were ideally rational.

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

$$q(x_1, x_2, \theta) = q(\theta)q(x_1|\theta)q(x_2|\theta) = q(x_1)q(\theta|x_1)q(x_2|\theta) \ , \tag{2.170}$$

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

$$p_1(\theta) = q(\theta|x_1') = q(\theta)\frac{q(x_1'|\theta)}{q(x_1')} \ . \tag{2.171}$$

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.(2.171) now plays the role of the prior and the new posterior distribution for θ is

$$p_{12}(\theta) = p_1(\theta|x_2') = p_1(\theta)\frac{q(x_2'|\theta)}{p_1(x_2')} \ , \tag{2.172}$$

therefore

$$p_{12}(\theta) \propto q(\theta)q(x_1'|\theta)q(x_2'|\theta) \ . \tag{2.173}$$

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,

$$p_{12}(\theta) = q(\theta|x_1', x_2') \propto q(\theta)q(x_1', x_2'|\theta) \ . \tag{2.174}$$

From the symmetry of eq.(2.173) 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 commutivity 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 commutivity will continue to hold.

2.10.4 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. When tackling familiar problems where we have experience and intuition this seems to work well. But when the problems are truly new and we have neither experience nor intuition then the guessing can be risky and 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.* [Jaynes 1968][Caticha Preuss 2004] and references therein.)

Our more 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 assuming θB is no different from assuming B alone; they are epistemically equivalent. Therefore $q(x|\theta B)=q(x|B)$ and Bayes' rule gives

$$p(\theta|B)=q(\theta|B)\frac{q(x|\theta B)}{q(x|B)}=1\ . \tag{2.175}$$

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: truly prejudiced minds do not, and indeed, *cannot* question their own beliefs.

Lots of data overwhelms the prior

As more and more data are 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

$$q(x|\theta) = \prod_{r=1}^{N} q(x_r|\theta) \ , \tag{2.176}$$

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

$$p(\theta) = \frac{q(\theta)}{q(x)} q(x|\theta) = \frac{q(\theta)}{q(x)} \prod_{r=1}^{N} q(x_r|\theta) \ . \tag{2.177}$$

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

$$\frac{p(\theta_1)}{p(\theta_2)} = \frac{q(\theta_1)}{q(\theta_2)} R(x) \ , \tag{2.178}$$

where we introduced the likelihood ratios

$$R(x) \overset{\text{def}}{=} \prod_{r=1}^{N} R_r(x_r) \quad \text{and} \quad R_r(x_r) \overset{\text{def}}{=} \frac{q(x_r|\theta_1)}{q(x_r|\theta_2)} \ . \tag{2.179}$$

We will prove the following theorem:

$$\textit{given } \theta_1, \quad R(x) \to \infty \quad \textit{in probability.} \tag{2.180}$$

Equivalently, this is expressed as

$$\lim_{N\to\infty} \Pr\left(R(x) > \Lambda|\theta_1\right) = 1 \tag{2.181}$$

for any arbitrarily large positive number Λ,

The significance of the theorem is that barring two trivial exceptions the accumulation of data will drive a rational agent to become more and more convinced of the truth — in this case the truth is θ_1 — and this happens irrespective of the prior $q(\theta)$. The first exception occurs when the prior $q(\theta_1)$ vanishes which reflects a mind that is so deeply prejudiced that it is incapable of learning despite overwhelming evidence to the contrary. Such an agent can hardly be called rational. The second exception occurs when $q(x_r|\theta_1) = q(x_r|\theta_2)$ for all x_r. This represents an experiment that is so poorly designed that it offers no possibility of distinguishing between θ_1 and θ_2.

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

$$\frac{1}{N} \log R(x) = \frac{1}{N} \sum_{r=1}^{N} \log R_r(x_r) \ . \tag{2.182}$$

Since the variables $\log R_r(x_r)$ are independent, eq.(2.118) gives

$$\lim_{N\to\infty} \Pr\left(\left|\frac{1}{N}\log R(x) - K(\theta_1,\theta_2)\right| \leq \varepsilon|\theta_1\right) = 1 \tag{2.183}$$

where ε is any small positive number and

$$\begin{aligned} K(\theta_1,\theta_2) &= \left\langle \frac{1}{N}\log R(x)|\theta_1 \right\rangle \\ &= \sum_{x_r} q(x_r|\theta_1)\log R_r(x_r) \ . \end{aligned} \tag{2.184}$$

In other words,

$$\text{given } \theta_1, \quad e^{N(K-\varepsilon)} \leq R(x) \leq e^{N(K+\varepsilon)} \quad \textit{in probability.} \tag{2.185}$$

In Chapter 4 we will prove that $K(\theta_1,\theta_2) \geq 0$ which is called the Gibbs inequality. 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,

$$K(\theta_1,\theta_2) = +\sum_{x_r} q(x_r|\theta_1)\log\frac{q(x_r|\theta_1)}{q(x_r|\theta_2)} \ , \tag{2.186}$$

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* — the entropy of $q(x_r|\theta_1)$ *relative* to $q(x_r|\theta_2)$.[13] It can be interpreted as a measure of the extent that the distribution $q(x_r|\theta_1)$ can be distinguished from $q(x_r|\theta_2)$.

The marginalization Paradox***

Non-informative priors***

The Stein shrinking phenomenon***

2.11 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,[14]

$$p(\theta|E) = p(\theta)\frac{p(E|\theta)}{p(E)} \ , \tag{2.187}$$

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

[14] 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.

and one can say that the hypothesis θ is partially 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

$$\frac{p(\theta_1|E)}{p(\theta_2|E)} = \frac{p(\theta_1)}{p(\theta_2)} \frac{p(E|\theta_1)}{p(E|\theta_2)} \ . \tag{2.188}$$

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

$$R(\theta_1, \theta_2) \overset{\text{def}}{=} \frac{p(E|\theta_1)}{p(E|\theta_2)} \ . \tag{2.189}$$

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,

$$w_E(\theta_1, \theta_2) \overset{\text{def}}{=} \log \frac{p(E|\theta_1)}{p(E|\theta_2)} \ . \tag{2.190}$$

This is called the *weight of evidence* for θ_1 against θ_2 [Good 1950].[15] A useful special case is when the second hypothesis θ_2 is the negation of the first. Then

$$w_E(\theta) \overset{\text{def}}{=} \log \frac{p(E|\theta)}{p(E|\tilde{\theta})} \ , \tag{2.191}$$

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

$$\text{Odds}(\theta) \overset{\text{def}}{=} \frac{p(\theta)}{p(\tilde{\theta})} \ . \tag{2.192}$$

The posterior and prior odds are related by

$$\frac{p(\theta|E)}{p(\tilde{\theta}|E)} = \frac{p(\theta)}{p(\tilde{\theta})} \frac{p(E|\theta)}{p(E|\tilde{\theta})} \ , \tag{2.193}$$

[15] 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.

and taking logarithms we have

$$\log \mathrm{Odds}(\theta|E) = \log \mathrm{Odds}(\theta) + w_E(\theta) \; . \tag{2.194}$$

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

$$\log \frac{p(E_1E_2|\theta)}{p(E_1E_2|\tilde{\theta})} = \log \frac{p(E_1|\theta)}{p(E_1|\tilde{\theta})} + \log \frac{p(E_2|E_1\theta)}{p(E_2|E_1\tilde{\theta})}$$

so that

$$w_{E_1E_2}(\theta) = w_{E_1}(\theta) + w_{E_2|E_1}(\theta) \; . \tag{2.195}$$

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 the observation of 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.

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,

$$w_{B|C}(\theta) = \log \frac{p(B|C\theta)}{p(B|C\tilde{\theta})} = \log \frac{1}{p(B|C\tilde{\theta})} \geq 0 \; , \tag{2.196}$$

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. Using eq.(2.195) the relevant weight of evidence concerns the joint probability of B and C,

$$w_{BC}(\theta) = w_C(\theta) + w_{B|C}(\theta) \; , \tag{2.197}$$

which, as we show below, is also positive. Indeed, using Bayes' theorem,

$$w_C(\theta) = \log \frac{p(C|\theta)}{p(C|\tilde{\theta})} = \log \left(\frac{p(C)p(\theta|C)}{p(\theta)} \frac{p(\tilde{\theta})}{p(C)p(\tilde{\theta}|C)} \right) \; . \tag{2.198}$$

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 $w_C(\theta) = 0$. Recalling eq.(2.196) leads to

$$w_{BC}(\theta) \geq 0 \ . \tag{2.199}$$

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

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

because

$$w_{\tilde{C}|\tilde{B}}(\theta) = \log \frac{p(\tilde{C}|\tilde{B}\theta)}{p(\tilde{C}|\tilde{B}\tilde{\theta})} = \log \frac{1}{p(\tilde{C}|\tilde{B}\tilde{\theta})} \geq 0 \tag{2.200}$$

and

$$w_{\tilde{B}}(\theta) = \log \frac{p(\tilde{B}|\theta)}{p(\tilde{B}|\tilde{\theta})} = \log \left(\frac{p(\tilde{B})p(\theta|\tilde{B})}{p(\theta)} \frac{p(\tilde{\theta})}{p(\tilde{B})p(\tilde{\theta}|\tilde{B})} \right) = 0 \ , \tag{2.201}$$

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

But we could have additional background information that establishes a connection between θ and C. 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

$$w_{BC}(\theta_1) = w_C(\theta_1) + w_{B|C}(\theta_1) \ .$$

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

$$w_{B|C}(\theta_1) = \log \frac{p(B|C\theta_1)}{p(B|C\theta_2)} = \log \frac{1}{1 - 10^{-3}} \approx 10^{-3} \tag{2.202}$$

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

$$w_C(\theta_1) = \log \frac{p(C|\theta_1)}{p(C|\theta_2)} = \log 10^{-1} \approx -2.303 \ . \tag{2.203}$$

Therefore $w_{BC}(\theta_1) = -2.302 < 0$. 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. Or shouldn't it?

The justification of 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 — perhaps not all the time but at least most of the time. Induction is risky.

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 do 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 and not absolute certainty.

2.12 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].)

2.12.1 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,

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \ . \tag{2.204}$$

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

$$p(\mu,\sigma|x)=\frac{p(\mu,\sigma)}{p\,(x)}\prod_{i=1}^{N}p(x_i|\mu,\sigma)\ . \tag{2.205}$$

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.(2.205) 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.

Returning to the Gaussian case, let us rewrite eq. (2.205) as

$$p(\mu,\sigma|x)=\frac{p(\mu,\sigma)}{p\,(x)}\frac{1}{\left(2\pi\sigma^2\right)^{N/2}}\exp\left[-\sum_{i=1}^{N}\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right] \tag{2.206}$$

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

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i \quad\text{and}\quad s^2=\frac{1}{N}\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\ , \tag{2.207}$$

eq.(2.206) becomes

$$p(\mu,\sigma|x)=\frac{p(\mu,\sigma)}{p\left(x\right)}\frac{1}{\left(2\pi\sigma^2\right)^{N/2}}\exp\left[-\frac{\left(\mu-\bar{x}\right)^2+s^2}{2\sigma^2/N}\right] . \tag{2.208}$$

It is interesting that the data appears here only in the particular combination given in eq.(2.207) – 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. (2.208) 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

$$p(\mu|x)=\frac{p(\mu)}{p\left(x\right)}\frac{1}{\left(2\pi\sigma_0^2\right)^{N/2}}\exp\left[-\sum_{i=1}^{N}\frac{\left(x_i-\mu\right)^2}{2\sigma_0^2}\right] \tag{2.209}$$

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

$$\propto p(\mu)\exp\left[-\frac{\left(\mu-\bar{x}\right)^2}{2\sigma_0^2/N}\right] . \tag{2.210}$$

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

$$\mu=\bar{x}\pm\frac{\sigma_0}{\sqrt{N}} . \tag{2.211}$$

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 agreement with the previous result in eq.(2.123).

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. But there are two important differences: First, a frequentist approach can yield an estimator but it cannot yield a probability distribution for a parameter that is not random but merely unknown. Second, the non-Bayesian approach has no mechanism to handle additional prior information and can only proceed by ignoring it. On the other hand, the Bayesian approach has yielded a full probability distribution, eq.(2.210), and it can easily take prior information into account. For example, if, on the basis of other physical considerations, we happen to know that μ has to be positive, then we just assign $p(\mu) = 0$ for $\mu < 0$ and we calculate the estimate of μ from the truncated Gaussian in eq.(2.210).

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,

$$p(\mu, \sigma) \propto \begin{cases} C & \text{for} \quad \sigma > 0 \\ 0 & \quad \text{otherwise.} \end{cases} \tag{2.212}$$

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, in chapter 7, we shall have much more to say about this misleadingly innocuous question.

To estimate μ we return to eq.(2.206) or (2.208). For the purpose of estimating μ the variable σ is an uninteresting nuisance which, as we saw in section 2.5.4, can be eliminated through marginalization,

$$p(\mu|x) = \int_0^\infty d\sigma \, p(\mu, \sigma|x) \tag{2.213}$$

$$\propto \int_0^\infty d\sigma \, \frac{1}{\sigma^N} \exp\left[-\frac{(\mu - \bar{x})^2 + s^2}{2\sigma^2/N} \right] . \tag{2.214}$$

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

$$p(\mu|x) \propto \int_0^\infty dt \, t^{N-2} \exp\left[-\frac{t^2}{2} N \left((\mu - \bar{x})^2 + s^2 \right) \right] . \tag{2.215}$$

Repeated integrations by parts lead to

$$p(\mu|x) \propto \left[N \left((\mu - \bar{x})^2 + s^2 \right) \right]^{-\frac{N-1}{2}} , \tag{2.216}$$

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

$$\langle \mu \rangle = \bar{x} . \tag{2.217}$$

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.(2.204),

$$\left. \frac{d^2}{dx^2} \log p(x|\mu, \sigma) \right|_{x_{\max}} = -\frac{1}{\sigma^2} . \tag{2.218}$$

Therefore, to the extent that eq.(2.216) approximates a Gaussian, we can write

$$(\Delta\mu)^2 \approx \left[-\frac{d^2}{d\mu^2} \log p(\mu|x) \Big|_{\mu_{\max}} \right]^{-1} = \frac{s^2}{N-1} \; . \tag{2.219}$$

(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,

$$p(\sigma|x) = \int_{-\infty}^{\infty} d\mu \, p(\mu, \sigma|x) \tag{2.220}$$

$$\propto \frac{1}{\sigma^N} \exp\left[-\frac{Ns^2}{2\sigma^2} \right] \int_{-\infty}^{\infty} d\mu \, \exp\left[-\frac{(\mu - \bar{x})^2}{2\sigma^2/N} \right] \; . \tag{2.221}$$

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

$$p(\sigma|X) \propto \frac{1}{\sigma^{N-1}} \exp\left[-\frac{Ns^2}{2\sigma^2} \right] \; . \tag{2.222}$$

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

$$\sigma_{\max} = \sqrt{\frac{N}{N-1} s^2} \; , \tag{2.223}$$

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

$$\frac{\sigma_{\max}^2}{N} = \frac{s^2}{N-1} \; . \tag{2.224}$$

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

$$\sigma = \sigma_{\max} \pm \frac{\sigma_{\max}}{\sqrt{2\,(N-1)}} \; . \tag{2.225}$$

2.12.2 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,

$$y_i = f_\theta(x_i) + \varepsilon_i \ . \tag{2.226}$$

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 $\left\langle \varepsilon_i^2 \right\rangle = \sigma^2$ 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

$$p(y|\theta) = \prod_{i=1}^{N} p(y_i|\theta) \ , \tag{2.227}$$

where

$$p(y_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f_\theta(x_i))^2}{2\sigma^2}\right) \ . \tag{2.228}$$

Bayes' theorem gives,

$$p(\theta|y) \propto p(\theta) \exp\left[-\sum_{i=1}^{N} \frac{(y_i - f_\theta(x_i))^2}{2\sigma^2}\right] \ . \tag{2.229}$$

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

$$f(x) = a + bx \ , \tag{2.230}$$

and suppose further that being ignorant about the values of $\theta = (a, b)$ we choose $p(\theta) = p(a, b) \sim$ const, then

$$p(a, b|y) \propto \exp\left[-\sum_{i=1}^{N} \frac{(y_i - a - bx_i)^2}{2\sigma^2}\right] \ . \tag{2.231}$$

A good estimate of a and b is the value that maximizes the posterior distribution, which we recognize as the Bayesian equivalent of the method of least squares. However, the Bayesian analysis can already take us beyond the scope of the least squares method because from $p(a, b|y)$ we can also estimate the uncertainties Δa and Δb.

2.12.3 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,

$$f_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_n x^n \ , \tag{2.232}$$

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 the 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.[16]

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,

$$p(m|x) = \frac{p(m)}{p(x)} p(x|m) \ , \tag{2.233}$$

where $p(m)$ is the prior for the model.

The factor $p(x|m)$ is the *prior probability for the data* given the model and plays the role of a *likelihood function.* This is precisely the quantity which, back in eq.(2.163), we had called the 'evidence',

$$p(x|m) = \int d\theta_m p(x, \theta_m|m) = \int d\theta_m p(\theta_m|m) \ p(x|m, \theta_m) \ . \tag{2.234}$$

Thus we see that while the evidence is of no significance in the problem of estimating parameters within a given model, it turns out to be the central quantity when choosing among different models.

Substituting back into (2.233) gives

$$p(m|x) \propto p(m) \int d\theta_m p(\theta_m|m) p(x|m, \theta_m) \ . \tag{2.235}$$

Thus, the problem of model selection 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.(2.235) 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

[16] For useful references on this topic see [Balasubramanian 1996, 1997], [Rodriguez 2005].

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

$$\int d\theta_m p(\theta_m|m) = 1 \quad \text{then} \quad p(\theta_m|m) \approx \frac{1}{V_m} \ , \tag{2.236}$$

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

$$p(m|x) \propto \frac{L_{\max} \times \delta\theta_m}{V_m} \ . \tag{2.237}$$

We can now interpret eq.(2.237) 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.(2.235)!

2.12.4 Maximum Likelihood

If one adopts the frequency interpretation of probabilities then most uses of Bayes' theorem are not allowed. The reason is simple: from a frequentist perspective 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*, 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 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' or the 'estimator', that relates the two: the estimate for θ is $\hat{\theta}(x)$. The problem is to estimate the unknown θ when what is known is the sampling distribution $p(x|\theta)$ and the data x. 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 where θ appears as a fixed parameter, Fisher introduced a function of θ, which he called the likelihood function, where the observed data x appear as fixed parameters,

$$L\left(\theta|x\right) \overset{\text{def}}{=} p(x|\theta) \ . \tag{2.238}$$

Thus, the estimator $\hat{\theta}(x)$ is the value of θ that maximizes the likelihood function and, accordingly, this method of parameter estimation is called the method of 'maximum likelihood'.

The likelihood function $L(\theta|x)$ is not the probability of θ; it is not normalized in any way; and it makes no sense to use it to compute an average or a variance of θ. Nevertheless, the same intuition that leads one to propose maximization of the likelihood to estimate θ also suggests using the width of the likelihood function to estimate an error bar. Fisher's somewhat ad hoc proposal turned out to be extremely useful and it dominated the field of statistics throughout the 20th century. Its success is readily explained within the Bayesian framework.

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

$$p(\theta) = \text{const} \Rightarrow p(\theta|x) \propto p(\theta)p(x|\theta) \propto p(x|\theta) \ . \tag{2.239}$$

This is why the Bayesian discussion in this section has reproduced so many of the standard results of the 'orthodox' theory. But then the Bayesian approach has many other advantages. In addition to greater conceptual clarity, unlike the likelihood function, the Bayesian posterior is a true probability distribution that allows estimation not just of θ but of all 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 to figure out how to construct a perpetual motion machine. They all failed. Ever since a rudimentary understanding of the laws of thermodynamics was achieved in the 19th century no competent scientist would waste time considering perpetual motion. Other scientists tried to demonstrate the impossibility of perpetual motion from the established principles of mechanics. They failed too: *there exist no derivations of the Second Law from purely mechanical principles*. It took a long time, and for many the subject remains controversial, but it has gradually become clear that the reason hinges on the fact that entropy is not a physical quantity to be derived from mechanics; 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 have emerged from a combination of purely mechanical notions. If anything it should be the other way around.

In this chapter we trace some of the early developments leading to the notion of entropy. Much of this chapter (including the title) is inspired by a beautiful article by E. T. Jaynes [Jaynes 1988]. I have also borrowed from historical papers by 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 the efficiency could be improved by either changing the working substance to something other than steam or by changing the operating temperatures and pressures.

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 own father Lazare Carnot) that had ended in failure. Carnot's brilliant idea was to proceed anyway and assume what he knew was true but could not prove as the postulate from which he would draw all sorts of useful conclusions about engines.[1]

At the time Carnot did his work the nature of heat as a form of energy transfer had not yet been understood. He adopted the 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 would use 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 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 correctly interpreted as a form of energy transfer it was understood that in order to operate continuously for any significant length of time an engine would have to repeat the same cycle over and over again, always returning to same initial state. This could only be achieved if the excess heat generated in each cycle were discarded into a low temperature reservoir.

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 operating between two temperatures can be more efficient than a reversible engine E_R that operates between the same temperatures.*"

The proof of Carnot's principle is quite straightforward but because he used the caloric model it was not correct—the necessary revisions were supplied later by Clausius in 1850. As a side remark, it is interesting that Carnot's notebooks,

[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 rulers 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.

which were made public by his family in about 1870 long after his death, 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 development of thermodynamics [Wilson 1981].

The following is Clausius' proof. Figure (3.1a) shows a heat engine E that 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, $w_S > w_R$. One could split the work w_S generated by E_S into two parts w_R and $w_S - w_R$. 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 $w_S - w_R$ could then be used for any other purposes. The net result is to extract heat $q_{2R} - q_{2S} > 0$ from the cold reservoir and convert it to work without any need for the hight temperature reservoir, that is, 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. Therefore the assumption that the latter do not exist implies Carnot's principle that heat engines cannot be more efficient than reversible ones.

The statement that perpetual motion is not possible is true but it is also incomplete in one important way. It blurs the 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 include
(1) that all reversible engines operating between the same temperatures are equally efficient;
(2) that their efficiency is a function of the temperatures only,

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

and is therefore independent of all other details of how the engine is constructed and operated;
(3) that the efficiency increases with the temperature difference [see eq.(3.5) below]; and finally
(4) that 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 proofs of these statements are left as an exercise for the reader.)

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

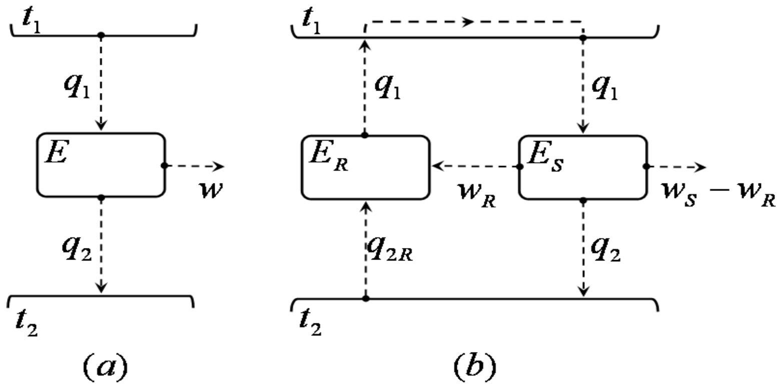

Figure 3.1: (a) An engine E operates between heat reservoirs at temperatures t_1 and t_2. (b) A perpetual motion machine can be built by coupling a super-efficient engine E_S to a reversible engine E_R.

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 transfer 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

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

and

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

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 $x_a = \log(1 - e_a)$, or $e_a = 1 - e^{x_a}$,

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

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 that the optimal choice of regraduating function, which leads to Kelvin's definition of absolute temperature, is

$$T = Cf\,(t) \ \ . \tag{3.11}$$

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,

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

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 heat, $Q_1 = q_1$, while outputs are negative heat, $Q_2 = -q_2$. 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,

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

It may be worth emphasizing that the T_i are the temperatures of the reservoirs. In an irreversible process the system will not in general be in thermal equilibrium and it may not be possible to assign a temperature to it.

The next non-trivial step, taken by Clausius, was to use eq.(3.16) to introduce the concept of entropy.

3.3 Clausius: entropy

By about 1850 both Kelvin and Clausius had realized that two laws (energy conservation and Carnot's principle) 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 in which the system is slowly taken through a continuous sequence of equilibrium states. In such a process T is both the temperature of the system and of the reservoirs. The equality implies that the integral from any state A to any other state B is independent of the path taken,

$$\oint \frac{dQ}{T} = 0 \Rightarrow \int_{A,R_{1AB}}^{B} \frac{dQ}{T} = \int_{A,R_{2AB}}^{B} \frac{dQ}{T} \ , \tag{3.18}$$

where R_{1AB} and R_{2AB} denote any two *reversible* paths linking the states A and B. Clausius realized 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

$$S_B = S_A + \int_{A,R_{AB}}^{B} \frac{dQ}{T} \ . \tag{3.19}$$

This first notion of entropy we will call the *Clausius entropy* or the *thermodynamic entropy.* Note that *the Clausius entropy is defined only for states of thermal equilibrium* which severely limits its range of applicability.

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 a reversible path. Then, the more general eq.(3.17) gives

$$\int_{A,\text{irrev}}^{B} \frac{dQ}{T} + \int_{B,R_{AB}}^{A} \frac{dQ}{T} \leq 0 \ . \tag{3.20}$$

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

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

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

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

which led Clausius to summarize the laws of thermodynamics as "*The energy of the universe is constant. The entropy of the universe tends to a maximum.*" This represents great progress: all restrictions to cyclic processes have disappeared. But a word of caution and restraint is however necessary. Glib pronouncements such as "the energy of the universe is constant" might have been the culmination of insight by 19th century standards but today we know better. It is not that the energy of the universe is increasing or decreasing, it is rather that the very notion of a total energy in a curved expanding universe is not a quantity that can be unambiguously defined. And similarly, "the entropy of the universe tends to a maximum," is a catchy phrase that captures our imagination but once one realizes that the thermodynamic entropy applies only to systems in thermal equilibrium one wonders what it could possibly mean for an expanding universe that is clearly not in equilibrium.

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. His derivation in1860 is particularly elegant because it relies on symmetry. Maxwell's first assumption is a symmetry requirement, the distribution should only depend on the actual magnitude $|\vec{v}| = v$ of the velocity and not on its direction,

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

A 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

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

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 even 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 into physics, 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 (allowing transfer of particles as well as heat) and to inhomogeneous systems (as, for example, in the equilibrium between a gas and a liquid phase). 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 correctly 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.[2] 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 that "... 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."

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

[2] For a discussion of the Gibbs' paradox from the more modern perspective of information theory see section 5.11.

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, just like Maxwell, he defined the probability of a molecule having a velocity $\vec{v}$ within a small cell d^3v as the fraction of particles with velocities within the cell. But then he also defined the probability as being proportional to the amount of time that the molecule spent within that particular cell. Both definitions have a clear origin in mechanics.

By 1868 he had managed to generalize Maxwell's work in several directions. He extended the theorem of equipartition of energy for point particles to complex molecules. And he also generalized the Maxwell distribution to particles in the presence of an external potential $U(\vec{x})$ which is the Boltzmann distribution. The latter, in modern notation, is

$$P(\vec{x},\vec{p})d^3xd^3p \propto \exp\left[-\frac{1}{kT}\left(\frac{p^2}{2m}+U(\vec{x})\right)\right]d^3xd^3p\ . \tag{3.34}$$

The argument was that in equilibrium the distribution should be stationary, that it should not change as a result of collisions among particles. The collision argument was indeed successful but it gave the distribution for individual molecules; it did not keep track of the correlations among molecules that would arise from the collisions themselves.

A better treatment of the interactions among molecules was needed and was found soon enough, also in 1868. The idea that naturally suggests itself is to replace the potential $U(\vec{x})$ due to a single external force by a potential $U(\vec{x}_1 \ldots \vec{x}_N)$ that includes all intermolecular forces. The change is enormous: it led Boltzmann to consider the probability of the microstate of the system as a whole rather than the probabilities of the microstates of individual molecules. Thus, the universe of discourse shifted from the one-particle phase space with volume element d^3xd^3p to the N-particle phase space with volume element $d^{3N}xd^{3N}p$ and Boltzmann was led to the microcanonical distribution in which the N-particle microstates are uniformly distributed over a hypersurface of constant energy — a subspace of $6N-1$ dimensions.

The question of probability was, once again, brought to the foreground. A notion of probability as the fraction of molecules in d^3v was no longer usable but Boltzmann could still identify 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 amount of time that the system would spend in the region. This obviously mechanical concept of probability is sometimes called the "time" ensemble.

Perhaps inadvertently, at least at first, Boltzmann also introduced another definition, according to which the probability that the state of the system is within a certain region of phase space at a given instant in time is proportional to the volume of the region. This is almost natural: just like the faces of a symmetric die are assigned equal probabilities, Boltzmann assigned equal probabilities to equal volumes in phase space.

At first Boltzmann did not think it was necessary to comment on whether the two definitions of probability 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 (and thus *erg*-odic). 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 the argument runs as follows. The energy of N interacting particles is given by

$$H=\sum_{i}^{N}\frac{p_i^2}{2m}+U\left(x_1,\ldots,x_N;V\right) \ , \tag{3.35}$$

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 (but not purely mechanical) terms that would correspond to the macroscopic internal energy. He opted for the "expectation"

$$E=\langle H\rangle=\int dz_N\,P_N\,H \ , \tag{3.36}$$

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

$$P_N=\frac{\exp\left(-\beta H\right)}{Z} \quad \text{where} \quad Z=\int dz_N\ e^{-\beta H} \ , \tag{3.37}$$

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

$$E=\frac{3}{2}NkT+\langle U\rangle \ . \tag{3.38}$$

The connection to the thermodynamic entropy, eq.(3.19), 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 and other types of motion such as plain displacements or rotations. The distribution function turns out to be the crucial ingredient. In any infinitesimal transformation the change in the internal energy separates into two contributions,

$$\delta E=\int dz_N\ H\delta P_N\ +\int dz_N\,P_N\delta H \ . \tag{3.39}$$

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,

$$\delta Q=\delta E-\langle\delta U\rangle \ . \tag{3.40}$$

An aside: The discrete version of this idea might be familiar from elementary quantum mechanics. If the probability that a quantum system is in a microstate i with energy eigenvalue ε_i is p_i, then the internal energy is

$$E = \langle H \rangle = \sum_i p_i \varepsilon_i \ . \tag{3.41}$$

The energy transferred in an infinitesimal process is

$$\delta E = \langle H \rangle = \sum_i \varepsilon_i \delta p_i + \sum_i p_i \delta \varepsilon_i \ . \tag{3.42}$$

The second term is the energy transfer that results from changing the energy levels while keeping the probabilities p_i fixed. This can be achieved, for example, by changing an external electric field, or by changing the volume of the box in which the particles are contained. The corresponding change in energy is called work. Then, the first term, which refers to an energy change which involves only a change in the probability of occupation δp_i and not in the energy levels is called heat. Thus, in quantum mechanics, a slow process in which the quantum system makes no transitions ($\delta p_i = 0$) while the energy levels are moved around ($\delta \varepsilon_i \neq 0$) is called an *adiabatic* process (*i.e.*, no heat is transferred).

Getting back to Boltzmann, substituting δE from eq.(3.38) into (3.40), one gets

$$\delta Q = \frac{3}{2} Nk\delta T + \delta \langle U \rangle - \langle \delta U \rangle \ . \tag{3.43}$$

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

$$\frac{\delta Q}{T} = \delta \left[\frac{3}{2} Nk \log T + \frac{\langle U \rangle}{T} + k \log \left(\int d^{3N}x \, e^{-\beta U} \right) + \text{const} \right] \ . \tag{3.44}$$

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

$$S = \frac{E}{T} + k \log Z + \text{const} \, , \tag{3.45}$$

which is recognized as the correct modern expression. Indeed, the free energy, $F = E - TS$, is such that $Z = e^{-F/kT}$.

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 labelled $n = 1, ..., m$. 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 we call a microstate, is defined by specifying the cell of each individual molecule. A macrostate is specified less completely by the distribution function, $w_1, \ldots, w_m$.

The probability of the macrostate is given by the probability of a microstate $1/m^N$ multiplied by the number W of microstates compatible with a given macrostate, which is called the "multiplicity",

$$P(w_1, \ldots, w_m) = \frac{W}{m^N} \quad \text{where} \quad W = \frac{N!}{w_1! \ldots w_m!} \ . \tag{3.46}$$

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 any microstate is as likely as any other – the 'equal a priori probability postulate'. Thus, the most probable macrostate is that which maximizes P or equivalently W subject to the constraints of a fixed total number of particles N and a fixed total energy E,

$$\sum_{n=1}^{m} w_n = N \quad \text{and} \quad \sum_{n=1}^{m} w_n \varepsilon_n = E. \tag{3.47}$$

where ε_n is the energy of a particle in the nth cell.

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

$$\log W = N \log N - N - \sum_{n=1}^{m} (w_n \log w_n - w_n) \ , \tag{3.48}$$

which, using $N = \sum w_n$, can be written as

$$\log W = -N \sum_{n=1}^{m} \frac{w_n}{N} \log \frac{w_n}{N} \ , \tag{3.49}$$

or

$$\log W = -N \sum_{n=1}^{m} f_n \log f_n \ , \tag{3.50}$$

where $f_n = w_n/N$ is the fraction of molecules in the nth cell, 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 (3.47) is

$$f_n = \frac{w_n}{N} \propto e^{-\beta\varepsilon_n} \ , \tag{3.51}$$

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 the one-particle phase space. Therefore

$$\log W = -N \int dz_1 \, f(x,p) \log f(x,p) \ , \tag{3.52}$$

where $dz_1 = d^3x d^3p$ and the most probable distribution (3.51) is the same equilibrium distribution found earlier by Maxwell and generalized by Boltzmann.

The derivation of the Boltzmann distribution (3.51) from a purely probabilistic argument is a major accomplishment. However, although minimized, the role of dynamics it is not completely 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 is proportional to the multiplicity W provided the microstates are assigned equal probabilities, or equivalently, equal volumes in phase space are assigned equal a priori weights. As always, equal probabilities must at ultimately be justified in terms of some form of underlying symmetry. As we shall later see in chapter 5, the required symmetry follows from Liouville's theorem – under a Hamiltonian time evolution a region in phase space moves around and its shape is distorted but its volume remains 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 requirement that the dynamics be Hamiltonian but the functional form of the Hamiltonian is not important.

It is very surprising 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

$$S = k \log W \ . \tag{3.53}$$

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

There is, however, a serious problem with eq.(3.52): it involves the distribution function $f(x,p)$ in the one-particle phase space and therefore it cannot take correlations into account. Indeed, Boltzmann used his eq.(3.52) in the one case where it actually works, for ideal gases of non-interacting particles. The expression that applies to systems of interacting particles is[3]

$$\log W_G = - \int dz_N \ f_N \log f_N \ , \tag{3.54}$$

[3] 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 section 5.11.

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.(3.52) and (3.54) is enormous; it goes well beyond the issue of intermolecular interactions. The probability in Eq.(3.52) is the single-particle distribution, it can be interpreted as a "mechanical" property, namely, the relative number of molecules in each cell and then the entropy Eq.(3.52) can be interpreted as a mechanical property of the system. In contrast, eq.(3.54) involves the N-particle distribution which is not a property of any single individual system but at best a property of an ensemble of similarly prepared 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 thinks of probabilities in terms of frequencies. If the only available notion of probability requires an ensemble and real ensembles are nowhere to be found then either one gives up on probabilistic arguments altogether or one invents an imaginary ensemble. Gibbs opted for the second alternative.

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 a story, not a history. I have not mentioned many developments that are central to 20th century physics—for example, the Boltzmann equation, or the ergodic hypothesis, or all applications of statistical mechanics to the macroscopic properties of matter, phase transitions, transport properties, and so on and on. These topics 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. The goal in this chapter was to discuss the origins of thermodynamics and statistical mechanics in order to provide some background for the first historical examples of such an *entropic* 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 neat 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), Brillouin (1952), 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.

The decades following Gibbs (1902) were extremely fruitful for statistical mechanics but they centered in the systematic development of calculational methods and their application to a bewildering range of systems and phenomena, including the extension to the quantum domain by Bose, Einstein, Fermi, Dirac, and von Neumann. With the possible exception of Szilard there were no significant conceptual advances concerning the connection of entropy and information until the work of Shannon, Brillouin, and Jaynes around 1950. In this book we will approach statistical mechanics from the point of view of information and entropic inference. For an entry point to the extensive literature on alternative approaches based, for example, on Boltzmann's equation, the ergodic hypothesis, etc., see *e.g.* [Ehrenfest 2012] [ter Haar 1955] [Wehrl 1978] [Mackey 1989][Lebowitz 1993, 1999] and [Uffink 2001, 2003, 2006].

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' 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 a rational agent 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.

What the rules of probability theory allow us to do is to assign probabilities to some "complex" propositions on the basis of the probabilities of some other, perhaps more "elementary", propositions. The problem is that in order to get the machine running one must first assign probabilities to those elementary propositions. How does one do this?

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, Brillouin 1952, 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 some incomplete information we have somehow assigned probabilities p_i. In order to figure out which is the actual state within the set $\{i\}$ we need more information. The question we address here is how much more information is needed. Note that we are not asking which particular piece of information is missing; we are merely asking the *quantity* of information that is missing. It seems reasonable that the amount of missing information in a sharply peaked distribution is smaller than the amount missing in a broad distribution, but how much smaller?[1] Is it possible to quantify the notion of amount of information? Can one find a function $S[p]$ of the probabilities 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. The restriction to probabilities defined over a discrete space of alternatives is an important limitation. 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 missing in p relative to the reference distribution q.

Axiom 2. If all the p_i's are equal, $p_i = 1/n$. Then $S = S(1/n, \ldots, 1/n)$ is some function $F(n)$ that 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. It also means 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 $S[p]$ is meant to measure the amount of additional information beyond the incomplete information already codified in the p_i that will be needed to pinpoint the actual state of the system. Imagine that this missing informa-

[1] If probabilities are subjective then this intuition is itself questionable. The subjective probablities could be sharply peaked around a completely wrong value and the actual amount of missing information could be substantial.

tion 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.

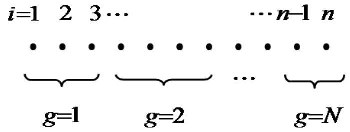

Figure 4.1: The n states are divided into N groups to formulate the grouping axiom.

Imagine the n states are divided into N groups labeled by $g = 1 \dots N$ as shown in Fig 4.1. 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 probability that the system is in state i conditional on its being in group g. For $i \in g$ we have

$$p_i = p_{ig} = P_g p_{i|g} \quad \text{so that} \quad p_{i|g} = \frac{p_i}{P_g} \, . \tag{4.2}$$

Suppose we were to obtain the missing 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 which is 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 $g = 1 \dots N$. 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. Then the *expected amount of missing information* to take us from the gs to the actual i 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 $S = \langle \ldots \rangle$ is the expected value of something, it is not clear, at this point, what that something and its interpretation would be . We will return to this below.

The solution to Shannon's constraints is obtained in two steps. The power of Shannon's axioms arises from their universality; they are meant to hold for all choices of n and N, for all probability distributions, and for all possible groupings. 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,

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

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

and

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

Then, axiom 3 gives

$$F\left(mN\right) = F\left(N\right) + F\left(m\right) \ . \tag{4.7}$$

This should be true for all integers N and m. It is easy to see that one solution of this equation is

$$F\left(m\right) = k \log m \ , \tag{4.8}$$

where k is any positive constant (just substitute), but it is also easy to see that eq.(4.7) has infinitely many other solutions. To single out (4.8) as the unique solution we must further impose the additional requirement that $F(m)$ be monotonic increasing in m (axiom 2).

The uniqueness proof that we give below is due to [Shannon Weaver 1949] (see also [Jaynes 2003]). Its details might not be of interest to most readers and may be skipped. First we show that (4.8) is not the only solution of eq.(4.7). 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

$$F\left(m\right) = \textstyle\sum_r \alpha_r F(q_r) \tag{4.9}$$

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). 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

$$F\,(n) = S_G[P] + \sum_g P_g\, F(m_g), \tag{4.15}$$

Therefore,

$$S_G[P] = F\,(n) - \sum_g P_g F\,(m_g) = \sum_g P_g\ [F\,(n) - F\,(m_g)] \ . \tag{4.16}$$

Substituting our previous expression for F we get

$$S_G[P] = \sum_g P_g\, k\, \log \frac{n}{m_g} = -k\, \sum_g P_g\, \log P_g \, . \tag{4.17}$$

Therefore Shannon's quantitative measure of the amount of missing information, the entropy of the probability distribution $p_1, \ldots, p_n$ is

$$S[p] = -k\, \sum_{i=1}^{n} p_i\, \log\, p_i \, . \tag{4.18}$$

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 the Grouping Axiom which makes explicit reference to two entropies $S[p]$ and $S_G[P]$ referring to two different descriptions of the same system — a fine-grained and a coarse-grained description. 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 by *the* entropy one means the particular entropy obtained 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 $k_B = 1.38 \times 10^{-16}\,\text{erg/K}$ 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

$$S[p] = -\sum_{i=1}^{N} p_i \, \log_2 \, p_i \ . \tag{4.19}$$

The base of the logarithm is 2, and the entropy is said to measure information in units called 'bits'.

Next we turn to the question of interpretation. Earlier we mentioned that from the Grouping Axiom 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 would be $\log 1/p_i$. But this is not always satisfactory because it clashes with the intuition that in general large messages will carry large amounts of information while short messages will carry small amounts. Indeed, 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 physically conveyed by a very short one bit message, namely "0". This shows that interpreting $\log 1/p$ as an *actual amount* of information is not quite right. It may perhaps be better to interpret $\log 1/p$ as a measure of how unexpected or how surprising the piece of information is. Some authors do just this and call $\log 1/p_i$ the "surprise" of i, but then the direct interpretation of S as an amount of expected "information" is lost.

The standard practice consists of defining the technical term 'information' as whatever is measured by (4.18). There is nothing wrong with this — definitions are not true or false, they are just more or less useful. Suppose we interpret $S[p]$ as the 'lack of information' or the 'uncertainty' implicit in p — here the term 'uncertainty' is used as synonymous to 'lack of information' so that more information implies less uncertainty. Unfortunately, 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 information has increased the entropy rather than decreased it. (This question is further discussed in section 4.7.)

The point of these remarks is not to suggest that there is something wrong with the mathematical derivation — there is not, eq.(4.18) does follow from the axioms. The point rather is to suggest caution when interpreting S. In fact, at this point the notion of information itself is too imprecise, too vague. Any attempt to define its amount 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? The remarks above constitute a warning that this technical meaning of information as whatever is measured by S does not coincide with the more colloquial meaning of information as something that induces us to change our minds.

In their later writings both Shannon and Jaynes agreed that one should not place too much significance on the axiomatic derivation of eq.(4.18), that its use can be *fully* justified a posteriori by its formal properties, for example, by the various inequalities it satisfies. This position can, however, be criticized 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.

On one hand, interpreting entropy as an amount of missing information is a convenient and intuitive shortcut — much like interpreting probability as a frequency, or interpreting temperature as a measure of expected kinetic energy. It is usually safe as long as one is aware of its limitations. On the other hand, the problem of interpretation can be quite serious because as long as the notion of information is kept imprecise and vague it is possible to introduce other measures of information. Indeed, such measures have been introduced by Renyi and by Tsallis, creating a whole industry of alternative theories [Renyi 1961, Tsallis 1988]. If the ultimate goal is to design a framework for inference this situation is very unsatisfactory: whenever one can reach a conclusion using Shannon's entropy, one can equally well reach different conclusions using any one of Renyi-Tsallis entropies. Which, among all those alternatives, should one choose? This is a problem to which we will return in chapter 6.

The two-state case

To gain intuition about $S[p]$ consider the case of a variable that can take two values. The paradigmatic 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

$$S(p) = -p \log p - (1-p) \log (1-p) \ , \tag{4.20}$$

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 $S_{\max} = \log 2$.

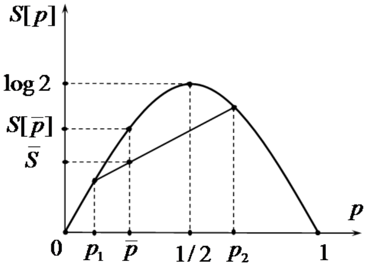

Figure 4.2: Showing the concavity of the entropy $S(\bar{p}) \geq \bar{S}$ for the case of two states.

An important set of properties follows from the concavity of the entropy which itself follows from the concavity of the logarithm.

Suppose we are told the biased coin was drawn from a box that contains a fraction q of coins for which the probability of heads is p_1 and the remaining fraction $1-q$ are coins for which the probability of heads is p_2. The probability of heads for a random coin is given by the mixture

$$\bar{p} = q p_1 + (1-q) p_2 \ . \tag{4.21}$$

The concavity of entropy implies that the entropies corresponding to p_1, to p_2, and to $\bar{p}$ satisfy the inequality

$$S(\bar{p}) \geq qS\left(p_1\right) + (1-q)\, S\left(p_2\right) = \bar{S} \ , \tag{4.22}$$

with equality in the extreme cases where $p_1 = p_2$, or $q = 0$, or $q = 1$. The interpretation is that if all we know is $\bar{p}$ then the amount of missing information is given by $S(\bar{p})$. If, in addition to knowing $\bar{p}$, we are further told that $\bar{p}$ is the result of mixing p_1 and p_2 with probabilities q and $1-q$, then we actually know more and this leads to the inequality $\bar{S} \leq S(\bar{p})$.

4.2 Relative entropy

The following entropy-like quantity, which we earlier met in eq.(2.186),

$$K[p,q] = +\sum_i p_i \, \log \frac{p_i}{q_i} \ , \tag{4.23}$$

turns out to be useful. Despite the positive sign K is sometimes read as the 'entropy of p relative to q,' and often 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.(4.23) becomes

$$K[p,1/n] = \sum_i p_i \, (\log p_i + \log n) = \log n - S[p] \ . \tag{4.24}$$

The relative entropy is also known by many other names including information divergence, information for discrimination, and Kullback-Leibler divergence [Kullback 1959]. The expression (4.23) 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, eq.(2.190) [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 $K[p,1/n]$ *as the amount of information in* p*.*

As we saw in section (2.11) the weight of evidence factor in favor of hypothesis θ_1 against θ_2 provided by data x is

$$w(\theta_1 : \theta_2) \overset{\text{def}}{=} \log \frac{p(x|\theta_1)}{p(x|\theta_2)} \ . \tag{4.25}$$

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,

$$\int dx\, p(x|\theta_1) \log \frac{p(x|\theta_1)}{p(x|\theta_2)} = K(\theta_1, \theta_2) \ , \tag{4.26}$$

is interpreted as *the expected amount of information per observation drawn from* $p(x|\theta_1)$ *in favor of* θ_1 *against* θ_2. Any such interpretations can be heuristically useful but they ultimately suffer from the same conceptual difficulties mentioned earlier concerning the Shannon entropy. Later, in chapter 6, we shall see that these interpretational difficulties can be avoided and that the relative entropy turns out to be the fundamental quantity for inference – indeed, more fundamental, more general, and therefore, more useful than the Shannon entropy itself. (We will also redefine it with a negative sign, $S[p,q] \overset{\text{def}}{=} -K[p,q]$, so that it includes thermodynamic entropy as a special case.) In this chapter we just derive some properties and consider some applications.

The Gibbs inequality – An important property of the relative entropy is the Gibbs inequality,

$$K[p,q] \geq 0 \ , \tag{4.27}$$

with equality if and only if $p_i = q_i$ for all i. The proof uses the concavity of the logarithm,

$$\log x \leq x - 1 \ . \tag{4.28}$$

(The graph of the curve $y = \log x$ lies under the straight line $y = x - 1$.) Therefore

$$\log \frac{q_i}{p_i} \leq \frac{q_i}{p_i} - 1 \ , \tag{4.29}$$

which implies

$$\sum_i p_i \log \frac{q_i}{p_i} \leq \sum_i (q_i - p_i) = 0 \ . \tag{4.30}$$

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 suggestive, this language is not 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 $K[p+\delta p, p]$ turns out to be symmetric and satisfies all the requirements of a metric. Indeed, up to a constant factor, it is the only natural Riemannian metric on the manifold of probability distributions. It is variously known as the Fisher metric, the Fisher-Rao metric and more commonly as the information metric.

The two inequalities $S[p] \geq 0$ and $K[p,q] \geq 0$ together with eq.(4.24) imply

$$0 \leq S[p] \leq \log n \ , \tag{4.31}$$

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 Sufficiency*

4.4 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

$$S_x \overset{\text{def}}{=} S[p_x] = -\sum_x p_x \log p_x \ . \tag{4.32}$$

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,

$$-\sum_x p_x \log p_x = -\sum_i p_x(x_i) \log p_x(x_i) \ . \tag{4.33}$$

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

$$S_{xy} = -\sum_{xy} p_{xy} \log p_{xy} \ . \tag{4.34}$$

When the variables x and y are independent, $p_{xy} = p_x p_y$, the joint entropy is *additive*

$$S_{xy} = -\sum_{xy} p_x p_y \log(p_x p_y) = S_x + S_y \ , \tag{4.35}$$

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,

$$\mathrm{var}(x+y) = \mathrm{var}(x) + \mathrm{var}(y) \ . \tag{4.36}$$

In thermodynamics additivity leads to *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 (4.35) can be generalized into an inequality. Consider the joint distribution $p_{xy} = p_x p_{y|x} = p_y p_{x|y}$. The entropy K of p_{xy} relative to the product distribution $p_x p_y$

that would represent uncorrelated variables is given by

$$\begin{aligned} K[p_{xy}, p_x p_y] &= \sum_{xy} p_{xy} \, \log \frac{p_{xy}}{p_x p_y} \\ &= -S_{xy} - \sum_{xy} p_{xy} \, \log p_x - \sum_{xy} p_{xy} \, \log p_y \\ &= -S_{xy} + S_x + S_y \ . \end{aligned} \tag{4.37}$$

Therefore, the Gibbs inequality, $K \geq 0$, leads to

$$S_{xy} \leq S_x + S_y \ , \tag{4.38}$$

with the equality holding when the two variables x and y are independent. The inequality (4.38) is referred to as *subadditivity.* Its interpretation is clear: entropy increases when information about correlations among subsystems is discarded.

4.5 Conditional entropy and mutual information

Consider again two variables x and y. We want to measure the amount of information about one variable x when we have some limited information about some other 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

$$S_{x|y} = \sum_y p_y S[p_{x|y}] = -\sum_y p_y \sum_x p_{x|y} \, \log \, p_{x|y} = -\sum_{xy} p_{xy} \, \log \, p_{x|y} \ , \tag{4.39}$$

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

$$\log \, p_{xy} = \log \, p_y + \log \, p_{x|y} \ , \tag{4.40}$$

and take the expectation over x and y to get

$$S_{xy} = S_y + S_{x|y} \ . \tag{4.41}$$

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,

$$S_{xy} \geq S_{x|y} \ . \tag{4.42}$$

A related 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.(4.37),

$$M_{xy} \overset{\text{def}}{=} K[p_{xy}, p_x p_y] = \sum_{xy} p_{xy} \log \frac{p_{xy}}{p_x p_y} \\ = S_x + S_y - S_{xy} \geq 0 \ , \tag{4.43}$$

where we used eq.(4.37). Note that M_{xy} is symmetrical in x and y. Using eq.(4.41) the mutual information is related to the conditional entropies by

$$M_{xy} = S_x - S_{x|y} = S_y - S_{y|x} \ . \tag{4.44}$$

An important application of mutual information to the problem of experimental design is given below in section 4.7.

4.6 Continuous distributions

Shannon's derivation of the expression for entropy, eq.(4.18), applies to probability distributions of discrete variables. The generalization to continuous variables is not 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

$$-\int dx\, p(x) \log p(x) \tag{4.45}$$

is unsatisfactory. A change of variables $x \to y = y(x)$ changes the probability density $p(x)$ to $p'(y)$ but the actual probabilities do not change, $p(x)dx = p'(y)dy$. The problem is that the transformation $x \to y$ does not represent a loss or gain of information so the entropy should not change either. However, one can check that (4.45) is not invariant,

$$-\int dx\, p(x) \log p(x) = -\int dy\, p'(y) \log \left[p'(y) \left| \frac{dy}{dx} \right| \right] \\ \neq -\int dy\, p'(y) \log p'(y) \, . \tag{4.46}$$

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

$$p_n = p(x_n)\Delta x \ , \tag{4.47}$$

where $x_n = x_a + n\Delta x$ and n is an integer. The discrete entropy is

$$S_N = -\sum_{n=1}^{N} \Delta x\, p(x_n) \, \log\left[p(x_n)\Delta x\right] \ , \tag{4.48}$$

and as $N \to \infty$ we get

$$S_N \to \log N - \int_{x_a}^{x_b} dx\, p(x) \, \log \left[\frac{p(x)}{1/\left(x_b - x_a\right)} \right] \tag{4.49}$$

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 might be tempted to consider

$$-\int_{x_a}^{x_b} dx\, p(x) \, \log \left[\frac{p(x)}{1/\left(x_b - x_a\right)} \right] \tag{4.50}$$

as a candidate for the continuous entropy, until we realize that, except for an additive constant, it coincides with the unacceptable expression (4.45) 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

$$-\int_{y_a}^{y_b} dy\, p'(y) \, \log \left[\frac{p'(y)}{1/\left(y_b - y_a\right)} \right] \ . \tag{4.51}$$

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 (4.50) represents a probability density that is constant in the variable x, but not in y. Similarly, the density $1/\left(y_b - y_a\right)$ in (4.51) is constant 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 about the particular problem at hand we must identify a privileged set of coordinates that will define what we mean by equal intervals or by equal volumes. Equivalently, we must identify one preferred probability distribution $\mu(x)$ we are willing to define as uniform — where by "uniform" we mean a distribution that assigns equal probabilities to equal volumes. Then, and only then, it makes sense to propose the following definition

$$S[p,\mu] \overset{\text{def}}{=} -\int_{x_a}^{x_b} dx\, p(x) \, \log \frac{p(x)}{\mu(x)} \ . \tag{4.52}$$

It is easy to check that this is invariant,

$$\int_{x_a}^{x_b} dx\, p(x) \, \log \frac{p(x)}{\mu(x)} = \int_{y_a}^{y_b} dy\, p'(y) \, \log \frac{p'(y)}{\mu'(y)} \ . \tag{4.53}$$

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 about what we mean by equal volumes: express x in Cartesian coordinates, that is, replace x by the triple (x, y, z)

and choose $\mu(x,y,z)$ =constant. If we were to translate into spherical coordinates the corresponding $\mu'(r,\theta,\phi)$ would no longer be *constant*, but we would still call it *uniform* because it assigns equal probabilities to equal volumes.

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 $d\ell^2 = g_{ij}dx^i dx^j$, and 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,

$$\mu(x)d^D x \propto (\det g)^{1/2}\, d^D x \ . \tag{4.54}$$

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 volumes. The corresponding choice of a μ that is uniform in phase space is called the postulate of "equal a priori probabilities." (See the discussion in section 5.2.)

Notice that the expression in eq.(4.52) is a relative entropy $-K[p,\mu]$. 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. Relative entropy is the more fundamental quantity — a theme that will be fully developed in chapter 6. Indeed there is no difficulty in obtaining the continuum limit from the discrete version of relative entropy. We can check that

$$K_N = \sum_{n=0}^{N} p_n \log \frac{p_n}{q_n} = \sum_{n=0}^{N} \Delta x\, p(x_n) \log \frac{p(x_n)\Delta x}{q(x_n)\Delta x} \tag{4.55}$$

has a well defined limit,

$$K[p,q] = \int_{x_a}^{x_b} dx\, p(x) \log \frac{p(x)}{q(x)} \ , \tag{4.56}$$

which is manifestly invariant under coordinate transformations.

4.7 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 address now 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 do we 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 seeks to maximize subject to the appropriate constraints. The basic idea was proposed in [Lindley 1956]; a more modern application 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 $q_\varepsilon(x_\varepsilon|\theta)$.

The amount of information before the experiment is performed is given by

$$K_b = K[q,\mu] = \int d\theta\, q(\theta)\, \log\frac{q(\theta)}{\mu(\theta)}\ , \tag{4.57}$$

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_ε were obtained the amount of information after the experiment would be

$$K_a(x_\varepsilon) = K[q_\varepsilon,\mu] = \int d\theta\, q_\varepsilon(\theta|x_\varepsilon)\, \log\frac{q_\varepsilon(\theta|x_\varepsilon)}{\mu(\theta)}\ . \tag{4.58}$$

But all decisions must be made before the data x_ε is available; the expected amount of information to be obtained from experiment ε is

$$\langle K_a\rangle = \int dx_\varepsilon\, q_\varepsilon(x_\varepsilon)\int d\theta\, q_\varepsilon(\theta|x_\varepsilon)\, \log\frac{q_\varepsilon(\theta|x_\varepsilon)}{\mu(\theta)}\ , \tag{4.59}$$

where $q_\varepsilon(x_\varepsilon)$ is the prior probability that data x_ε is observed in experiment ε,

$$q_\varepsilon(x_\varepsilon) = \int d\theta\, q_\varepsilon(x_\varepsilon,\theta) = \int d\theta\, q(\theta) q_\varepsilon(x_\varepsilon|\theta)\,. \tag{4.60}$$

Using Bayes theorem $\langle K_a\rangle$ can be written as

$$\langle K_a\rangle = \int dx_\varepsilon d\theta\, q_\varepsilon(x_\varepsilon,\theta)\, \log\frac{q_\varepsilon(x_\varepsilon,\theta)}{q_\varepsilon(x_\varepsilon)q(\theta)} + K_b\ . \tag{4.61}$$

Therefore, the expected information gained in experiment ε, which is $\langle K_a\rangle - K_b$, turns out to be

$$M(\varepsilon) = \int dx_\varepsilon d\theta\, q_\varepsilon(x_\varepsilon,\theta)\, \log\frac{q_\varepsilon(x_\varepsilon,\theta)}{q_\varepsilon(x_\varepsilon)q(\theta)}\ , \tag{4.62}$$

which we recognize as the mutual information of the data from experiment ε and the variable θ to be inferred, eq.(4.43). Clearly the best experiment is that which maximizes $M(\varepsilon)$ subject to whatever conditions (*e.g.*, limited resources, etc.) apply to the situation at hand.

Incidentally, the mutual information, eq.(4.43), satisfies the Gibbs inequality $M(\varepsilon) \geq 0$. Therefore, unless the data x_ε and the variable θ are statistically independent (which represents a totally useless experiment because 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 samples of 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 example 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_ε represents light intensity; and ε represents the time of observation and the orientation of the telescope.

The mutual information $M(\varepsilon)$ can be written in terms of conditional entropy as in eq.(4.44). Explicitly,

$$\begin{aligned} M(\varepsilon) &= \int dx_\varepsilon d\theta \, q_\varepsilon(x_\varepsilon, \theta) \, \log \frac{q_\varepsilon(x_\varepsilon|\theta)}{q_\varepsilon(x_\varepsilon)} \\ &= \int dx_\varepsilon d\theta \, q_\varepsilon(x_\varepsilon, \theta) \left[\log \frac{q_\varepsilon(x_\varepsilon|\theta)}{\mu(x_\varepsilon)} - \log \frac{q_\varepsilon(x_\varepsilon)}{\mu(x_\varepsilon)} \right] \\ &= \int d\theta \, q(\theta) \int dx_\varepsilon \, q_\varepsilon(x_\varepsilon|\theta) \, \log \frac{q_\varepsilon(x_\varepsilon|\theta)}{\mu(x_\varepsilon)} - \int dx_\varepsilon \, q_\varepsilon(x_\varepsilon) \, \log \frac{q_\varepsilon(x_\varepsilon)}{\mu(x_\varepsilon)} \ , \end{aligned}$$

where $\mu(x_\varepsilon)$ defines what we mean by the uniform distribution in the space of x_εs. 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

$$M(\varepsilon) = \text{const} - \textstyle\int dx_\varepsilon \, q_\varepsilon(x_\varepsilon) \, \log \dfrac{q_\varepsilon(x_\varepsilon)}{\mu(x_\varepsilon)} = \text{const} + S_x(\varepsilon) \qquad (4.63)$$

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.

4.8 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. Shannon's great insight was that, as far as the engineer is concerned, 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. The expected number of zeros and ones is Np and $N(1-p)$ respectively. For large N the probability of any one of these *typical* sequences is approximately

$$P(x_1, \ldots, x_N) \approx p^{Np}(1-p)^{N(1-p)} \ , \tag{4.64}$$

so that

$$-\log P(x_1, \ldots, x_N) \approx -N[p\log p - (1-p)\log(1-p)] = NS(p) \tag{4.65}$$

where $S(p)$ is the two-state entropy, eq.(4.20), the maximum value of which is $S_{\max} = \log 2$. Therefore, the probability of typical sequences is roughly

$$P(x_1, \ldots, x_N) \approx e^{-NS(p)} \, . \tag{4.66}$$

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

$$e^{NS(p)} = 2^{NS(p)/\log 2} = 2^{NS(p)/S_{\max}} \tag{4.67}$$

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

$$-\frac{1}{N}\log P(x_1, \ldots, x_N) \to S[p] \quad \text{in probability.} \tag{4.68}$$

Proof: If the variables x_i are independent, so are functions of them such the logarithms of their probabilities, $\log p(x_i)$,

$$-\frac{1}{N}\log P(x_1, \ldots, x_N) = -\frac{1}{N}\sum_i^N \log p(x_i)\,, \tag{4.69}$$

and the law of large numbers (see section 2.8) gives

$$\lim_{N\to\infty} \text{Prob}\left[\left|-\frac{1}{N}\log P(x_1, \ldots, x_N) + \langle \log p(x)\rangle\right| \le \varepsilon\right] = 1\,, \tag{4.70}$$

where

$$-\langle \log p(x)\rangle = S[p]\ . \tag{4.71}$$

This concludes the proof.

We can elaborate on the AEP idea further. The typical sequences are those for which eq.(4.66) or (4.68) 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

$$e^{-N[S(p)+\varepsilon]} \le P(x_1, \ldots, x_N) \le e^{-N[S(p)-\varepsilon]}\,. \tag{4.72}$$

Theorem of typical sequences:

(1) For N sufficiently large $\text{Prob}[A_{N,\varepsilon}] > 1 - \varepsilon$.

(2) $|A_{N,\varepsilon}| \le e^{N[S(p)+\varepsilon]}$ where $|A_{N,\varepsilon}|$ is the number of sequences in $A_{N,\varepsilon}$.

(3) For N sufficiently large $|A_{N,\varepsilon}| \ge (1-\varepsilon)e^{N[S(p)-\varepsilon]}$.

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... at least most of them).

Proof: Eq.(4.70) states that for fixed ε, for any given δ there is an N_δ such that for all $N > N_\delta$, we have

$$\text{Prob}\left[\left|-\frac{1}{N}\log P(x_1,\ldots,x_N) + S[p]\right| \le \varepsilon\right] \ge 1-\delta\,. \tag{4.73}$$

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

$$\begin{aligned} 1 \ge \text{Prob}[A_{N,\varepsilon}] &= \sum_{(x_1,\ldots,x_N)\in A_{N,\varepsilon}} P(x_1,\ldots,x_N) \\ &\ge \sum_{(x_1,\ldots,x_N)\in A_{N,\varepsilon}} e^{-N[S(p)+\varepsilon]} = e^{-N[S(p)+\varepsilon]}\,|A_{N,\varepsilon}|\ . \end{aligned} \tag{4.74}$$

Finally, from part **(1)**,

$$\begin{aligned} 1-\varepsilon < \text{Prob}[A_{N,\varepsilon}] &= \sum_{(x_1,\ldots,x_N)\in A_{N,\varepsilon}} P(x_1,\ldots,x_N) \\ &\le \sum_{(x_1,\ldots,x_N)\in A_{N,\varepsilon}} e^{-N[S(p)-\varepsilon]} = e^{-N[S(p)-\varepsilon]}\,|A_{N,\varepsilon}|\ , \end{aligned} \tag{4.75}$$

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.*

4.9 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. The idea is simple: It is just as important to seek truth as to avoid error. Wishful thinking is not allowed: we ought to assign probabilities that do not reflect more knowledge than you actually have. More explicitly:

> *Among all possible probability distributions we ought to adopt the distribution that represents what we do in fact know while honestly reflecting ignorance about all else that we do not know.*

The mathematical implementation of this idea involves entropy:

> *Since least information is expressed as maximum entropy, the preferred 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.[2] Ultimately, the method of maximum entropy expresses an ethical principle of intellectual honesty that demands that one should not assume information one does not have. This justification of the MaxEnt method is compelling but it relies on interpreting entropy as a measure of missing information and therein lies its weakness: are we sure that entropy is the unique measure of information or of uncertainty? This flaw will be addressed and resolved later in chapter 6.

As a first example 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

$$0 = \delta \left[S[p] - \alpha \left(\textstyle\sum_i p_i - 1 \right) \right] = - \textstyle\sum_i \left(\log\, p_i + 1 + \alpha \right) \delta p_i \ , \tag{4.76}$$

so that independent variations δp_i lead to

$$\log\, p_i + 1 + \alpha = 0 \qquad \text{or} \qquad p_i = e^{-1-\alpha} \ , \tag{4.77}$$

which agrees with the intuition that maximum uncertainty is described by a uniform distribution. The multiplier α is determined from the normalization constraint. The result is

$$p_i = \frac{1}{n} \ . \tag{4.78}$$

We can check that the corresponding entropy,

$$S_{\max} = -\sum_i \frac{1}{n} \log \frac{1}{n} = \log n \,, \tag{4.79}$$

[2] The presentation in these sections (4.9) and (4.10) follows the pioneering work of E.T. Jaynes [Jaynes 1957b, 1957c] and particularly [Jaynes 1963]. Other relevant papers are reprinted in [Jaynes 1983] and collected online at http://bayes.wustl.edu.

is the maximum value allowed by eq.(4.31).

Remark: The distribution of maximum ignorance turns out to be uniform and 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 because from the very beginning the Shannon axioms give a privileged status to the uniform distribution. It would be more appropriate to say that the method of maximum entropy has been designed so as to reproduce Laplace's principle.

4.10 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. To simplify the notation we assume that the variables x are discrete, $x=x_i$ for $i=1\ldots n$, and we set $f^k(x_i)=f_i^k$.

To maximize $S[p]$ subject to the constraints

$$\left\langle f^k\right\rangle = \textstyle\sum_i p_i f_i^k = F^k \quad \text{with} \quad k=1,2,\ldots\ , \tag{4.80}$$

and the normalization, $\sum p_i = 1$, introduce Lagrange multipliers α and λ_k,

$$\begin{aligned} 0 &= \delta\left(S[p]-\alpha\textstyle\sum_i p_i - \lambda_k \textstyle\sum_i p_i f_i^k\right) \\ &= -\textstyle\sum_i \left(\log\, p_i + 1 + \alpha + \lambda_k f_i^k\right)\delta p_i\ , \end{aligned} \tag{4.81}$$

where we adopt the Einstein summation convention that repeated upper and lower indices are summed over. Independent variations δp_i lead to the so-called 'canonical' distribution,

$$p_i = \exp -(\lambda_0 + \lambda_k f_i^k)\ , \tag{4.82}$$

where we have set $1+\alpha=\lambda_0$. The normalization constraint determines λ_0,

$$e^{\lambda_0} = \textstyle\sum_i \exp(-\lambda_k f_i^k) \overset{\text{def}}{=} Z\left(\lambda_1,\lambda_2,\ldots\right) \tag{4.83}$$

where we have introduced the so-called "partition" function $Z(\lambda)$. The remaining multipliers λ_k are determined by eqs.(4.80): substituting eqs.(4.82) and (4.83) into eqs.(4.80) gives

$$-\frac{\partial \log Z}{\partial \lambda_k} = F^k\ . \tag{4.84}$$

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.(4.82) into $S[p] = -\sum p_i \log p_i$ yields the value of the maximized entropy,

$$S_{\max} = \textstyle\sum_i p_i(\lambda_0 + \lambda_k f_i^k) = \lambda_0 + \lambda_k F^k \ . \tag{4.85}$$

Equations (4.82-4.84) 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 (4.85) is indeed the absolute maximum rather than just a local extremum or a stationary point. Consider any other distribution p'_i that satisfies the same constraints (4.80). According to the basic Gibbs inequality for the entropy of p' relative to the canonical p is

$$K(p', p) = \textstyle\sum_i p'_i \log \dfrac{p'_i}{p_i} \geq 0 \ , \tag{4.86}$$

or

$$S[p'] \leq -\textstyle\sum_i p'_i \log p_i \ . \tag{4.87}$$

Substituting eq.(4.82) and using the fact that p'_i satisfies the same constraints (4.80) gives

$$S[p'] \leq \textstyle\sum_i p'_i(\lambda_0 + \lambda_k f_i^k) = \lambda_0 + \lambda_k F^k \ . \tag{4.88}$$

Therefore, recalling (4.85), we have

$$S[p'] \leq S[p] = S_{\max} \ . \tag{4.89}$$

In words: within the family $\{p\}$ of all distributions that satisfy the constraints (4.80) the distribution that achieves the maximum entropy is the canonical distribution p given in eq.(4.82).

Having found the maximum entropy distribution we can now develop the MaxEnt formalism along lines that closely parallel statistical mechanics. Each distribution within the family of distributions of the form (4.82) can be thought of as a point in a continuous space — the "statistical 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 also associate a number, the value of the maximized entropy. Therefore, $S_{\max}(F^1, F^2, \ldots) = S_{\max}(F)$ is a scalar field on the statistical manifold.

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 quantity

$$\lambda_0 = \log Z(\lambda_1, \lambda_2, \ldots) = \log Z(\lambda) \tag{4.90}$$

is sometimes called the "free energy" because it is closely related to the thermodynamic free energy ($Z = e^{-\beta F}$). The quantities $S(F)$ and $\log Z(\lambda)$ are Legendre transforms of each other,

$$S(F) = \log Z(\lambda) + \lambda_k F^k . \tag{4.91}$$

They contain the same information and therefore just as the Fs are obtained from $\log Z(\lambda)$ from eq.(4.84), the λs can be obtained from $S(F)$,

$$\frac{\partial S(F)}{\partial F^k} = \lambda_k \ . \tag{4.92}$$

The proof is straightforward: write

$$\frac{\partial S(F)}{\partial F^k} = \frac{\partial \log Z(\lambda)}{\partial \lambda_j} \frac{\partial \lambda_j}{\partial F^k} + \frac{\partial \lambda_j}{\partial F^k} F^j + \lambda_k \ , \tag{4.93}$$

and use eq.(4.84). Equation (4.92) shows that the multipliers λ_k are the components of the gradient of the entropy $S(F)$ on the manifold of canonical distributions. Thus, the change in entropy when the constraints are changed by δF^k while the functions f^k held fixed is

$$\delta S = \lambda_k \delta F^k \ . \tag{4.94}$$

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 a particular f^k refers to the energy of the system, then the parameter v could represent the volume of the system or perhaps an externally applied magnetic field. A general change in the expected value F^k can be induced by changes in both f^k and λ_k,

$$\delta F^k = \delta \left\langle f^k \right\rangle = \textstyle\sum_i \left(p_i \delta f_i^k + f_i^k \delta p_i \right) \ . \tag{4.95}$$

The first term on the right is

$$\left\langle \delta f^k \right\rangle = \textstyle\sum_i p_i \dfrac{\partial f_i^k}{\partial v} \delta v = \left\langle \dfrac{\partial f^k}{\partial v} \right\rangle \delta v \ . \tag{4.96}$$

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,

$$\delta W^k \overset{\text{def}}{=} \left\langle \delta f^k \right\rangle = \left\langle \frac{\partial f^k}{\partial v} \right\rangle \delta v \, . \tag{4.97}$$

The second term in eq.(4.95),

$$\delta Q^k \overset{\text{def}}{=} \textstyle\sum_i f_i^k \delta p_i = \delta \left\langle f^k \right\rangle - \left\langle \delta f^k \right\rangle \tag{4.98}$$

is a kind of "generalized heat", and

$$\delta F^k = \delta W^k + \delta Q^k \tag{4.99}$$

is a "generalized first law." However, there is no implication that the quantities f^k are conserved (e.g., energy is a conserved quantity but magnetization is not).

The corresponding change in the entropy is obtained from eq.(4.91),

$$\begin{aligned}\delta S &= \delta \log Z(\lambda) + \delta(\lambda_k F^k) \\ &= -\frac{1}{Z}\textstyle\sum_i \left[\delta\lambda_k f_i^k + \lambda_k \delta f_i^k\right] e^{-\lambda_k f_i^k} + \delta\lambda_k F^k + \lambda_k \delta F^k \\ &= \lambda_k \left(\delta \left\langle f^k \right\rangle - \left\langle \delta f^k \right\rangle\right) ,\end{aligned} \tag{4.100}$$

which, using eq.(4.98), gives

$$\delta S = \lambda_k \delta Q^k \ . \tag{4.101}$$

It is easy to see that this is equivalent to eq.(4.92) 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.(4.98)] this result is a triviality: the amount of information in a distribution cannot change when the probabilities do not change,

$$\delta p_i = 0 \Rightarrow \delta Q^k = 0 \Rightarrow \delta S = 0 \ . \tag{4.102}$$

4.11 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 broader question "What is information?" shall be addressed in more detail in section 6.1 (see also [Caticha 2007, 2014a]). The MaxEnt 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 its numerical value F? When can we expect the inferences to be reliable?[3]

When using the MaxEnt method to obtain, say, the canonical Boltzmann distribution it has been common to adopt the following language:

[3] This material follows the presentation in [Caticha 2012a].

> 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.

The issue raised by **O1** strikes at the very heart of what physical theories are supposed to be and what purpose they are meant to serve. For the sake of argument let us grant that there is such a thing as an external reality (why not?), that actual phenomena out there are what they are independently of our thoughts about them. Then the issue raised by **O1** is whether the purpose of our theories is to provide *models that faithfully mirror this external reality* or whether the connection to reality is considerably more indirect and the *models are merely pragmatic tools for manipulating information about reality* for the purposes of prediction, control, explanation, etc.

In the former case, **O1** is a legitimate objection because if theories mirror reality then the information available to us should play no role. In the latter case, however, the purpose of the theory is not to mirror reality. It is still true that external realities remain independent of whatever information we might have, but the theory is just a means to produce models and make predictions. The objection **O1** originates in a failure to recognize that in this latter conception of 'theory' the success of our models and predictions — including the successful modeling and prediction of the black body spectrum — depends critically on possessing the right information, the relevant information.

To address objections **O2** and **O3** it is useful to distinguish four epistemically different types of constraints:

(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 highly informative 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 much less informative 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 for 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 blackbody 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 highly *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 unknown numerical value of $\langle\varepsilon\rangle = E$ handled as a free parameter. This leads us to a *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.

However, 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.[4]

(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 one should 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.

4.12 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? No. What they are is valuable opportunities for learning. They teach us how to use the method and warn us about how

[4] It is this case C that Jaynes adopted for his foundations of entropic inference leading to what he called predictive statistical mechanics. See [Jaynes 1963, 1986].

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 8.3).

4.12.1 MaxEnt cannot fix flawed information

An important point is that the issue of how a piece of information was obtained in the first place should not be confused with the issue of how that piece of information is to be processed. 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 is uncertain, whether the messenger is reliable — 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 then, and only then, 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 to do it.

4.12.2 MaxEnt cannot supply missing information

It is not uncommon that we may find ourselves in situations where our intuition insists that our MaxEnt inferences are not right — and this also applies to all other forms of inference, whether based on Bayes' rule or on entropy and its generalizations (see chapter 6). The right way to proceed is to ask: why do we feel that something is wrong?

The answer is that we must have some expectations which we have not yet fully recognized that cause our intuitions to clash with the inferences from MaxEnt. Further analysis will inevitably indicate that either those expectations were misguided — we have here an opportunity to educate our intuition and learn. Or, alternatively, the analysis might vindicate the earlier expectations. This tells us that we had additional prior information that happened to be relevant but, not having recognized it, we neglected to incorporate it into the MaxEnt analysis. Either way, the right way to handle such situations is not to blame the method: first blame the user.

4.12.3 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. Expected values are epistemic; sample values are not.

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 examples of using the MaxEnt method to process incorrect information.

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. We might, for example, try

$$C_1 \approx \frac{1}{n} \textstyle\sum_i x_i \ . \tag{4.103}$$

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 for 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 raw 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. Later, in chapter 6, we shall return to the problem of processing information in the form of data using entropic methods. The answer, unsurprisingly, will establish the deep connection between Bayesian and entropic inference.

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 $f(x)$ as being "relevant." 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

$$p(x_\alpha|\lambda) = \frac{e^{-\lambda f_\alpha}}{Z} \quad \text{where} \quad Z = \sum_{\alpha=1}^{k} e^{-\lambda f_\alpha} \tag{4.104}$$

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,

$$p(\lambda|D) = \frac{p(\lambda)}{p(D)} \prod_{i=1}^{n} \frac{e^{-\lambda f_i}}{Z} \ , \tag{4.105}$$

where $p(\lambda)$ is the prior. It is convenient to consider the logarithm of the posterior,

$$\begin{aligned} \log p(\lambda|D) &= \log \frac{p(\lambda)}{p(D)} - \sum_{i=1}^{n} (\log Z + \lambda f_i) \\ &= \log \frac{p(\lambda)}{p(D)} - n(\log Z + \lambda \bar{f}) \ , \end{aligned} \tag{4.106}$$

where $\bar{f}$ is the sample average,

$$\bar{f} = \frac{1}{n} \sum_{i=1}^{n} f_i \ . \tag{4.107}$$

The value of λ that maximizes the posterior $p(\lambda|D)$ is such that

$$\frac{\partial \log Z}{\partial \lambda} + \bar{f} = \frac{1}{n} \frac{\partial \log p(\lambda)}{\partial \lambda} \ , \tag{4.108}$$

or, using (4.84),

$$\langle f \rangle = \bar{f} - \frac{1}{n} \frac{\partial \log p(\lambda)}{\partial \lambda} \ . \tag{4.109}$$

As $n \to \infty$ we see that the optimal λ is such that $\langle f \rangle \to \bar{f}$. This is to be expected: for large n the data overwhelms the prior $p(\lambda)$ and $\bar{f}$ tends to $\langle f \rangle$ (in probability). But the result eq.(4.109) 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 $g(x)$.[5]

[5] 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

> *"There is no description or model that does not also reflect an interest or a purpose."*
>
> Anonimous[1]

> *"... but it is important to note that the whole content of the theory depends critically on just what we mean by 'probability'."*
>
> E. T Jaynes [1957c]

Among the various theories that make up what we call physics, statitical mechanics and thermodynamics hold a very special place because they provided the first example [Jaynes 1957b, 1957c, 1963, 1965] 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.[2]

[1] I read this somewhere. Who wrote it? Probably H. Putnam or perhaps W. James. If they did not then I will claim it as my own.

[2] We approach statistical mechanics from the point of view of entropic inference. (See

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, starting in chapter 11, we will revisit this question and show that quantum mechanics and its Newtonian limit are themselves derivable 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 point in the N-particle phase space, $z = (\vec{x}_1, \vec{p}_1 \ldots \vec{x}_N, \vec{p}_N)$ with coordinates z^α, $\alpha = 1 \ldots 6N$. 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 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,

$$\frac{\partial}{\partial t} f(z,t) = -\nabla_\alpha J^\alpha(z,t)\ , \tag{5.3}$$

where J^α is the probability current,

$$J^\alpha(z,t) = f(z,t)\dot{z}^\alpha \tag{5.4}$$

also [Balian 1991, 1992, 1999].) For an entry point to the extensive literature on alternative approaches based, for example, on the ergodic hypothesis see *e.g.* [Ehrenfest 1912] [Khinchin 1949] [ter Haar 1955] [Wehrl 1978] [Mackey 1989] [Lebowitz 1993, 1999] and [Uffink 2001, 2003, 2006]. For a discussion of why ergodic arguments are irrelevant to statistical mechanics see [Earman Redei 1996].

with

$$\dot{z} = \left(\dots \frac{d\vec{x}_i}{dt}, \frac{d\vec{p}_i}{dt} \dots \right) = \left(\dots \frac{\partial H}{\partial \vec{p}_i}, -\frac{\partial H}{\partial \vec{x}_i} \dots \right) \tag{5.5}$$

Since

$$\nabla_\alpha \dot{z}^\alpha = \sum_{i=1}^{N} \left(\frac{\partial}{\partial \vec{x}_i} \cdot \frac{\partial H}{\partial \vec{p}_i} - \frac{\partial}{\partial \vec{p}_i} \cdot \frac{\partial H}{\partial \vec{x}_i} \right) = 0 \ , \tag{5.6}$$

evaluating the divergence in eq.(5.3) gives

$$\frac{\partial f}{\partial t} = -\dot{z}^\alpha \nabla_\alpha f = -\sum_{i=1}^{N} \left(\frac{\partial f}{\partial \vec{x}_i} \cdot \frac{\partial H}{\partial \vec{p}_i} - \frac{\partial f}{\partial \vec{p}_i} \cdot \frac{\partial H}{\partial \vec{x}_i} \right) \overset{\text{def}}{=} \{H, f\} \ . \tag{5.7}$$

where $\{H, f\}$ is Poisson bracket. 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.(5.7) 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,

$$\frac{d}{dt} f\left(z(t),t\right) = \frac{\partial}{\partial t} f(z,t) + \dot{z}^\alpha \nabla_\alpha f \ . \tag{5.8}$$

Using (5.7) we see that

$$\frac{d}{dt} f\left(z(t),t\right) = 0 \ , \tag{5.9}$$

which means that f is constant along a flow line. Explicitly,

$$f\left(z(t),t\right) = f\left(z(t'),t'\right) \ . \tag{5.10}$$

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,

$$\frac{d}{dt} \operatorname{Prob}[\Delta z(t)] = \frac{d}{dt} [\Delta z(t) f\left(z(t),t\right)] = 0 \ . \tag{5.11}$$

But $f\left(z(t),t\right)$ itself is constant, eq.(5.9), therefore

$$\frac{d}{dt} \Delta z(t) = 0 \ , \tag{5.12}$$

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 4.6, we pointed out that a proper definition of entropy in a continuum, eq.(4.52), requires that one specify a privileged background measure $\mu(z)$,

$$S[f,\mu]=-\int dz\, f(z)\log\frac{f(z)}{\mu(z)}\ , \tag{5.13}$$

where $dz=d^{3N}xd^{3N}p$ in Cartesian coordinates. 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. (At one point Boltzmann himself tried this. Needless to say, the idea did not succeed and he abandoned it.) In this section we want to *derive* $\mu(z)$ by proving the following theorem

> **Theorem on Equal a Priori Probabilities:** *Since* a deterministic 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

$$S(t)=-\int dz\, f(z,t)\log\frac{f(z,t)}{\mu(z)}\ . \tag{5.14}$$

The first input from Hamiltonian dynamics is that information is not lost and therefore we must require that $S(t)$ be constant,

$$\frac{d}{dt}S(t)=0\ . \tag{5.15}$$

Therefore,

$$\frac{d}{dt}S(t)=-\int dz\,\left[\frac{\partial f(z,t)}{\partial t}\log\frac{f(z,t)}{\mu(z)}+\frac{\partial f(z,t)}{\partial t}\right]\ . \tag{5.16}$$

The second term vanishes,

$$\int dz\,\frac{\partial f(z,t)}{\partial t}=\frac{d}{dt}\int dz\, f(z,t)=0\ . \tag{5.17}$$

A second input from Hamiltonian dynamics is that probabilities are not merely conserved; they are locally conserved. This is expressed by eqs.(5.3) and (5.4). The first term of eq.(5.16) can be rewritten,

$$\frac{d}{dt}S(t) = \int dz\, \nabla_\alpha (f\dot{z}^\alpha) \log \frac{f}{\mu} \ , \tag{5.18}$$

so that integration by parts (the surface term vanishes) gives

$$\frac{d}{dt}S(t) = -\int dz\, f\dot{z}^\alpha \nabla_\alpha \log \frac{f}{\mu} = \int dz\, \left[-\dot{z}^\alpha \nabla_\alpha f + f\dot{z}^\alpha \nabla_\alpha \log \mu \right] \ . \tag{5.19}$$

Hamiltonian dynamics enters here once again: the first term vanishes by Liouville's equation (5.7),

$$-\int dz\, \dot{z}^\alpha \nabla_\alpha f = \int dz\, \frac{\partial f}{\partial t} = 0 \ , \tag{5.20}$$

and therefore, imposing (5.15), leads to

$$\frac{d}{dt}S(t) = \int dz\, f\dot{z}^\alpha \nabla_\alpha \log \mu = 0 \ . \tag{5.21}$$

This condition must hold for any arbitrary choice of $f(z,t)$, therefore

$$\dot{z}^\alpha \nabla_\alpha \log \mu(z) = 0 \ . \tag{5.22}$$

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 $\dot{z}^\alpha$ at will. The time derivative dS/dt must still vanish irrespective of the particular choice of the vector $\dot{z}^\alpha$. We conclude that

$$\nabla_\alpha \log \mu(z) = 0 \quad \text{or} \quad \mu(z) = \text{const}\,. \tag{5.23}$$

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,

$$\frac{d}{dt}S(t) = 0 \ , \tag{5.24}$$

and this singles out the Gibbs entropy,

$$S(t) = -\int dz\, f(z,t) \log f(z,t) \ , \tag{5.25}$$

(in $6N$-dimensional configuration space) as the correct *information* entropy.

It is sometimes objected that (5.24) implies that the Gibbs entropy (5.25) cannot be identified with the *thermodynamic* entropy of Clausius, eq.(3.19),

because this would be in contradiction with the Second Law.[3] This is true but it should not an objection; it is further evidence that there is no such thing as *the* unique entropy of a system. Different entropies attach to different descriptions of the system and, as we shall see in Section 5.7, equations (5.24) and (5.25) will turn out to be crucial elements in the derivation of the Second Law.
Remark: In section 4.1 we pointed out that the interpretation of entropy $S[f,\mu]$ 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 $S[f,\mu]$ requires no interpretation; it is a tool designed for updating from a prior μ to a posterior f distribution. More explicitly the entropy $S[f,\mu]$ is introduced to rank candidate distributions f according to some criterion of "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.(5.15) so the proof of the Equal a Priori Theorem proceeds exactly as above.

5.3 The constraints for thermal equilibrium

Thermodynamics is mostly concerned with situations of thermal equilibrium. What is the relevant information needed to make inferences in these special cases?[4] A problem here is that the notion of relevance is relative — a particular piece of information might be relevant for one specific question and irrelevant for 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 $\{f,H\}=0$ 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. (The symmetry of most situations is such that the same probability will be assigned to molecules moving to the left as to those moving to the right.) To simplify the situation even more we will

[3] See *e.g.* [Mackey 1989].

[4] The presentation here follows [Caticha 2008]. See also [Lee and Presse 2012].

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 4.10 and the remarks in 4.11 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 4.11.)

Now we are ready to deploy the MaxEnt method. The argument depends crucially on using a measure $\mu(z)$ that is constant in the phase-space variables. Maximize the entropy,

$$S[f] = -\int dz\, f(z) \log f(z) \ , \tag{5.26}$$

of the composite system C subject to normalization and the fixed energy constraint,

$$f(z) = 0 \qquad \text{if} \qquad \varepsilon(z) \neq \varepsilon_C \ . \tag{5.27}$$

To simplify the discussion it is convenient to divide phase space into discrete cells a of equal a priori probability. By the theorem of section 5.2 these cells are of equal phase-space volume Δz. Then we can use the discrete entropy,

$$S = -\sum_c p_c \log p_c \qquad \text{where} \qquad p_c = f(z_c)\Delta z \ . \tag{5.28}$$

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 $\Omega^B(\varepsilon_b)$. We assume that the microstates c of the composite system, $C = A + B$, are labelled by specifying the state of A and the state of B, $c = (a, b)$. This condition looks innocent but this may be deceptive; it implies A and B are not quantum mechanically entangled. Our relevant information also includes the fact that A and B interact very weakly, that is, any interaction potential V_{ab} depending on both microstates a and b can be neglected. The interaction must be weak but cannot be strictly zero, just barely enough to attain equilibrium. It is this condition of weak interaction that justifies us in talking about a system A separate from the bath B. Under these conditions the total energy ε_c constrains the allowed microstates of $C = A + B$ to the subset that satisfies

$$\varepsilon_a + \varepsilon_b = \varepsilon_c \ . \tag{5.29}$$

The total number of such microstates is

$$\Omega^C(\varepsilon_c) = \sum_a \Omega^B(\varepsilon_c - \varepsilon_a) \ . \tag{5.30}$$

At this point 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 $\Omega^C(\varepsilon_c)$ allowed microstates. This is precisely the problem tackled in section 4.9. Thus, the distribution of maximum entropy gives $p_c = 0$ when $\varepsilon_b \neq \varepsilon_c - \varepsilon_a$, and is uniform, eq.(4.78), over the allowed microstates – those with $\varepsilon_b = \varepsilon_c - \varepsilon_a$. The probability of any allowed microstate of C is $1/\Omega^C(\varepsilon_c)$, and the corresponding entropy is $S^C = \log \Omega^C(\varepsilon_c)$. 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

$$p_a = \sum_b p_{ab} = \sum_{\{b|\varepsilon_b = \varepsilon_c - \varepsilon_a\}} \frac{1}{\Omega^C(\varepsilon_c)} \ , \tag{5.31}$$

where the sum is over all states b with energy $\varepsilon_b = \varepsilon_c - \varepsilon_a$, and since $1/\Omega^C(\varepsilon_c)$ is just a constant,

$$p_a = \frac{\Omega^B(\varepsilon_c - \varepsilon_a)}{\Omega^C(\varepsilon_c)} \ . \tag{5.32}$$

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, $\varepsilon_c \gg \varepsilon_a$, which suggests a Taylor expansion. Since $\Omega^B(\varepsilon_b)$ varies vary rapidly with ε_b it is convenient to rewrite p_a as

$$p_a \propto \exp \log \Omega^B(\varepsilon_c - \varepsilon_a) \ . \tag{5.33}$$

and then Taylor expand the logarithm,

$$\log \Omega^B(\varepsilon_c - \varepsilon_a) = \log \Omega^B(\varepsilon_c) - \beta\varepsilon_a + \ldots \ , \tag{5.34}$$

where the inverse temperature $\beta = 1/kT$ of the bath has been introduced according to the standard thermodynamic definition,

$$\left.\frac{\partial \log \Omega^B}{\partial \varepsilon_b}\right|_{\varepsilon_c} \overset{\text{def}}{=} \beta \ . \tag{5.35}$$

and we conclude that the distribution that codifies the relevant information about equilibrium is

$$p_a = \frac{1}{Z} \exp(-\beta\varepsilon_a) \ , \tag{5.36}$$

which has the canonical form of eq.(4.82). (Being independent of a the factor $\Omega^B(\varepsilon_c)/\Omega^C(\varepsilon_c)$ has been absorbed into the normalization Z.)

Remark: It may be surprising that strictly speaking a system such as A *does not have a temperature.* The temperature T is not an ontic property of the system but an epistemic property that characterizes the probability distribution (5.36). Indeed, although we often revert to language to the effect that *the system is in a macrostate with temperature T,* we should note that in actual fact the system is in a particular microstate and not in a probability distribution. The latter refers to our state of knowledge and not to the ontic state of the system.

Our goal in this section was to identify the relevant variables. We are now in a position to give 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.(4.82) which coincides with (5.36).

But our analysis has also disclosed an important limitation. Eq.(5.32) 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 (a) 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.(5.29), and (b) 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.(5.29) 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. The system A 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 will be briefly addressed 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 microcanonical 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 as *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), their underlying measure, plus an identification of the relevant constraints. Indeed, the central assumption in the previous microcanonical argument — the assignment of equal probabilities to cells of equal volume *in phase space* — is justified by the information approach which replaces the ad hoc equal a priori *postulate* by an equal a priori *theorem*. Furthermore, as we shall see below, the recognition of the informational character of entropy leads to an unprecedented conceptual clarification of the foundations of statistical mechanics – including the second law. But 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 an unqualified yes.

5.4 The canonical formalism

We consider a system in thermal equilibrium [Jaynes 1957b, 1957c, 1963]. 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) information entropy,

$$S[p] = -\sum_a p_a \log p_a \quad \text{where} \quad p_a = f(z_a)\Delta z \ , \tag{5.37}$$

subject to constraints on normalization and energy $\langle\varepsilon\rangle = E$ yields, eq.(4.82),

$$p_a = \frac{1}{Z} e^{-\beta\varepsilon_a} \tag{5.38}$$

where the Lagrange multiplier β is determined from

$$-\frac{\partial \log Z}{\partial \beta} = E \quad \text{and} \quad Z(\beta, V) = \sum_a e^{-\beta\varepsilon_a} \ . \tag{5.39}$$

The maximized value of the Gibbs entropy is, eq.(4.85),

$$S_G(E,V) = kS(E,V) = k\log Z + k\beta E \ , \tag{5.40}$$

where we reintroduced the constant k. Differentiating with respect to E we obtain the analogue of eq.(4.92),

$$\left(\frac{\partial S_G}{\partial E}\right)_V = k\frac{\partial \log Z}{\partial \beta}\frac{\partial \beta}{\partial E} + k\frac{\partial \beta}{\partial E}E + k\beta \ = k\beta \ , \tag{5.41}$$

where eq.(5.39) has been used to cancel the first two terms.

The connection between the statistical formalism 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.(3.38), and identify the expected energy $\langle\varepsilon\rangle$ with the thermodynamical internal energy E: $\langle\varepsilon\rangle = E$. Next we consider a small change in the internal energy,

$$\delta E = \delta\sum_a p_a\varepsilon_a = \sum_a p_a\delta\varepsilon_a + \sum_a \varepsilon_a\delta p_a\,. \tag{5.42}$$

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,

$$\langle\delta\varepsilon\rangle = \sum_a p_a\frac{\partial\varepsilon_a}{\partial V}\delta V = \left\langle\frac{\partial\varepsilon}{\partial V}\right\rangle\delta V\ . \tag{5.43}$$

Thus, it is reasonable to identify $\langle\delta\varepsilon\rangle$ with mechanical work,

$$\langle\delta\varepsilon\rangle = \delta W = -P\delta V\ , \tag{5.44}$$

where P is the pressure,

$$P = -\left\langle\frac{\partial\varepsilon}{\partial V}\right\rangle\ . \tag{5.45}$$

Remark: This is an interesting expression in its own right. Notice that this definition of pressure makes no reference of particles colliding with the wall of a container; it is much more general. It applies to particles and to radiation both in the classical and quantum regimes. It also applies to the zero-point fluctuations of quantum fields where the resulting negative pressure is known as the Casimir effect.

Having identified the work δW, the second term in eq.(5.42) must therefore represent heat,

$$\delta Q = \delta E - \delta W = \delta\langle\varepsilon\rangle - \langle\delta\varepsilon\rangle\,. \tag{5.46}$$

The corresponding change in entropy is obtained from eq.(5.40),

$$\begin{aligned}\frac{1}{k}\delta S_G &= \delta\log Z + \delta(\beta E)\\ &= -\frac{1}{Z}\sum_a e^{-\beta\varepsilon_a}\left(\varepsilon_a\delta\beta + \beta\delta\varepsilon_a\right) + E\delta\beta + \beta\delta E\\ &= \beta(\delta E - \langle\delta\varepsilon\rangle)\ ,\end{aligned} \tag{5.47}$$

therefore,

$$\delta S_G = k\beta\delta Q \quad\text{or}\quad \delta S_G = \frac{\delta Q}{T}\ , \tag{5.48}$$

where we introduced the suggestive notation

$$k\beta = \frac{1}{T} \quad\text{or}\quad \beta = \frac{1}{kT}\ . \tag{5.49}$$

Integrating eq.(5.48) from an initial state A to a final state B gives

$$S_G(B) - S_G(A) = \int_A^B \frac{dQ}{T} \tag{5.50}$$

where every intermediate state along the path from A to B is a maximum entropy state.

We are now ready to complete the correspondence between this canonical formalism and thermodynamics. The thermodynamic entropy introduced by Clausius S_C is defined only for equilibrium states,

$$S_C(B) - S_C(A) = \int_A^B \frac{dQ}{T} \ , \tag{5.51}$$

where the integral is along a reversible path — the states along the path are equilibrium states — and where the temperature T is defined by

$$\left(\frac{\partial S_C}{\partial E}\right)_V \overset{\text{def}}{=} \frac{1}{T} \quad \text{so that} \quad \delta S_C = \frac{\delta Q}{T} \ . \tag{5.52}$$

Comparing eqs.(5.50) and (5.51) we see that the *maximized* Gibbs entropy S_G and the Clausius S_C differ only by an additive constant. Adjusting the constant so that S_G matches the Clausius entropy S_C for one equilibrium state they will match for all equilibrium states. We can therefore conclude that

> A macrostate of thermal equilibrium is described by a maximum entropy distribution. The maximized Gibbs entropy, $S_G(E,V)$, corresponds to the thermodynamic entropy S_C originally introduced by Clausius. The Lagrange multiplier β corresponds to the inverse temperature.

Thus, the framework of entropic inference provides a natural explanation for thermodynamic quantities such as temperature and entropy in terms of those theoretical concepts — Lagrange multipliers, information entropies — that must inevitably appear in all theories of inference.

Remark: It might not be a bad idea to stop for a moment and let this marvelous notion sink in: temperature, that which we associate with hot things being hot and cold things being cold is, in the end, nothing but a Lagrange multiplier. It turns out that in some common cases temperature also happens to be a measure of the mean kinetic energy per molecule. This latter conception is useful but it is too limited; it fails to capture the full significance of the concept of temperature. For example, it does not apply to relativistic particles or to photons or to black holes.

Substituting (5.44) and (5.48) into eq.(5.46), yields the *fundamental thermodynamic identity*,

$$\delta E = T\delta S - P\delta V \ , \tag{5.53}$$

where we dropped the G and C subscripts. Incidentally, this identity shows that the "natural" variables for energy are S and V, that is, $E = E(S,V)$. Similarly, writing

$$\delta S = \frac{1}{T}\delta E + \frac{P}{T}\delta V \tag{5.54}$$

confirms that $S = S(E,V)$.

Equation (5.53) 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 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,

$$F(T,V) = E - TS \; , \tag{5.55}$$

so that

$$\delta F = -S\delta T - P\delta V \; . \tag{5.56}$$

For processes at constant T we have $\delta F = \delta W$ which justifies the name 'free' energy – the amount of energy that is free to be converted to useful work when the system is not isolated but in contact with a bath at temperature T. Eq.(5.40) then leads to

$$F = -kT \log Z(T,V) \quad \text{or} \quad Z = e^{-\beta F} \; . \tag{5.57}$$

Several useful thermodynamic relations can be easily obtained from eqs.(5.53), (5.54), and (5.56). For example, the identities

$$\left(\frac{\partial F}{\partial T}\right)_V = -S \quad \text{and} \quad \left(\frac{\partial F}{\partial V}\right)_T = -P \, , \tag{5.58}$$

can be read directly from eq.(5.56).

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 with only 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. Below we offer some brief remarks on this topic.

One might, for example, account for finite bath size effects by keeping additional terms in the expansion (5.34),

$$\log \Omega^B(\varepsilon_c - \varepsilon_a) = \log \Omega^B(\varepsilon_c) - \beta\varepsilon_a - \frac{1}{2}\gamma\varepsilon_a^2 \ldots \, , \tag{5.59}$$

leading to corrections to the Boltzmann distribution,

$$p_a = \frac{1}{Z} \exp(-\beta\varepsilon_a - \frac{1}{2}\gamma\varepsilon_a^2 \ldots) \; . \tag{5.60}$$

An alternative path is to provide a more detailed model of the bath [Plastino and 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 can be quite realistic. For example, when the bath consists of N harmonic oscillators we have $\alpha = N$, and when the bath is an ideal gas of N molecules we have $\alpha = 3N/2$.

Then the number of microstates of B in a narrow energy range $\delta\varepsilon$ is

$$\Omega^B(\varepsilon) = W(\varepsilon + \delta\varepsilon) - W(\varepsilon) = \alpha C\varepsilon^{\alpha-1}\delta\varepsilon \ , \tag{5.61}$$

and the probability that A is in a particular microstate a of energy ε_a is given by eq.(5.32),

$$p_a \propto \Omega^B(\varepsilon_c - \varepsilon_a) \propto (1 - \frac{\varepsilon_a}{\varepsilon_c})^{\alpha-1} \ , \tag{5.62}$$

so that

$$p_a = \frac{1}{Z}(1 - \frac{\varepsilon_a}{\varepsilon_c})^{\alpha-1} \quad \text{with} \quad Z = \sum\nolimits_a (1 - \frac{\varepsilon_a}{\varepsilon_c})^{\alpha-1} \, . \tag{5.63}$$

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.(5.60). Indeed, using

$$\log(1+x) = x - \frac{1}{2}x^2 + \ldots \tag{5.64}$$

we expand

$$(1 - \frac{\varepsilon_a}{\varepsilon_c})^{\alpha-1} = \exp\left[(\alpha - 1)(-\frac{\varepsilon_a}{\varepsilon_c} - \frac{1}{2}(\frac{\varepsilon_a}{\varepsilon_c})^2 + \ldots)\right] \ , \tag{5.65}$$

to get eq.(5.60) with

$$\beta = \frac{\alpha - 1}{\varepsilon_c} \quad \text{and} \quad \gamma = \frac{\alpha - 1}{\varepsilon_c^2} \ . \tag{5.66}$$

We will not pursue the subject any further except to comment that distributions such as (5.63) have been proposed by C. Tsallis on the basis of a very different logic [Tsallis 1988, 2011].

Non-extensive thermodynamics

The idea proposed by Tsallis is to generalize the Boltzmann-Gibbs canonical formalism by adopting a different "non-extensive entropy",

$$T_\eta(p_1, \ldots, p_n) = \frac{1 - \sum_i p_i^\eta}{\eta - 1} \ ,$$

that depends on a parameter η [Tsallis 1988, 2011].[5] Equivalent versions of such "entropies" have been proposed as alternative measures of information by several other authors; see, for example [Renyi 1961], [Aczel 1975], and [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

$$p_i^\delta = e^{\delta \log p_i} = 1 + \delta \log p_i + \ldots \ . \tag{5.67}$$

As $\delta \to 0$ we get

$$\begin{aligned} T_{1+\delta} &= \frac{1}{\delta}(1 - \textstyle\sum_i p_i^{1+\delta}) \\ &= \frac{1}{\delta}[1 - \textstyle\sum_i p_i(1 + \delta \log p_i)] = -\textstyle\sum_i p_i \log p_i \ . \end{aligned} \tag{5.68}$$

The distribution that maximizes the Tsallis entropy subject to the usual normalization and energy constraints,

$$\textstyle\sum_i p_i = 1 \quad \text{and} \quad \textstyle\sum_i \varepsilon_i p_i = E \, ,$$

is

$$p_i = \frac{1}{Z_\eta}[1 - \lambda \varepsilon_i]^{1/(\eta - 1)} \, , \tag{5.69}$$

where Z_η is a normalization constant and the constant λ is a ratio of Lagrange multipliers. This distribution is precisely of the form (5.63) with $\lambda = 1/\varepsilon_c$ and $\eta = 1 + (\alpha - 1)^{-1}$.

Our conclusion is that Tsallis distributions make perfect sense within the canonical Gibbs-Jaynes approach to statistical mechanics. However, 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 distinguishes the canonical Boltzmann-Gibbs distributions from (5.63) or (5.69) is the relevant information on the basis of which we draw inferences and not the inference method. An added advantage is that the free and undetermined parameter η can, within the standard MaxEnt formalism advocated here, be calculated in terms of the size of the bath.

5.6 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

[5] Criticism of Tsallis' non-extensive entropy formalism is given in [La Cour and Schieve 2000] [Nauenberg 2003] [Presse et al 2013].

the few macroscopic variables over which we have some control. Most other questions are deemed 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 a particular molecule. The answer is that the expected position of the molecule is 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 very 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, $X = E = \langle\varepsilon\rangle$. The generalization to more than one variable is not difficult. The microstates z can be divided into typical and atypical microstates. The typical microstates are those contained within a region $\mathcal{R}_\delta$ defined by imposing upper and lower bounds on $f(z)$.

In this section we shall explore a few properties of the typical region. We will show that the probability of the typical region turns out to be "high", that is, $\text{Prob}[\mathcal{R}_\delta] = 1 - \delta$ where δ is a small positive number. We will also show that the thermodynamic entropy S_C and the "phase" volume W_δ of the typical region are related through Boltzmann's equation,

$$S_C \approx k \log W_\delta \ , \tag{5.70}$$

where

$$W_\delta = \text{Vol}(\mathcal{R}_\delta) = \textstyle\int_{\mathcal{R}_\delta} dz \ . \tag{5.71}$$

The surprising feature is that S_C turns out to be essentially independent of δ. The following theorems which are adaptations of the Asymptotic Equipartition Property [Shannon 1948, Shannon Weaver 1949] state this result in a mathematically precise way. (See also [Jaynes 1965] and section 4.8.)

The Asymptotic Equipartition Theorem: Let $f(z)$ be the canonical distribution and $S = S_G/k = S_C/k$ the corresponding entropy,

$$f(z) = \frac{e^{-\beta\varepsilon(z)}}{Z} \quad \text{and} \quad S = \beta E + \log Z \ . \tag{5.72}$$

If $\lim_{N\to\infty} \Delta\varepsilon/N = 0$, that is, the energy fluctuations $\Delta\varepsilon$ (Δ is the standard deviation) may increase with N but they do so less rapidly than N, then, as $N \to \infty$,

$$-\frac{1}{N}\log f(z) \longrightarrow \frac{S}{N} \quad \text{in probability,} \tag{5.73}$$

Since S/N is independent of z the theorem roughly states 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 — they are deemed "inaccessible" — or they might individually have a

high probability but are too few to contribute significantly. The term 'essentially' is tricky because $f(z)$ may differ from e^{-S} by a huge *multiplicative* factor — perhaps several billion — but $\log f(z)$ will still differ from $-S$ by an amount that is unimportant because it grows less rapidly than N.
Remark: The left hand side of (5.73) 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 is some ontological property associated to the individual microstate z rather than a property of the macrostate. But this is not so: the entropy S is a property of a whole probability distribution $f(z)$ and not of the individual zs. Any given microstate z_0 can lie within the support of several different distributions possibly describing different physical situations and having different entropies. The mere act of finding that the system is in state z_0 at time t_0 is not sufficient to allow us to figure out whether the system is best described by a macrostate of equilibrium as in (5.73) or whether it was undergoing some dynamical process that just happened to pass through z_0 at t_0.

Next we prove the theorem. Apply the Tchebyshev inequality, eq.(2.109),

$$P\left(|x-\langle x\rangle| \geq \delta\right) \leq \left(\frac{\Delta x}{\delta}\right)^2 , \tag{5.74}$$

to the variable

$$x = \frac{-1}{N}\log f(z) . \tag{5.75}$$

Its expected value is the entropy per particle,

$$\begin{aligned}\langle x\rangle &= \frac{-1}{N}\langle \log f\rangle \\ &= \frac{S}{N} = \frac{1}{N}\left(\beta E + \log Z\right) .\end{aligned} \tag{5.76}$$

To calculate the variance,

$$(\Delta x)^2 = \frac{1}{N^2}\left[\left\langle(\log f)^2\right\rangle - \langle\log f\rangle^2\right] , \tag{5.77}$$

use

$$\begin{aligned}\left\langle(\log f)^2\right\rangle &= \left\langle(\beta\varepsilon + \log Z)^2\right\rangle \\ &= \beta^2\left\langle\varepsilon^2\right\rangle + 2\beta\langle\varepsilon\rangle\log Z + (\log Z)^2 ,\end{aligned} \tag{5.78}$$

so that

$$(\Delta x)^2 = \frac{\beta^2}{N^2}\left(\left\langle\varepsilon^2\right\rangle - \langle\varepsilon\rangle^2\right) = \left(\frac{\beta\Delta\varepsilon}{N}\right)^2 . \tag{5.79}$$

Collecting these results gives

$$\text{Prob}\left[\left|-\frac{1}{N}\log f(z) - \frac{S}{N}\right| \geq \delta\right] \leq \left(\frac{\beta}{\delta}\right)^2\left(\frac{\Delta\varepsilon}{N}\right)^2 . \tag{5.80}$$

For systems such that the relative energy fluctuation $\Delta\varepsilon/N$ tends to 0 as $N \to \infty$ the limit on the right is zero,

$$\lim_{N\to\infty} \text{Prob}\left[\left|-\frac{1}{N}\log f(z) - \frac{S}{N}\right| \geq \delta\right] = 0 \ , \tag{5.81}$$

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

$$\frac{\Delta\varepsilon}{N} = \frac{\Delta\varepsilon}{E}\frac{E}{N} \propto \frac{1}{N^{1/2}} \to 0 \ . \tag{5.82}$$

Typically this happens when the spatial correlations among particles fall sufficiently fast with distance — distant particles are uncorrelated. Under these conditions both energy and entropy are extensive quantities.

The following theorem elaborates on these ideas further. To be precise let us define the typical region $\mathcal{R}_\delta$ as the set of microstates with probability $f(z)$ such that

$$e^{-S-N\delta} \leq f(z) \leq e^{-S+N\delta} \ , \tag{5.83}$$

or, using eq.(5.72),

$$\frac{1}{Z}e^{-\beta E-N\delta} \leq f(z) \leq \frac{1}{Z}e^{-\beta E+N\delta} \ . \tag{5.84}$$

This last expression shows that the typical microstates have energy within a narrow δ range

$$\varepsilon(z) \approx E \pm NkT\delta \ . \tag{5.85}$$

Remark: Even though states z with energies lower than typical 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) $\text{Vol}(\mathcal{R}_\delta) = W_\delta \leq e^{S+N\delta}$.

(3) $W_\delta \geq (1-\delta)e^{S-N\delta}$.

(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, $S = \log W$.*

For large N the entropy is a measure of the logarithm of the phase volume of typical states**,**

$$S = \log W_\delta \pm N\delta \ , \tag{5.86}$$

where $\log W_\delta = N \times O(1)$ while $\delta \ll 1$. The results above are not very sensitive to the value of δ. A broad range of values $1/N \ll \delta \ll 1$ are allowed. This means that δ can be "microscopically large" (*e.g.*, $\delta \approx 10^{-6}, 10^{-12} \gg 10^{-23}$) provided it remains "macroscopically small" (*e.g.*, $\delta \approx 10^{-6}, 10^{-12} \ll 1$). 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.(5.81) states that for fixed δ, for any given η there is an N_η such that for all $N > N_\eta$, we have

$$\text{Prob}\left[\left|-\frac{1}{N}\log f(z) - \frac{S}{N}\right| \leq \delta\right] \geq 1 - \eta\,. \tag{5.87}$$

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

$$f(z) > e^{-S+N\delta} = \frac{1}{Z}e^{-\beta E+N\delta} \quad \text{or} \quad \varepsilon(z) < E - \frac{N}{\beta}\delta \tag{5.88}$$

is negligible — states that individually are more probable than typical occupy a negligible volume. To prove **(2)** write

$$\begin{aligned} 1 &\geq \text{Prob}[\mathcal{R}_\delta] = \textstyle\int_{\mathcal{R}_\delta} dz\, f(z) \\ &\geq e^{-S-N\delta} \textstyle\int_{\mathcal{R}_\delta} dz \ = e^{-S-N\delta} W_\delta\,. \end{aligned} \tag{5.89}$$

Similarly, to prove **(3)** use **(1)**,

$$\begin{aligned} 1 - \delta &< \text{Prob}[\mathcal{R}_\delta] = \textstyle\int_{\mathcal{R}_\delta} dz\, f(z) \\ &\leq e^{-S+N\delta} \textstyle\int_{\mathcal{R}_\delta} dz = e^{-S+N\delta} W_\delta\,, \end{aligned} \tag{5.90}$$

Finally, from (2) and (3),

$$(1-\delta)e^{S-N\delta} \leq W_\delta \leq e^{S+N\delta} \ , \tag{5.91}$$

which is the same as

$$-\delta + \frac{\log(1-\delta)}{N} \leq \frac{\log W_\delta - S}{N} \leq \delta \ , \tag{5.92}$$

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

$$\langle -\log f \rangle = S = \lambda_k \left\langle X^k \right\rangle + \log Z \ , \tag{5.93}$$

and its variance is

$$(\Delta \log f)^2 = \lambda_k \lambda_m \left(\left\langle X^k X^m \right\rangle - \left\langle X^k \right\rangle \left\langle X^m \right\rangle\right) \ . \tag{5.94}$$

5.7 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.

However, since it makes no sense to assign a thermodynamic entropy to a universe that is not in thermal equilibrium we should at the very least be a bit more explicit. First recall some definitions. A process in which every intermediate state is a state of equilibrium — which is achieved if the process is very slow or quasi-static — is said to be reversible. If no heat is exchanged the process is said to be adiabatic. Then the Second Law can be stated as follows:

> In an adiabatic irreversible process that starts and ends in equilibrium the total entropy increases; if the process is adiabatic and reversible the total entropy remains constant.

The Second Law was amended into a stronger form by Gibbs (1878):

> In an adiabatic irreversible process not only does the entropy tend to increase, but it does increase to the maximum value allowed by the constraints imposed on the system.

In this and the following two sections we derive and comment on the Second Law following the argument in [Jaynes 1963, 1965]. Jaynes' derivation is deceptively simple: the mathematics is trivial.[6] But it is conceptually subtle so 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, which is defined only for equilibrium states.

Consider a system at time t in a state of equilibrium defined by certain thermodynamic variables $X(t)$. As we saw in section 5.4 the macrostate of equilibrium is described by the canonical probability distribution $f^{\text{can}}(z,t)$ obtained by maximizing the Gibbs entropy S_G subject to the constraints $X(t) = \langle x(t)\rangle$ where the quantities $x = x(z)$ are functions of the microstate such as energy, density, etc. The thermodynamic entropy S_C is then given by

$$S_C(t) = S_G^{\text{can}}(t) \ . \tag{5.95}$$

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 need not be slow. It could, for example, be the free expansion of a gas

[6] The mathematical arguments can be traced to the work of Gibbs [Gibbs 1902]. Unfortunately, Gibbs' treatment of the conceptual foundations left much to be desired and was promptly criticized in an extremely influential review by Paul and Tatyana Ehrenfest [Ehrenfest 1912]. Jaynes' decisive contribution was to place the subject on a completely different foundation based on improved conceptual understandings of probability, entropy, and information.

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. Since no heat was exchanged with the environment the process is adiabatic but not necessarily reversible. We further assume that a new equilibrium is eventually reached at some later time t'. This is a non-trivial condition to which we will briefly return below. Under these circumstances the initial canonical distribution $f^{\text{can}}(t)$, e.g. eq.(4.82) or (5.38), evolves according to Liouville's equation, eq.(5.7),

$$f^{\text{can}}(t) \overset{H(t)}{\longrightarrow} f(t') \ , \tag{5.96}$$

and, according to eq.(5.24), the corresponding Gibbs entropy remains constant,

$$S_G^{\text{can}}(t) = S_G(t') \ . \tag{5.97}$$

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 the new $f(t')$ is not necessarily of the canonical form, eq.(4.82).

From the new distribution $f(t')$ we can, however, compute the new expected values $X(t') = \langle x(t') \rangle$ that apply to the state of equilibrium at t'. Of all distributions agreeing with the same new values $X(t')$ the canonical distribution $f^{\text{can}}(t')$ is that which has maximum Gibbs entropy, $S_G^{\text{can}}(t')$. Therefore

$$f(t') \overset{\text{MaxEnt}}{\longrightarrow} f^{\text{can}}(t') \tag{5.98}$$

implies

$$S_G(t') \leq S_G^{\text{can}}(t') \ . \tag{5.99}$$

But $S_G^{\text{can}}(t')$ coincides with the thermodynamic entropy of the new equilibrium state,

$$S_G^{\text{can}}(t') = S_C(t') \ . \tag{5.100}$$

Collecting all these results, eqs.(5.95)-(5.100), we conclude that the thermodynamic entropy has increased,

$$S_C(t) \leq S_C(t') \ . \tag{5.101}$$

This is the Second Law. The equality applies when the time evolution is quasistatic so that the distribution remains canonical at all intermediate instants through the process.

To summarize, the chain of steps is

$$S_C(t) \underset{(1)}{=} S_G^{\text{can}}(t) \underset{(2)}{=} S_G(t') \underset{(3)}{\leq} S_G^{\text{can}}(t') \underset{(4)}{=} S_C(t') \ . \tag{5.102}$$

Steps (1) and (4) hinge on identifying the maximized Gibbs entropy with the thermodynamic entropy — which is justified provided we have correctly identified the relevant macrovariables X for the particular problem at hand. Step (2) follows from the constancy of the Gibbs entropy under Hamiltonian evolution

— since this is a mathematical theorem 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 — such would be an entropy increase over and above the Second Law. 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 back to the initial time t and including the initial values $X(t)$. In contrast, a description in terms of the distribution $f^{\text{can}}(t')$ contains information about the macrovariables $X(t')$ at time t' *and nothing else.* In a truly 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 evolution implied by Hamiltonian dynamics leads to distributions, such as $f(t')$, that include information beyond what is allowed in a purely thermodynamic description. It is the act of discarding such extra information that lies at the foundation of the Second Law. *The irreversibility implicit in the Second Law arises from the restriction to thermodynamic descriptions.*[7]

The fact that the Second Law refers to the thermodynamic entropies of initial and final states of equilibrium implies that many important processes lie outside its purview. One example is that of a gas that forever expands into vacuum and never reaches equilibrium. Another example, this time borrowed from cosmology, is that of an expanding universe. These are not states to which one can assign a thermodynamic entropy and therefore the Second Law does not apply. It is not that the Second Law is in any way violated; it is rather that in these matters it remains silent.

Thus, the Second Law of Thermodynamics is not a Law of "Nature." The Second Law is a law but its connection to Nature is indirect. It is a law within the very useful but also very limited class of models restricted to thermodynamic descriptions of thermal equilibrium.

It is important to emphasize what has just been proved: in an irreversible adiabatic process *from an initial to a final state of equilibrium* the *thermodynamic* entropy increases — this is the Second Law. Many questions have been left unanswered; some we will briefly address in the next two sections. Other questions we will not address: 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.

Remark: It is sometimes stated that it is the Second Law that drives a system towards equilibrium. This is not correct. The approach to equilibrium is not

[7] On the topic of descriptions see [Grad 1961, 1967, Balian Veneroni 1987, Balian 1999].

a consequence of the Second Law. If anything, it is the other way around: the existence of a final state of equilibrium is a pre-condition for the Second Law.

The extension of entropic methods of inference beyond situations of equilibrium is, of course, highly desirable. I will offer two comments on this matter. The first is that there is at least one theory of extreme non-equilibrium that is highly successful and very well known. It is called quantum mechanics — the ultimate framework for a probabilistic time-dependent non-equilibrium dynamics. The derivation of quantum mechanics as an example of an "entropic dynamics" will be tackled in Chapter 11. The second comment is that term 'non-equilibrium' is too broad and too vague to be useful. In order to make progress it is important to be very specific about which type of non-equilibrium process one is trying to describe.[8]

5.8 Interpretation of the Second Law: Reproducibility

First a summary of the previous sections: We saw that for macroscopic systems fluctuations of the variables $X(t)$ are negligible — all microstates within the typical region $\mathcal{R}(t)$ are characterized by essentially the same values of $X(t)$, eq.(5.84). We also saw that given $X(t)$ the typical region has probability one — it includes essentially all possible initial microstates 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 original 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 $\mathcal{R}'(t')$ which includes all microstates compatible with the information $X(t')$ has volume $W'(t') > W(t)$ and entropy $S_C(t') > S_C(t)$ — this is the Second law.

The volume $W(t) = e^{S_C(t)/k}$ of the typical region $\mathcal{R}(t)$ 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 control the actual microstate of the system when the $X(t)$ are the only variables we can experimentally 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 $\mathcal{R}(t)$ and therefore that in due course of time it will evolve to be within $\mathcal{R}(t')$. (See Figure 5.1) In order for the process $X(t) \to X(t')$ to be experimentally reproducible it must be that all

[8] I once heard the remark that "it is conceivable that one might be able to formulate a theory of elephants; but how could one ever come up with a theory of non-elephants?"

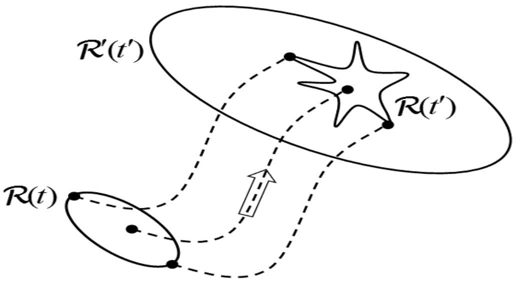

Figure 5.1: Entropy increases towards the future: The microstate of a system in equilibrium with macrovariables $X(t)$ at the initial time t lies somewhere within $\mathcal{R}(t)$. A constraint is removed and the system spontaneously evolves to a new equilibrium at $t' > t$ in the region $\mathcal{R}(t')$ characterized by values $X(t')$ and with the same volume as $\mathcal{R}(t)$. The maximum entropy region that describes equilibrium with the same values $X(t')$ irrespective of the prior history is $\mathcal{R}'(t')$. The experiment is reproducible because all states within the larger region $\mathcal{R}'(t')$ are characterized by the same $X(t')$.

the microstates in $\mathcal{R}(t)$ will also evolve to be within $\mathcal{R}'(t')$ which means that $W(t) = W(t') \leq W'(t')$. Conversely, if it happened that $W(t) > W'(t')$ we would sometimes observe that an initial microstate within $\mathcal{R}(t)$ would evolve into a final microstate lying outside $\mathcal{R}'(t')$, that is, sometimes we would observe that $X(t)$ would not evolve to $X(t')$. Such an experiment would definitely not be 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 physical constraints Y that significantly restrict the microstates compatible with the initial macrostate and which therefore provide an even better description? The answer is that to the extent that we are *only interested in the*

X variables, it is unlikely that the inclusion of additional Y variables in the description will lead to improved predictions. The reason is that since the process $X(t) \to X(t')$ is already reproducible when no particular care has been taken to control the values of Y it is unlikely that the additional information provided by Y would be relevant. Thus 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 for our purposes or not.*

5.9 On reversibility, irreversibility, and the arrow of time

A considerable source of confusion on the question of reversibility originates in the fact 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 former refers to equilibrium macrostates, the latter refers to microstates. 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 its very 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 the system might have evolved from a non-equilibrium situation in the relatively recent past. Thus, 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 still happen. 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. In principle the situation is 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 — a non-equilibrium state — might have been observed in the very recent past. This can seem paradoxical because the formalisms of mechanics and of statistical mechanics do not introduce any time asymmetry. The solution to puzzles of this kind hinges on realizing that if the system was in a highly improbable state in the recent past, then it is most likely that the state did not arise 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 — this is not uncommon; we do it all the time. Thus, the time asymmetry is not introduced by the laws of mechanics. It is introduced through the asymmetry between our information about external interventions in the past versus our information about spontaneous processes in the future.

We can pursue this matter further. It is not unusual to hear that the arrow of time is defined by the Second Law. The claim is that since the laws of dynamics are invariant under time-reversal,[9] in order to distinguish the past from the future we need a criterion from outside mechanics and the Second Law is proposed as a potential candidate. One objection to this proposal is that the thermodynamic entropy applies only to equilibrium states which is too limited as it excludes most irreversible non-equilibrium processes of interest. Another is that if the future is defined as the direction in which entropy increases, then it is impossible for entropy not to increase with time. In other words, either the Second Law is a tautology or one needs to seek elsewhere for an arrow of time.

Which raises the question: does the Jaynes' derivation lead to a tautological Second Law, or was an arrow of time effectively introduced somewhere else? The answer is that a non-thermodynamic arrow was indeed introduced so that the Second Law is not tautological.

One might wonder whether the arrow was introduced by the mere action of asking a question that was itself already asymmetrical. Well, yes, the system starts in an *initial* equilibrium state, *a constraint is removed*, and we asked to which *final* equilibrium does the system spontaneously evolve. The conclusion was that entropy increased into the future. As said earlier, the asymmetry consists of an external intervention in the past — the removal of a constraint — and spontaneous evolution to the future. A more symmetrical question would reflect spontaneous evolution both into the future and from the past. Suppose we ask the reverse question: given the final equilibrium state at time t', which

[9] The violations of time reversal symmetry that are found in particle physics are not relevant to the issues discussed here.

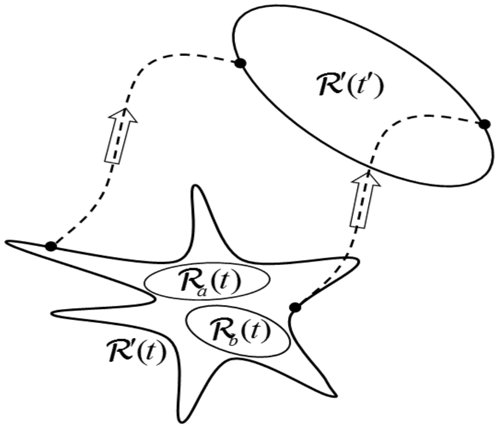

Figure 5.2: The reproducibility arrow of time leads to the Second law: we can guarantee that the system will reproducibly evolve to $\mathcal{R}'(t')$ by controlling the initial microstate to be in region $\mathcal{R}_a(t)$.

initial equilibrium state at time $t < t'$ did it come from? The answer is that once an equilibrium has been reached at time t' then, by the very definition of equilibrium, the spontaneous evolution into the future $t'' > t'$ will maintain the same equilibrium state. And vice versa: to the extent that the system has evolved spontaneously — that is, *to the extent that there are no external interventions* — the time reversibility of the Hamiltonian dynamics leads us to that if the system is in equilibrium at t'', then it must have been in equilibrium at any all other previous time t'. In other words, if the equilibrium at time t' is defined by variables $X(t')$, then the spontaneous evolution both into the future $t'' > t'$ and from the past $t < t'$, lead to the same equilibrium state, $X(t) = X(t') = X(t'')$.

But, of course, our interest lies precisely in the effect of external interventions. A *reproducible* experiment that starts in equilibrium and ends in the equilibrium state of region $\mathcal{R}'(t')$ defined by $X(t')$ is shown in Figure 5.2. We can guarantee that the initial microstate will end somewhere in $\mathcal{R}'(t')$ by controlling the initial equilibrium macrostate to have values $X(t)$ that define a region such as $\mathcal{R}_a(t)$. Then we have an external intervention: a constraint is removed and the system is allowed to evolve spontaneously. The initial region $\mathcal{R}_a(t)$ will necessarily have a volume very very much smaller than $\mathcal{R}'(t)$. Indeed, the region $\mathcal{R}_a(t)$ would be highly atypical within $\mathcal{R}'(t)$. The entropy of the initial region $\mathcal{R}_a(t)$ is lower than that of the final region $\mathcal{R}'(t)$ which leads to the Second Law once again.

Figure 5.2 also shows that there are many other initial equilibrium macrostates such as $\mathcal{R}_b(t)$ defined by values $X_b(t)$ that lead to the same final equilibrium macrostate. For example, if the system is a gas in a box of given volume and the final state is equilibrium, the gas might have initially been in equilibrium confined by a partition to the left half of the box, or the upper half, or the right third, or any of many other such constrained states.

Finally, we note that introducing the notion of reproducibility is something that goes beyond the laws of mechanics. Reproducibility refers to our capability to control the initial microstate by deliberately manipulating the values $X_a(t)$ in order to reproduce the later values $X(t')$. The relation is that of a cause $X_a(t)$ leading to an effect $X(t')$. To the extent that causes are supposed to precede their effects we conclude that the *reproducibility arrow of time* is the *causal arrow of time.*

Our goal has been to derive the Second Law which requires an arrow of time but, at this point, the origin of the latter remains unexplained. In chapter 11 we will revisit this problem within the context of a dynamics conceived as an application of entropic inference. There we will find that the time associated to such an "entropic dynamics" is intrinsically endowed with the directionality required for the Second Law.

Overview

It may be useful to collect the main arguments of the previous three sections in a more condensed form along with a short preview of things to come later in chapter 11. To understand what is meant by the Second Law — roughly that entropy increases as time increases — one must specify what entropy we are talking about, and we must specify an arrow of time so that it is clear what we mean by 'time increases'.

The entropy in the Second Law of Thermodynamics is the thermodynamic entropy of Clausius, eq.(5.51), which is only defined for equilibrium states. One can invent all sorts of other entropies, which might increase or not. One might, for example, define an entropy associated to a microstate. The increase of such an entropy could be due to some form of coarse graining, or could be induced by unknown external perturbations. Or we could have the entropy of a probability distribution that increases in a process of diffusion. Or the entropy of the distribution $|\Psi|^2$ as it might arise in quantum mechanics, which increases just as often as it decreases. Or any of many other possibilities. But none of these refer to the Second Law, nor do they violate it.

Then there is the question of the arrow of time: either it is defined by the Second Law or it is not. If it is, then the Second law is a tautology. While this is a logical possibility one can offer a pragmatic objection: an arrow of time linked to the entropy of equilibrium states is too limited to be useful — thermal equilibrium is too rare, too local, too accidental a phenomenon in our universe. It is more fruitful to pursue the consequences of an arrow of time that originates through some other mechanism.

Entropic dynamics (ED) offers a plausible mechanism in the spirit of inference and information (see chapter 11). In the ED framework, time turns out to be intrinsically endowed with directionality; ultimately this is what sets the direction of causality. The arrow of such an "entropic time" is linked to an entropy but not to thermodynamics; it is linked to the dynamics of probabilities. The causal arrow of time is the arrow of entropic time.

Thus, the validity of the Second Law rests on three elements: (1) the thermodynamic entropy given by the maximized Gibbs entropy; (2) the existence of an arrow of time; and (3) the existence of a time-reversible dynamical law that involves no loss of information. The inference/information approach to physics contributes to explain all three of these elements.

5.10 Avoiding pitfalls – II: is this a 2$^{\text{nd}}$ law?

Canonical distributions have many interesting properties that can be extremely useful but can also be extremely misleading. Here is an example.

We adopt the notation of Section 5.1 and consider the canonical distribution

$$f_s(z) = \frac{1}{Z_s} e^{-\lambda_k x^k(z)} \ . \tag{5.103}$$

The λ_k are Lagrange multipliers chosen to enforce the expected value constraints

$$\langle x^k \rangle = \int dz\, f(z) x^k(z) = X_s^k \tag{5.104}$$

where $x^k(z)$ are some functions of the microstate z^α ($\alpha = 1 \ldots 6N$) and the partition function is

$$Z_s = \int dz\, e^{-\lambda_k x^k(z)} = e^{-\Phi_s} \tag{5.105}$$

where the potential Φ_s is a Mathieu function — a Legendre transform of the entropy that is somewhat analogous to a free energy. Indeed, the entropy of f_s is

$$S[f_s] = \lambda_k X_s^k - \Phi_s \ . \tag{5.106}$$

Here is a theorem:

Theorem: The distribution $f_s(z)$, eq.(5.103), is a stationary solution of the Fokker-Planck equation

$$\partial_t f = \partial_\alpha [f \partial^\alpha (\lambda_k x^k)] + \partial_\alpha \partial^\alpha f \ , \tag{5.107}$$

where $\partial_t = \partial/\partial t$, $\partial_\alpha = \partial/\partial z^\alpha$, $\partial^\alpha = \delta^{\alpha\beta}\partial_\beta$.

Proof: Rewrite eq.(5.107) as a continuity equation,

$$\partial_t f = -\partial_\alpha (f v^\alpha) \ , \tag{5.108}$$

according to which the probability flows with a velocity v^α given by

$$v^\alpha(z) = -\partial^\alpha [\lambda_k x^k(z) + \log f(z)] \ . \tag{5.109}$$

From (5.103) and (5.105) for $f = f_s$ we have

$$\lambda_k x^k(z) + \log f_s(z) = \Phi_s = \text{const} \tag{5.110}$$

and the corresponding current velocity, $v^\alpha = -\partial^\alpha \Phi_s$, vanishes.

Definition: In analogy to (5.106) for any arbitrary distribution $f(z)$ we can define a potential

$$\Phi[f] = \lambda_k X^k[f] - S[f] \ , \tag{5.111}$$

where

$$X^k[f] = \int dz\, f(z) x^k(z) \quad \text{and} \quad S[f] = -\int dz\, f(z) \log f(z) \ . \tag{5.112}$$

When the X^k are conserved quantities the potential Φ can be interpreted as follows. Assume that the system is in contact with reservoirs described by the Lagrange multipliers λ_k. From (5.111) a small change $\delta\Phi$ is given by

$$\delta\Phi = \lambda_k \delta X^k - \delta S \tag{5.113}$$

where the δX^k represent the amount of X^k *withdrawn* from the reservoirs and supplied to the system. The corresponding change in the thermodynamic entropy of the reservoirs is given by eq.(4.94),

$$\delta S_{\text{res}} = -\lambda_k \delta X^k \ , \tag{5.114}$$

so that

$$\delta\Phi = -\delta \left(S_{\text{res}} + S \right) \ . \tag{5.115}$$

Therefore, $-\delta\Phi$ represents the increase of the entropy of the combined system plus reservoirs.

Here is another theorem:

Theorem: If $f_t(z)$ is a solution of the Fokker-Planck equation (5.107), then the potential $\Phi[f]$ is a decreasing function of t.

Proof: From (5.111),

$$\frac{d\Phi}{dt} = \int dz\ \left(\lambda_k x^k + \log f + 1 \right) \partial_t f \ . \tag{5.116}$$

Next, use (5.108), (5.109), and integrate by parts,

$$\frac{d\Phi}{dt} = \int dz\, f \partial_\alpha \left(\lambda_k x^k + \log f \right) v^\alpha = -\int dz\, f\, v_\alpha v^\alpha \leq 0 \ , \tag{5.117}$$

which concludes the proof.

Discussion We have just shown that starting from any arbitrary initial distribution f_0 at time t_0, the solutions f_t of (5.107) evolve irreversibly towards a final state of equilibrium f_s. Furthermore, there is a potential Φ that monotonically decreases towards its minimum value Φ_s and that can (sometimes) be interpreted as the monotonic increase of the total entropy of system plus reservoirs.

What are we to make of all this? We appear to have a derivation of the Second Law with the added advantage of a dynamical equation that describes not only the approach to equilibrium but also the evolution of distributions arbitrarily far from equilibrium. If anything, this is too good to be true, where is the mistake?

The mistake is not to be found in the mathematics which is rather straightforward. The theorems are indeed true. The problem lies in the physics. To see this suppose we deal with a single system in thermal contact with a heat bath at temperature T. What is highly suspicious is that the process of thermalization described by eq.(5.107) appears to be of universal validity. It depends only on the bath temperature and is independent of all sorts of details about the thermal contact. This is blatantly wrong physics: surely the thermal conductivity of the walls that separate the system from the bath — whether the conductivity is high or low, whether it is uniform or not — must be relevant.

Here is another problem with the physics: the parameter that describes the evolution of f_t has been called t and this might mislead us to think that t has something to do with time. But this need not be so. In order for t to deserve being called time — and for the Fokker-Planck equation to qualify as a true dynamical equation, even if only an effective or phenomenological one — one must establish how the evolution parameter is related to properly calibrated clocks. As it is, eq.(5.107) is not a *dynamical* equation. It is just a cleverly constructed equation for those curves in the space of distributions that have the peculiar property of admitting canonical distributions as stationary states.

5.11 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?

Should add a more extended introduction to the Gibbs paradox.

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, 1967][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 (5.38)

$$p_a=\frac{e^{-\beta\varepsilon_a}}{Z_H}\,, \qquad (5.118)$$

where the partition function Z_H and the Lagrange multiplier β are determined from eqs.(5.39),

$$Z_H=\sum_i e^{-\beta\varepsilon_i}\quad\text{and}\quad\frac{\partial\log Z_H}{\partial\beta}=-E\;. \qquad (5.119)$$

The corresponding entropy, eq.(5.40) is (setting $k=1$)

$$S_H=\beta E+\log Z_H\,, \qquad (5.120)$$

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,

$$P_g = \frac{e^{-\beta E_g}}{Z_L}\,, \tag{5.121}$$

where

$$Z_L = \sum_g e^{-\beta E_g} \quad \text{and} \quad \frac{\partial \log Z_L}{\partial \beta} = -\bar{E}\,. \tag{5.122}$$

The corresponding entropy, eq.(5.40) is (setting $k = 1$)

$$S_L = \beta \bar{E} + \log Z_L\,, \tag{5.123}$$

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

$$Z_H = \sum_i e^{-\beta E_i} = N! \sum_g e^{-\beta E_g} = N! Z_L \quad \text{or} \quad Z_L = \frac{Z_H}{N!}\,. \tag{5.124}$$

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

$$S_L = S_H - \log N!\,. \tag{5.125}$$

Incidentally, note that

$$S_H = -\textstyle\sum_a p_a \log p_a = -\sum_g P_g \log P_g/N!\ . \tag{5.126}$$

Equations (5.124) and (5.125) 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.(5.124) and (5.125) 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

$$\mu = -kT\left(\frac{\partial S_T}{\partial N}\right)_{E,V,\ldots} \tag{5.127}$$

The two identifications $S_H = S_T$ or $S_L = S_T$, lead to two different chemical potentials, related by

$$\mu_L = \mu_H - NkT\,. \tag{5.128}$$

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

$$Z_L = \frac{Z_H}{N!} \quad \text{with} \quad Z_H = \sum_i e^{-\beta E_i} \,, \tag{5.129}$$

and the connection to the thermodynamics is established by postulating

$$S_T = S_L = S_H - \log N! \,. \tag{5.130}$$

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

$$\hat{Z}_L = \frac{\hat{Z}_H}{N_a!N_b!} \quad \text{with} \quad \hat{Z}_H = \sum_{\hat{\imath}} e^{-\beta E_i} \, , \tag{5.131}$$

and the connection to the new phenomenology is established by postulating

$$\hat{S}_T = \hat{S}_L = \hat{S}_H - \log N_a!N_b! \, . \tag{5.132}$$

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: one description is the macrostate, the other is 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 those fortunate circumstances when new information becomes available. *The theory of probability, if it is to be useful at all, demands a method 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 consists of data and 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 ought to be revised but only to the minimal 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 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]. One could introduce other entropies justified by different choices of axioms (see, for example, [Renyi 1961] and [Tsallis 1988]). Which one should we adopt? If different systems are to handled using different Renyi entropies, how do we handle composite systems?

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 diversity of the messages that could potentially be generated by a source (section 4.8). 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 [Shore and Johnson 1980; Skilling 1988-1990; Csiszar 1991, 2008; Caticha 2003, 2014a]. 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

[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.

basis of 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.[2]

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 proceed to 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 question whether degrees of belief can be interpreted as probabilities. Similarly, here 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 research failed to find an 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 will 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]) before the use of relative entropy as a tool for inference had been properly understood. It is not, therefore, surprising that 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

[2] The presentation below is based on work presented in a sequence of papers [Caticha 2003, Caticha Giffin 2006, Caticha 2007, 2014a, Vanslette 2017] and in earlier versions of these lectures [Caticha 2008, 2012c].

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 2.10.2). Another is that it has implications for physics: it provides an important missing piece for the old puzzles of quantum mechanics concerning the so-called collapse of the wave function and the measurement problem(see Chapter 11).

There is yet 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 topic, which completes the formulation of the ME method, will be addressed in chapter 8.

6.1 What is information?

The term 'information' is used with a wide variety of different meanings [Cover Thomas 1991; Landauer 1991; Jaynes 2003; Caticha 2007, 2014a; Golan 2008, 2018; Floridi 2011]. 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. 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 a message affects the beliefs of an ideally rational agent.

There is also an algorithmic notion of information, which captures the notion of complexity [Cover and Thomas 1991] and originates in the work of [Solomonov 1964], [Kolmogorov 1965] and [Chaitin 1975], and the related Minimum Description Length principle of [Rissanen 1978, 1986]. The algorithmic approach 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 their potential relevance to our subject, these algorithmic approaches will not be pursued here.

It is not unusual to hear that systems "carry" or "contain" information and that "information is physical".[3] This mode of expression can perhaps be traced to the origins of information theory in Shannon's theory of communication. We

[3] The general context is the thermodynamics of computation. See [Landauer 1991, Bennett 1982, 2003] and references therein. For a critical appraisal see [Norton 2011, 2013].

say that we have received information when among the vast variety of messages that could 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, however, exhaust the meaning of the word 'information'.

Here we will follow a different path. We seek an epistemic notion of information that is somewhat closer to the everyday colloquial use of the term — roughly, information is what I get when my question has been answered. Indeed, a fully Bayesian information theory requires an explicit account of the relation between information and the beliefs of ideally rational agents. Furthermore, 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 assessment of what beliefs we ought to hold is driven by the conviction that not all beliefs, not all probability assignments, are equally good. The concern with 'good' and 'better' bears on the issue of whether probabilities are subjective, objective, or somewhere in between. We argued earlier (in Chapter 1) that what makes one probability assignment better than another is 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". Thus, objectivity is desirable; objectivity is the goal. Probabilities are useful to the extent that they incorporate some degree of epistemic objectivity.[4] 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 2.10.3 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

[4] We recall from Section 1.1.3 that probabilities are ontologically subjective but epistemically they can span the range from being fully subjective to fully objective.

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 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) becomes "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 answer is pragmatic:

> *Information is what information does.*

Information is defined by its effects: (a) it induces us to update from prior beliefs to posterior beliefs, and (b) it restricts our options as to what we are honestly and rationally allowed to believe. This, I propose, is the defining characteristic of information.

> *Information is that which induces a change from one state of rational belief to another state that is more appropriately constrained.*

One significant 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.

Another aspect is that the notion that information is directly related to changing our minds does not involve any talk about *amounts* of information. Nevertheless it allows precise quantitative calculations. Indeed, constraints on the acceptable posteriors are precisely the kind of information the method of maximum entropy is designed to handle.

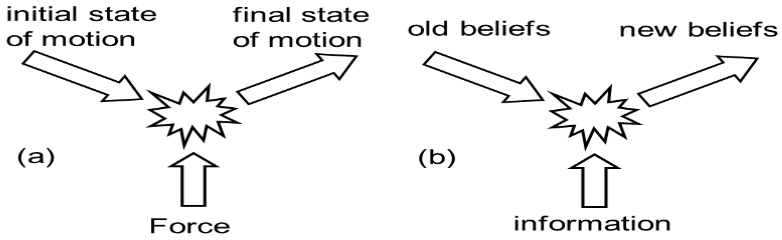

Figure 6.1: (a) In mechanics force is defined as that which affects motion. (2) Inference is dynamics too: information is defined as that which affects rational beliefs.

> *The mathematical representation of information is in the form of constraints on posterior probability distributions. The constraints are the information.*

Constraints can take a wide variety of forms including, in addition to the examples mentioned above, anything capable of affecting beliefs. For example, in Bayesian inference the likelihood function constitutes information because it contributes to constrain our posterior 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 and 11.)

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 or impulse. Similarly, in Bayesian inference a state of belief is described in terms of probabilities — a "degree" of belief — and the change from one state to another is due to information (see Fig.6.1). 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. In Chapter 11 we will reverse the logic and derive dynamical laws of physics as examples of entropic updating of probabilities — an entropic 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 would an ideally rational agent 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 — to be useful — we require 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 valuable 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 (PMU): *Beliefs should be updated only to the minimal extent required by the new information.*

This version of the principle generalizes the earlier version presented in section 2.10.2 which was restricted to information in the form of data.

The special case of updating in the absence of new information deserves a comment. The PMU states that when there is no new information ideally rational agents should not change their minds.[5] 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 agents 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 rational agents to change their minds for no

[5] Our concern here is with ideally rational agents who have fully processed all information acquired in the past. Our subject is not the psychology of actual humans who often change their minds by processes that are not fully conscious.

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 be compromised. 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 be merely 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 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: *Whenever two systems are a priori believed to be independent and we receive information about one it should not matter if the other is included in the analysis or not.* This amounts to requiring that independence be preserved unless information about correlations is explicitly introduced.

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 set of propositions $\{x\}$ about which we are uncertain. The proposition x can be discrete or continuous, in one or in several dimensions. It could, for example, represent the microstate 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],[6] 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 such ranking 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. Transitive rankings are implemented by assigning to each p 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 strategy of ranking distributions cannot be overestimated: it implies that the updating method will take the form of a 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 or some origin of potential energy but there is no need to constantly refer to a 'relative energy' — it is just not done.

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. Candidate functionals that fail to satisfy the criteria are discarded — hence the qualifier 'eliminative'. As we shall see the criteria adopted below are sufficiently constraining that there is a single entropy functional $S[p,q]$ that survives the process of elimination.

[6] [Skilling 1988] deals with the more general problem of ranking positive additive distributions which includes the intensity of images as well as probability distributions.

This approach has a number of virtues. First, to the extent that the design criteria are universally desirable, 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 candidates demonstrably fail to perform as desired.*

6.2.3 Specific design criteria

Consider a lattice of propositions generated by a set $\mathcal{X}$ of atomic propositions (mutually exclusive and exhaustive propositions) labeled by a discrete index $i = 1, 2 \ldots n$. The extension to infinite sets and to continuous labels will turn out to be straightforward. The index i might, for example, label the microstates of a physical system but, since the argument below is supposed to be of general validity, we shall not assume that the labels themselves carry any particular significance. We can always permute labels and this should have no effect on the updating of probabilities.

The structure of the lattice — its members are related to each other by disjunctions (OR) and conjunctions (AND) — is reflected in the consistency of the web of beliefs which is implemented through the sum and product rules. We adopt two design criteria one of which refers to propositions that are mutually exclusive and the other to propositions that are independent. These represent two extreme situations. At one end we have highly correlated propositions (if one proposition is true the other is false and vice versa); at the other end we have totally uncorrelated propositions (the truth or falsity of one proposition has no effect on the truth or falsity of the other). One relation is described by a simplified sum rule, $p(i \vee j) = p(i) + p(j)$, and the other by a simplified product rule, $p(i \wedge j) = p(i)p(j)$.[7]

Two design criteria and their consequences for the functional form of the entropy are given below. Detailed proofs are deferred to the next section.

Mutually exclusive subdomains

DC1 *Probabilities that are conditioned on one subdomain are not affected by information about other non-overlapping subdomains.*

Consider a subdomain $\mathcal{D} \subset \mathcal{X}$ composed of atomic propositions $i \in \mathcal{D}$ and suppose the information to be processed refers to some other subdomain $\mathcal{D}' \subset \mathcal{X}$ that does not overlap with $\mathcal{D}$, $\mathcal{D} \cap \mathcal{D}' = \emptyset$. 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 $q(i|\mathcal{D})$, the prior probability of i conditioned on $i \in \mathcal{D}$, is not updated. The selected

[7] For an alternative approach to the foundations of inference that exploits the various symmetries of the lattice of propositions see [Knuth 2005, 2006; Knuth Skilling 2012].

conditional posterior is

$$P(i|\mathcal{D})=q(i|\mathcal{D})\ . \tag{6.1}$$

(We adopt the following notation: priors are denoted by q, candidate posteriors by lower case p, and the selected posterior by upper case P. We shall write either $p(i)$ or p_i.)

We emphasize: the point is not that we make the unwarranted assumption that keeping $q(i|\mathcal{D})$ 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 i contribute additively to the entropy,

$$S(p,q)=\sum_i F\left(p_i,q_i\right)\ , \tag{6.2}$$

where F is some unknown function of two arguments. The proof is given in section 6.3.

Comment 1: It is essential that DC1 refers to *conditional* probabilities: local information about a domain $\mathcal{D}'$ can have a non-local effect on the total probability of another domain $\mathcal{D}$. An example may help to see why: Consider a loaded die with faces $i=1\ldots 6$. A priori we have no reason to favor any face, therefore $q(i)=1/6$. Then we are told that the die is loaded in favor of 2. The criterion DC1 tells nothing about how to update the $P(i)$s. If the die were *very* loaded in favor of 2, say, $P(2)=0.9$ then it must be that $P(i)<1/6$ for $i\neq 2$ and therefore all $P(i)$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 the conditional probabilities remain unchanged $P(i|\mathcal{D})=q(i|\mathcal{D})=1/4$; it says nothing about whether $P(i)$ for $i\in\mathcal{D}$ is less or more than $1/6$.[8]

Comment 2: An important special case is the "update" from a prior $q(i)$ to a posterior $P(i)$ in a situation in which no new information is available. The locality criterion DC1 applied to a situation where the subdomain $\mathcal{D}$ covers the whole space of is, $\mathcal{D}=\mathcal{X}$, requires that *in the absence of any new information the prior conditional probabilities are not to be updated:* $P(i|\mathcal{X})=q(i|\mathcal{X})$ *or* $P(i)=q(i)$.

Comment 3: The locality criterion DC1 includes Bayesian conditionalization as a special case. Indeed, if the information is given through the constraint $p(\tilde{\mathcal{D}})=0$ where $\tilde{\mathcal{D}}$ is the complement of $\mathcal{D}$ then $P(i|\mathcal{D})=q(i|\mathcal{D})$, which is referred to as Bayesian conditionalization. More explicitly, if θ is the variable to be inferred on the basis of information about a likelihood function $q(i|\theta)$ and observed data i', then the update from the prior q to the posterior P,

$$q(i,\theta)=q(i)q(\theta|i)\rightarrow P(i,\theta)=P(i)P(\theta|i) \tag{6.3}$$

[8] For $i\in\mathcal{D}$, if $p_2<2/9$ then $P_i>1/6$; if $p_2>2/9$ then $P_i<1/6$.

consists of updating $q(i) \to P(i) = \delta_{ii'}$ to agree with the new information and invoking the PMU so that $P(\theta|i') = q(\theta|i')$ remains unchanged. Therefore,

$$P(i,\theta) = \delta_{ii'} q(\theta|i') \quad \text{and} \quad P(\theta) = q(\theta|i') \ , \tag{6.4}$$

which is Bayes' rule (see sections 2.10.2 and 6.6 below). Thus, *entropic inference is designed to include Bayesian inference as a special case.* Note however that imposing DC1 is not identical to imposing Bayesian conditionalization: DC1 is not restricted to information in the form of absolute certainties such as $p(\mathcal{D}) = 1$.

Comment 4: If the label i is turned into a continuous variable x 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 (common in *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. The inference process should not be expected to discover and replicate information with which it was not supplied.

Subsystem independence

DC2 *When two systems are a priori believed to be independent and we receive independent information about one then it should not matter if the other is included in the analysis or not.*

Consider a system of propositions labelled by a composite index, $i = (i_1, i_2) \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$. For example, $\{i_1\} = \mathcal{X}_1$ and $\{i_2\} = \mathcal{X}_2$ might describe the microstates of two separate physical systems. Assume that all prior evidence led us to believe the two subsystems are independent, that is, any two propositions $i_1 \in \mathcal{X}_1$ and $i_2 \in \mathcal{X}_2$ are believed to be independent. This belief is reflected in the prior distribution: if the individual subsystem priors $q_1(i_1)$ and $q_2(i_2)$, then the prior for the whole system is $q_1(i_1)q_2(i_2)$. Next suppose that new information is acquired such that $q_1(i_1)$ would by itself be updated to $P_1(i_1)$ and that $q_2(i_2)$ would itself be updated to $P_2(i_2)$. DC2 requires that $S[p,q]$ be such that the joint prior $q_1(i_1)q_2(i_2)$ updates to the product $P_1(i_1)P_2(i_2)$ so that inferences about one subsystem do not affect inferences about the other.

The consequence of DC2 is to fully determine the unknown function F in (6.2) so that probability distributions $p(i)$ should be ranked relative to the prior $q(i)$ according to the relative entropy,

$$S[p,q] = -\sum_i p(i) \log \frac{p(i)}{q(i)} . \tag{6.5}$$

Comment 1: 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. Induction involves risk; the systems might in actual fact be correlated through some unknown interaction potential. The point is rather that if the joint prior reflected independence and the new evidence is silent on the matter of correlations, then the evidence we actually have — namely, 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 2: We also emphasize that *DC2 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. DC2 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.

Comment 3: The generalization to continuous variables $x \in \mathcal{X}$ is approached as a Riemann limit from the discrete case (see Section 4.6). A continuous probability density $p(x)$ or $q(x)$ can be approximated by the discrete distributions. Divide the region of interest $\mathcal{X}$ into a large number N of small cells. The probabilities of each cell are

$$p_i = p(x_i)\Delta x_i \quad \text{and} \quad q_i = q(x_i)\Delta x_i \ , \tag{6.6}$$

where Δx_i is an appropriately small interval. The discrete entropy of p_i relative to q_i is

$$S_N = -\sum_{i=1}^{N} \Delta x_i \, p(x_i) \, \log \left[\frac{p(x_i)\Delta x_i}{q(x_i)\Delta x_i} \right] \ , \tag{6.7}$$

and in the limit as $N \to \infty$ and $\Delta x_i \to 0$ we get the Riemann integral

$$S[p,q] = -\int dx \, p(x) \, \log \left[\frac{p(x)}{q(x)} \right] \ , \tag{6.8}$$

(To simplify the notation we include multi-dimensional integrals by writing $d^n x = dx$.)

It is easy to check that the ranking of distributions induced by $S[p,q]$ is invariant under coordinate transformations.[9] More explicitly, 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,

$$dx = \gamma(x')dx' \quad \text{where} \quad \gamma(x') = \left| \frac{\partial x}{\partial x'} \right| . \tag{6.9}$$

[9] The insight that coordinate invariance could be derived as a consequence of the requirement of subsystem independence first appeared in [Vanslette 2017].

Since $p(x)$ is a density; the transformed function $p'(x')$ is such that $p(x)dx = p'(x')dx'$ is invariant. Therefore

$$p(x) = \frac{p'(x')}{\gamma(x')} \quad \text{and} \quad q(x) = \frac{q'(x')}{\gamma(x')} \,, \tag{6.10}$$

and (6.8) gives

$$S[p', q'] = S[p, q] \ . \tag{6.11}$$

This shows that the two rankings, the one according to $S[p, q]$ and the other according to $S[p', q']$ coincide.

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. The preferred posterior P within the family $\{p\}$ is that which maximizes the relative entropy,*

$$S[p, q] = -\sum_i p_i \log \frac{p_i}{q_i} \quad \textit{or} \quad S[p, q] = -\int dx\, p(x) \, \log \left[\frac{p(x)}{q(x)}\right] \ , \tag{6.12}$$

subject to the constraints $\mathcal{C}$.

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. *No interpretation for $S[p, q]$ is given and none is needed.* 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 that are close to the entropy maximum ought to be ruled out or should 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, the one-parameter family of entropies proposed in [Renyi 1961, Aczel Daróczy 1975, Amari 1985, Tsallis 1988], see Section 6.5.3 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 for the purpose of updating in that they do not perform the functions stipulated by the design criteria DC1 and DC2.

6.3 The proofs

In this section we establish the consequences of the two criteria leading to the final result eq.(6.12). 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 for mutually exclusive subdomains

Here we prove that criterion DC1 leads to the expression eq.(6.2) for $S[p,q]$. Consider the case of a discrete variable, p_i with $i = 1 \ldots n$, so that $S[p,q] = S(p_1 \ldots p_n, q_1 \ldots q_n)$. 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 $\tilde{\mathcal{D}}$,

$$\sum_{j\in\tilde{\mathcal{D}}} a_j p_j = A \ . \tag{6.13}$$

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,

$$\sum_{i\in\mathcal{D}} p_i = P_{\mathcal{D}} \quad \text{and} \quad \sum_{j\in\tilde{\mathcal{D}}} p_j = P_{\tilde{\mathcal{D}}} \ , \tag{6.14}$$

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,

$$\begin{aligned} 0 = \Bigg[& \delta S - \lambda\left(\sum_{i\in\mathcal{D}} p_i - P_{\mathcal{D}}\right) + \\ & - \tilde{\lambda}\left(\sum_{j\in\tilde{\mathcal{D}}} p_i - P_{\tilde{\mathcal{D}}}\right) + \mu\left(\sum_{j\in\tilde{\mathcal{D}}} a_j p_j - A\right)\Bigg] \ , \end{aligned}$$

leading to

$$\frac{\partial S}{\partial p_i} = \lambda \quad \text{for} \quad i \in \mathcal{D} \ , \tag{6.15}$$

$$\frac{\partial S}{\partial p_j} = \tilde{\lambda} + \mu a_j \quad \text{for} \quad j \in \tilde{\mathcal{D}} \ . \tag{6.16}$$

Eqs.(6.13-6.16) 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

$$\frac{\partial S}{\partial p_i}=f_i(p_1\ldots p_n,q_1\ldots q_n) \tag{6.17}$$

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 conditioned on $\mathcal{D}$ can be shielded from arbitrary changes in the constraints pertaining to $\tilde{\mathcal{D}}$ is that for any $i\in\mathcal{D}$ the function f_i depends only on p_js with $j\in\mathcal{D}$. 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 including the partition into atomic propositions. Therefore f_i can depend only on p_i,

$$\frac{\partial S}{\partial p_i}=f_i(p_i,q_1\ldots q_n)\ . \tag{6.18}$$

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 (6.18) becomes

$$\frac{\partial S}{\partial p_i}=f_i(p_i,q_1\ldots q_j+\delta q_j\ldots q_n) \tag{6.19}$$

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 possible partitions into $\mathcal{D}$ and $\tilde{\mathcal{D}}$, and therefore,

$$\frac{\partial S}{\partial p_i}=f_i(p_i,q_i)\quad\text{for all}\quad i\in\mathcal{X}\ . \tag{6.20}$$

The choice of the functions $f_i(p_i,q_i)$ can be restricted further. If we were to maximize $S[p,q]$ subject to constraints

$$\textstyle\sum_i p_i=1\quad\text{and}\quad\sum_i a_ip_i=A \tag{6.21}$$

we get

$$\frac{\partial S}{\partial p_i}=f_i(p_i,q_i)=\lambda+\mu a_i\quad\text{for all}\quad i\in\mathcal{X}\ , \tag{6.22}$$

where λ and μ are Lagrange multipliers. Solving for p_i gives the posterior,

$$P_i=g_i(q_i,\lambda,\mu,a_i) \tag{6.23}$$

for some functions g_i. As stated in Section 6.2.3 we do not assume that the labels i themselves carry any particular significance. This means, in particular, that for any proposition labelled i we want the selected posterior P_i to depend only on the prior q_i and on the constraints – that is, on λ, μ, and a_i. We do not want to have different updating rules for different propositions: two different propositions i and i' with the same priors $q_i = q_{i'}$ and the same constraints $a_i = a_{i'}$ should be updated to the same posteriors, $P_i = P_{i'}$. In other words the functions g_i and f_i must be independent of i. Therefore

$$\frac{\partial S}{\partial p_i} = f(p_i, q_i) \quad \text{for all} \quad i \in \mathcal{X} \ . \tag{6.24}$$

Integrating, one obtains

$$S[p,q] = \textstyle\sum_i F(p_i, q_i) + \text{constant} \,. \tag{6.25}$$

for some still undetermined function F. The constant has no effect on the maximization and can be dropped.

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),

$$S[p,q] = \int dx\, F\left(p(x), q(x)\right) \ . \tag{6.26}$$

Comment: 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,

$$F \overset{?}{=} F(p, q, \frac{dp}{dx}, \frac{dq}{dx}, \ldots) \ . \tag{6.27}$$

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 4.12) is to recognize the existence of additional constraints concerning continuity that must be explicitly taken into account.

DC2: Independent subsystems

Let the microstates of a composite system be labeled by $(i_1, i_2) \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$. We shall consider two special cases.

Case (a) — Suppose nothing is said about subsystem 2 and the information about subsystem 1 is extremely constraining: for subsystem 1 we maximize $S_1[p_1, q_1]$ subject to the constraint that $p_1(i_1)$ is $P_1(i_1)$, the selected posterior being, naturally, $p_1(i_1) = P_1(i_1)$. For subsystem 2 we maximize $S_2[p_2, q_2]$ subject only to normalization so there is no update, $P_2(i_2) = q_2(i_2)$.

When the systems are treated jointly, however, the inference is not nearly as trivial. We want to maximize the entropy of the joint system,

$$S[p,q]=\sum_{i_1,i_2}F\left(p(i_1,i_2),q_1(i_1)q_2(i_2)\right)\ , \tag{6.28}$$

subject to normalization,

$$\sum_{i_1,i_2}p(i_1,i_2)=1\ , \tag{6.29}$$

and the constraint on subsystem 1,

$$\textstyle\sum_{i_1}p(i_1,i_2)=P_1(i_1)\ . \tag{6.30}$$

Notice that this is not just one constraint: we have one constraint for each value of i_1, and each constraint must be supplied with its own Lagrange multiplier, $\mu_1(i_1)$. Then,

$$\delta\left[S-\textstyle\sum_{i_1}\mu_1(i_1)\left(\sum_{i_2}p(i_1,i_2)-P_1(i_1)\right)-\lambda\left(\sum_{i_1,i_2}p(i_1,i_2)-1\right)\right]=0\ . \tag{6.31}$$

The independent variations $\delta p(i_1,i_2)$ yield

$$f\ \left(p(i_1,i_2),q_1(i_1)q_2(i_2)\right)=\lambda+\mu_1(i_1)\ , \tag{6.32}$$

where f is given in (6.24),

$$\frac{\partial S}{\partial p}=\frac{\partial}{\partial p}F\left(p,q_1q_2\right)=f\left(p,q_1q_2\right)\ . \tag{6.33}$$

Next we impose that the selected posterior is the product $P_1(i_1)q_2(i_2)$. The function f must be such that

$$f\ \left(P_1q_2,q_1q_2\right)=\lambda+\mu_1\ . \tag{6.34}$$

Since the RHS is independent of the argument i_2, the f function on the LHS must be such that the i_2-dependence cancels out and this cancellation must occur for all values of i_2 and all choices of the prior q_2. Therefore we impose that for any value of x the function $f(p,q)$ must satisfy

$$f(px,qx)=f(p,q)\ . \tag{6.35}$$

Choosing $x=1/q$ in the first equation we get

$$f\left(\frac{p}{q},1\right)=f(p,q)\quad\text{or}\quad\frac{\partial F}{\partial p}=f(p,q)=\phi\left(\frac{p}{q}\right)\ . \tag{6.36}$$

Thus, the function $f(p,q)$ has been constrained to a function $\phi(p/q)$ of a single argument.

Case (b) — Next we consider a situation in which both subsystems are updated and the information is assumed to be extremely constraining: when the subsystems are treated separately $q_1(i_1)$ is updated to $P_1(i_1)$ and $q_2(i_2)$ is updated to $P_2(i_2)$. When the systems are treated jointly we require that the joint prior for the combined system $q_1(i_1)q_2(i_2)$ be updated to $P_1(i_1)P_2(i_2)$.

First we treat the subsystems separately. Maximize the entropy of subsystem 1,

$$S[p_1,q_1]=\sum\nolimits_{i_1}F\ (p_1(i_1),q_1(i_1))\quad\text{subject to}\quad p_1(i_1)=P_1(i_1)\ . \tag{6.37}$$

To each constaint — one constraint for each value of i_1 — we must supply one Lagrange multiplier, $\lambda_1(i_1)$. Then,

$$\delta\left[S-\sum\nolimits_{i_1}\lambda_1(i_1)\left(p(i_1)-P_1(i_1)\right)\right]=0\ . \tag{6.38}$$

Using eq.(6.36),

$$\frac{\partial S}{\partial p_1}=\frac{\partial}{\partial p_1}F\left(p_1,q_1\right)=\phi\left(\frac{p_1}{q_1}\right)\ , \tag{6.39}$$

and, imposing that the selected posterior be $P_1(i_1)$, we find that the function ϕ must obey

$$\phi\left(\frac{P_1(i_1)}{q_1(i_1)}\right)=\lambda_1(i_1)\ . \tag{6.40}$$

Similarly, for system 2 we find

$$\phi\left(\frac{P_2(i_2)}{q_2(i_2)}\right)=\lambda_2(i_2)\ . \tag{6.41}$$

Next we treat the two subsystems jointly. Maximize the entropy of the joint system,

$$S[p,q]=\sum_{i_1,i_2}F\ \left(p(i_1,i_2),q_1(i_1)q_2(i_2)\right)\ , \tag{6.42}$$

subject to the following constraints on the joint distribution $p(i_1,i_2)$:

$$\sum\nolimits_{i_2}p(i_1,i_2)=P_1(i_1)\qquad\text{and}\qquad\sum\nolimits_{i_1}p(i_1,i_2)=P_2(i_2)\ . \tag{6.43}$$

Again, there is one constraint for each value of i_1 and of i_2 and we introduce Lagrange multipliers, $\eta_1(i_1)$ or $\eta_2(i_2)$. Then,

$$\delta\left[S-\sum\nolimits_{i_1}\eta_1(i_1)\left(\sum\nolimits_{i_2}p(i_1,i_2)-P_1(i_1)\right)-\{1\leftrightarrow 2\}\right]=0, \tag{6.44}$$

where $\{1\leftrightarrow 2\}$ indicates a third term, similar to the second, with 1 and 2 interchanged. Using eq.(6.36),

$$\frac{\partial S}{\partial p}=\frac{\partial}{\partial p}F\left(p,q_1q_2\right)=\phi\left(\frac{p}{q_1q_2}\right)\ , \tag{6.45}$$

the independent variations $\delta p(i_1,i_2)$ yield

$$\phi\left(\frac{p(i_1,i_2)}{q_1(i_1)q_2(i_2)}\right)=\eta_1(i_1)+\eta_2(i_2)\ , \tag{6.46}$$

and we impose that the selected posterior be the product $P_1(i_1)P_2(i_2)$. Therefore, the function ϕ must be such that

$$\phi\left(\frac{P_1P_2}{q_1q_2}\right)=\eta_1+\eta_2\ . \tag{6.47}$$

To solve this equation we take the exponential of both sides,

$$\xi\left(\frac{P_1P_2}{q_1q_2}\right)=e^{\eta_1}e^{\eta_2}\quad\text{where}\quad\xi=\exp\phi\ , \tag{6.48}$$

and rewrite it as

$$\xi\left(\frac{P_1P_2}{q_1q_2}\right)e^{-\eta_2(i_2)}=e^{\eta_1(i_1)}\ . \tag{6.49}$$

This shows that for any value of i_1, the dependences in the LHS on i_2 through P_2/q_2 and η_2 must cancel each other out. In particular, if for some subset of i_2s the subsystem 2 is updated so that $P_2=q_2$, which amounts to no update at all, the i_2 dependence on the left is eliminated but the i_1 dependence remains unaffected,

$$\xi\left(\frac{P_1}{q_1}\right)e^{-\eta_2'}=e^{\eta_1(i_1)}\ . \tag{6.50}$$

where η_2' is some constant independent of i_2. A similar argument with $\{1\leftrightarrow 2\}$ yields

$$\xi\left(\frac{P_2}{q_2}\right)e^{-\eta_1'}=e^{\eta_2(i_2)}\ , \tag{6.51}$$

where η_1' is a constant. Taking the exponential of (6.40) and (6.41) leads to

$$\xi\left(\frac{P_1}{q_1}\right)e^{-\eta_2'}=e^{\lambda_1-\eta_2'}=e^{\eta_1}\quad\text{and}\quad\xi\left(\frac{P_2}{q_2}\right)e^{-\eta_1'}=e^{\lambda_2-\eta_1'}=e^{\eta_2}\ . \tag{6.52}$$

Substituting back into (6.48), we get

$$\xi\left(\frac{P_1P_2}{q_1q_2}\right)=\xi\left(\frac{P_1}{q_1}\right)\xi\left(\frac{P_2}{q_2}\right)\ , \tag{6.53}$$

where a constant factor $e^{-(\eta_1'+\eta_2')}$ has been absorbed into a new function ξ. The general solution of this functional equation is a power,

$$\xi(xy)=\xi(x)\xi(y)\Longrightarrow\xi(x)=x^a\ , \tag{6.54}$$

so that

$$\phi(x)=a\log x+b\ , \tag{6.55}$$

where a and b are constants. Finally, we can integrate (6.36),

$$\frac{\partial F}{\partial p}=\phi\left(\frac{p}{q}\right)=a\log\frac{p}{q}+b\ , \tag{6.56}$$

to get

$$F[p,q] = ap\log\frac{p}{q} + b'p + c \tag{6.57}$$

where b' and c are constants.

At this point the entropy takes the general form

$$S[p,q] = \textstyle\sum_i \left(ap_i \log\frac{p_i}{q_i} + b'p_i + c\right) . \tag{6.58}$$

The additive constant c may be dropped: it contributes a term that does not depend on the probabilities and has no effect on the ranking scheme. Furthermore, since $S[p,q]$ will be maximized subject to constraints that include normalization which is implemented by adding a term $\lambda\sum_i p_i$, the b' constant can always be absorbed into the undetermined multiplier λ. Thus, the b' term has no effect on the selected distribution and can be dropped too. 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.(6.12) and concludes our derivation.

6.4 Consistency with the law of large numbers

Entropic methods of inference are of general applicability but there exist special situations — such as, for example, those involving large numbers of independent subsystems — where inferences can be made by purely probabilistic methods without ever invoking the concept of entropy. It is important to check that the two methods of calculation are consistent with each other.

Consider a system composed of a large number N of subsystems that are independent and identical. Let the microstates for each individual system be described by a discrete variable $i = 1 \ldots m$. Let the number of subsystems found in state i be n_i, and let $f_i = n_i/N$ be the corresponding frequency.[10] The probability of a particular frequency distribution $f = (f_1 \ldots f_m)$ generated by the prior q is given by the multinomial distribution,

$$Q_N\left(f|q\right) = \frac{N!}{n_1! \ldots n_m!} q_1^{n_1} \ldots q_m^{n_m} \quad \text{with} \quad \sum_{i=1}^{m} n_i = N \ . \tag{6.59}$$

When the n_i are sufficiently large we can use Stirling's approximation,

$$\log n! = n\log n - n + \log\sqrt{2\pi n} + O(1/n) \ . \tag{6.60}$$

[10] This is the "frequency" with which one observes the microstate i in the large sample N. Perhaps 'fraction' might be a better term.

Then

$$\begin{aligned}\log Q_N\left(f|q\right) &\approx N\log N - N + \log\sqrt{2\pi N} \\ &\quad -\sum_i \left(n_i\log n_i - n_i + \log\sqrt{2\pi n_i} - n_i\log q_i\right) \\ &= -N\sum_i \frac{n_i}{N}\log\frac{n_i}{Nq_i} - \sum_i \log\sqrt{\frac{n_i}{N}} - (N-1)\log\sqrt{2\pi N} \\ &= NS[f,q] - \sum_i \log\sqrt{f_i} - (N-1)\log\sqrt{2\pi N}\ , \end{aligned} \tag{6.61}$$

where $S[f,q]$ is given by eq.(6.12),

$$S[f,q] = -\sum_i f_i\log\frac{f_i}{q_i}\ . \tag{6.62}$$

Therefore for large N can be written as

$$Q_N\left(f|q\right) \approx C_N(\prod_i f_i)^{-1/2}\exp(NS[f,q]) \tag{6.63}$$

where C_N is a normalization constant. The Gibbs inequality $S[f,q] \leq 0$, eq.(4.27), shows that for large N the probability $Q_N\left(f|q\right)$ 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

$$\frac{1}{N}\log Q_N\left(f|q\right) \approx S[f,q] + r_N\ , \tag{6.64}$$

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.

We can take this calculation one step farther and ask what is the most probable frequency distribution among the subset of distributions that satisfies a constraint such as the sample average

$$\sum_i a_i f_i = \bar{a}\ . \tag{6.65}$$

The answer is given by (6.64): for large N maximizing the probability $Q_N\left(f|q\right)$ subject to the constraint $\bar{a} = A$, 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 $\langle a\rangle = A$.

[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.[11]

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 4.6) 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.

[11] It is significant that other "entropies", *e.g.* [Renyi 1961, Aczel Daróczy 1975, Amari 1985, Tsallis 1988] do not pass this test. (Section 6.5.3 below.)

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 2.10.4 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 Informative and non-informative priors*

Stein shrinking phenomenon

6.5.3 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, 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. This version of the method was developed for linear constraints and, therefore, one should not feel justified to perform *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]

$$S_\eta[p,q]=\frac{1}{\eta(\eta+1)}\int dx\,\left[(\eta+1)p-\eta q-p^{\eta+1}q^{-\eta}\right]\ , \tag{6.66}$$

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 term linear in q)

$$S_\eta=\frac{1}{\eta}\left(1-\frac{1}{\eta+1}\int dx\,p^\eta\right)\ . \tag{6.67}$$

A related entropy

$$S'_{\eta'}=\frac{1}{\eta'}\left(1-\int dx\,p^{\eta'+1}\right) \tag{6.68}$$

has been proposed in [Tsallis 1988] and forms the foundation of a so-called non-extensive statistical mechanics (see section 5.5). Clearly these two entropies are equivalent in that they generate equivalent variational problems – maximizing S_η is equivalent to maximizing $S'_{\eta'}$. 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 2.10 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 is available not only we do not know θ, we do not know x either. Thus, the relevant space for inference is not the space 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 an integral part of the *prior* distribution; and second, 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 numerical values of the arguments x and θ which are, at this point, still unknown.

Next data is collected 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',

$$p(x) = \int d\theta \, p(\theta, x) = \delta(x - x') \ . \tag{6.69}$$

This data information constrains but is not sufficient to determine the joint distribution

$$p(x,\theta) = p(x)p(\theta|x) = \delta(x - x')p(\theta|x') \ . \tag{6.70}$$

Any choice of $p(\theta|x')$ is in principle possible. So far the formulation of the problem parallels section 2.10 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,

$$S[p,q] = -\int dx d\theta \ p(x,\theta) \log \frac{p(x,\theta)}{q(x,\theta)} \ , \tag{6.71}$$

subject to the appropriate constraints. Note that the information in the data, eq.(6.69), 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, (6.71), subject to (6.69) and normalization,

$$\delta\left\{S + \alpha\left[\textstyle\int dx d\theta \ p(x,\theta) - 1\right] + \textstyle\int dx \, \lambda(x)\left[\textstyle\int d\theta \ p(x,\theta) - \delta(x - x')\right]\right\} = 0 \ , \tag{6.72}$$

yields the joint posterior,

$$P(x,\theta) = q(x,\theta) \, \frac{e^{\lambda(x)}}{Z} \ , \tag{6.73}$$

where Z is a normalization constant, and the multiplier $\lambda(x)$ is determined from (6.69),

$$\textstyle\int d\theta\; q(x,\theta)\frac{e^{\lambda(x)}}{Z} = q(x)\frac{e^{\lambda(x)}}{Z} = \delta(x-x') \;, \tag{6.74}$$

so that the joint posterior is

$$P(x,\theta) = q(x,\theta)\frac{\delta(x-x')}{q(x)} = \delta(x-x')q(\theta|x) \;. \tag{6.75}$$

The corresponding marginal posterior probability $P(\theta)$ is

$$P(\theta) = \textstyle\int dx\, P(\theta,x) = q(\theta|x') = q(\theta)\frac{q(x'|\theta)}{q(x')} \;, \tag{6.76}$$

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

$$p(x) = \textstyle\int d\theta\, p(\theta,x) = P_D(x) \;, \tag{6.77}$$

that are easily handled by ME. Maximizing S, (6.71), subject to (6.77) and normalization, leads to

$$P(x,\theta) = P_D(x)q(\theta|x) \;. \tag{6.78}$$

The corresponding marginal posterior,

$$P(\theta) = \textstyle\int dx\, P_D(x)q(\theta|x) = q(\theta)\int dx\, P_D(x)\frac{q(x|\theta)}{q(x)} \;, \tag{6.79}$$

is known as Jeffrey's rule which we met earlier in section 2.10.

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

$$\langle f\rangle = \int dx\, p(x)f(x) = F \ . \tag{6.80}$$

Maximizing S, (6.71), subject to (6.80) and normalization,

$$\delta\left\{S+\alpha\left[\int dxd\theta\ p(x,\theta)-1\right]+\lambda\int dxd\theta\ p(x,\theta)f(x)-F\right\}=0\ , \tag{6.81}$$

yields the joint posterior,

$$P(x,\theta)=q(x,\theta)\,\frac{e^{\lambda f(x)}}{Z}\ , \tag{6.82}$$

where the normalization constant Z and the multiplier λ are obtained from

$$Z=\int dx\ q(x)e^{\lambda f(x)} \quad \text{and} \quad \frac{d\log Z}{d\lambda}=F\ . \tag{6.83}$$

The corresponding marginal posterior is

$$P(\theta)=q(\theta)\int dx\ \frac{e^{\lambda f(x)}}{Z}q(x|\theta)\ . \tag{6.84}$$

These two examples (6.79) and (6.84) 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 $f(\theta)$,

$$\int dxd\theta\, P(x,\theta)f(\theta)=\langle f(\theta)\rangle = F\ . \tag{6.85}$$

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 (6.85) be satisfied by the posterior distribution.

Maximizing the entropy (6.71) subject to normalization, the data constraint (6.69), and the moment constraint (6.85) yields the joint posterior,

$$P(x,\theta)=q(x,\theta)\frac{e^{\lambda(x)+\beta f(\theta)}}{z}\ , \tag{6.86}$$

where z is a normalization constant,

$$z = \int dx d\theta \, e^{\lambda(x)+\beta f(\theta)} q(x,\theta) \, . \tag{6.87}$$

The Lagrange multipliers $\lambda(x)$ are determined from the data constraint, (6.69),

$$\frac{e^{\lambda(x)}}{z} = \frac{\delta(x-x')}{Zq(x')} \quad \text{where} \quad Z(\beta,x') = \int d\theta \, e^{\beta f(\theta)} q(\theta|x') \ , \tag{6.88}$$

so that the joint posterior becomes

$$P(x,\theta) = \delta(x-x') q(\theta|x') \frac{e^{\beta f(\theta)}}{Z} \ . \tag{6.89}$$

The remaining Lagrange multiplier β is determined by imposing that the posterior $P(x,\theta)$ satisfy the constraint (6.85). This yields an implicit equation for β,

$$\frac{\partial \log Z}{\partial \beta} = F \ . \tag{6.90}$$

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

$$P(\theta) = q(\theta|x') \frac{e^{\beta f(\theta)}}{Z} = q(\theta) \frac{q(x'|\theta)}{q(x')} \frac{e^{\beta f(\theta)}}{Z} \ . \tag{6.91}$$

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 vice 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.(6.77), we also know that the expected value over θ of a function $f(x,\theta)$ is

$$\langle f \rangle_x = \int d\theta \, p(\theta|x) f(x,\theta) = F(x) \ . \tag{6.92}$$

We seek a posterior $P(x,\theta)$ that maximizes (6.71). Introducing Lagrange multipliers α, $\lambda(x)$, and $\gamma(x)$,

$$\begin{aligned} 0 = \delta \big\{ & S + \alpha \left[\int dx d\theta \ p(x,\theta) - 1 \right] + \int dx \, \lambda(x) \left[\int d\theta \ p(x,\theta) - P_D(x) \right] \\ & + \int dx \, \gamma(x) \left[\int d\theta \ p(x,\theta) f(x,\theta) - P_D(x) F(x) \right] \big\} \ , \end{aligned} \tag{6.93}$$

the variation over $p(x,\theta)$ yields

$$P(x,\theta)=\frac{1}{\zeta}q(x)q(\theta)\,e^{\lambda(x)+\gamma(x)f(x,\theta)}\ , \tag{6.94}$$

where ζ is a normalization constant. The multiplier $\lambda(x)$ is determined from (6.77),

$$P(x)=\int d\theta\,P(\theta,x)=\frac{1}{\zeta}q(x)e^{\lambda(x)}\int d\theta\,q(\theta)\,e^{\gamma(x)f(x,\theta)}=P_D(x) \tag{6.95}$$

then,

$$P(x,\theta)=P_D(x)\frac{q(\theta)\,e^{\gamma(x)f(x,\theta)}}{\int d\theta'\,q(\theta')\,e^{\gamma(x)f(x,\theta')}} \tag{6.96}$$

so that

$$P(\theta|x)=\frac{P(x,\theta)}{P(x)}=\frac{q(\theta)\,e^{\gamma(x)f(x,\theta)}}{Z(x)}\quad\text{with}\quad Z(x)=\int d\theta'\,q(\theta')\,e^{\gamma(x)f(x,\theta')} \tag{6.97}$$

The multiplier $\gamma(x)$ is determined from (6.92)

$$\frac{1}{Z(x)}\frac{\partial Z(x)}{\partial\gamma(x)}=F(x)\ . \tag{6.98}$$

The corresponding marginal posterior is

$$P(\theta)=\int dx\,P_D(x)P(\theta|x)=q(\theta)\int dx\,P_D(x)\frac{e^{\gamma(x)f(x,\theta)}}{Z(x)}\ . \tag{6.99}$$

In the limit when the data are sharply determined $P_D(x)=\delta(x-x')$ the posterior takes the form of Bayes theorem,

$$P(\theta)=q(\theta)\,\frac{e^{\gamma(x')f(x',\theta)}}{Z(x')}\ , \tag{6.100}$$

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 several independent experiments. In this case the order in which the observed data $x' = \{x'_1, x'_2, \ldots\}$ is processed does not matter for the purpose of inferring θ. (See section 2.10.3) The proof that ME is completely compatible with Bayes' rule implies that data constraints implemented through δ functions, as in (6.69), 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 6.2. Next we receive a second piece of information described by the constraint $\mathcal{C}_2$. At this point we can proceed in two very 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.

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.

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

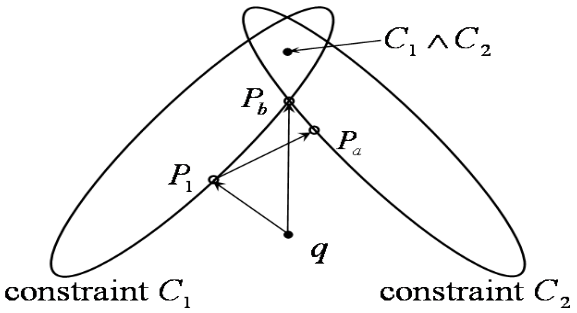

Figure 6.2: Illustrating the difference between processing two constraints C_1 and C_2 sequentially ($q \to P_1 \to P_a$) and simultaneously ($q \to P_b$ or $q \to P_1 \to P_b$).

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 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; and it includes MaxEnt and Bayes' rule as special cases.

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? The discussion of these matters in the next two chapters will 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 slightly different distribution? Would it have made any difference? 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 $\theta = (\theta^1 \ldots \theta^n)$ — 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 an intrinsic notion of distance. A metric structure is an optional addition that has to be supplied separately. This is done by singling out one privileged metric tensor — or just, *the metric* — from a family of potential candidates. Statistical manifolds are, however, an exception. One of the main goals of this chapter is to show that statistical manifolds already come endowed with a uniquely natural notion of distance — the so-called information metric. And the geometry is not optional — it is an intrinsic structural element that characterizes 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 that we also have a notion of volume and this in turn implies that there is a unique and epistemically objective notion of a probability distribution that is uniform over the space of parameters — that is, a distribution that assigns equal probabilities to equal volumes. Whether such distributions are maximally non-informative, or whether they define ignorance, or whether they reflect the actual prior beliefs of any rational agent, are all potentially important issues but they are quite beside the specific point that we want to emphasize, namely, that they are *uniform.*

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 coordinates (that is a map $\mathcal{M} \to \mathbb{R}^n$) so that each point $\theta \in \mathcal{M}$ is identified or labelled by n numbers, $\theta = (\theta^1 \dots \theta^n)$.

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 $p_\theta(x) = p(x|\theta)$.

Here are some examples:

The multinomial distributions are given by

$$p(\{m_i\}|\theta) = \frac{N!}{m_1!m_2!\dots m_n!}(\theta^1)^{m_1}(\theta^2)^{m_2}\dots(\theta^n)^{m_n} \ , \tag{7.2}$$

where $\theta = (\theta^1, \theta^2 \dots \theta^n)$,

$$N = \textstyle\sum_{i=1}^n m_i \quad \text{and} \quad \textstyle\sum_{i=1}^n \theta^i = 1 \ . \tag{7.3}$$

They form a statistical manifold of dimension $(n-1)$ called a simplex, $\mathcal{S}_{n-1}$. The parameters $\theta = (\theta^1, \theta^2 \dots \theta^n)$ are a convenient choice of coordinates.

The multivariate Gaussian distributions with means μ^a, $a = 1 \dots n$, and variance σ^2,

$$p(x|\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp -\frac{1}{2\sigma^2}\sum_{a=1}^n (x^a - \mu^a)^2 \ , \tag{7.4}$$

form an $(n+1)$-dimensional statistical manifold. A convenient choice of coordinates is $\theta = (\mu^1, \dots, \mu^n, \sigma)$.

[1] The use of superscripts rather than subscripts for the indices labelling coordinates is a common 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, $g_{ab}f^{ab} = \sum_a \sum_b g_{ab} f^{ab}$. Furthermore, we shall follow standard practice and indistinctly refer to both the metric tensor g_{ab} and the quadratic form (7.1) as the 'metric'.

The canonical distributions, eq.(4.82),

$$p(i|F) = \frac{1}{Z} e^{-\lambda_k f_i^k} \ , \tag{7.5}$$

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 $k = 1, 2, \ldots n$,

$$\left\langle f^k \right\rangle = \sum_i p_i f_i^k = F^k \ . \tag{7.6}$$

They form an n-dimensional statistical manifold. As coordinates we can either use the expected values $F = (F^1 \ldots F^n)$ or, equivalently, the Lagrange multipliers, $\lambda = (\lambda_1 \ldots \lambda_n)$.

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, force 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 define the infinitesimal displacement from a point $x = (x^1 \ldots x^n)$ by a vector,

$$d\vec{x} = (dx^1 \ldots dx^n) \ , \tag{7.7}$$

and then define finite vectors, such as velocities $\vec{v}$, through multiplication by a suitably large scalar $1/dt$, so that $\vec{v} = d\vec{x}/dt$.

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 would have to 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}.

For physics the consequences of curvature 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 in a curved space the operation of subtracting two vectors in different tangent spaces is no longer available to us. Similarly, there is no general definition of the total energy or the total momentum of an extended system of particles; the individual momenta are vectors that live in different tangent spaces and there is no unique way to add them.

One objection to using a displacement as the starting point to the construction of a vector that is visualized as an arrow is that it relies on our intuition of a curved space as being embedded in a flat space of larger dimension. A number of questions may be raised: What do we gain by introducing these larger embedding flat spaces? Do they physically exist? And what is so special about flatness anyway? So, while visualizing curved spaces in this way can be useful, it is also a good idea to pursue alternative approaches that do not rely on embedding.

Vectors as tangents to curves

One 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 $a = 1 \ldots n$. 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 $\{dx^a/d\lambda\}$,

$$\vec{v} \sim (\frac{dx^1}{d\lambda} \ldots \frac{dx^n}{d\lambda}) \ . \tag{7.8}$$

(The symbol $\sim$ stands for "is represented by".) This definition relies on the notion of coordinates which depend on the choice of frame and therefore so do the components of vectors. When we change to new coordinates,

$$x^{a'} = f^{a'}(x^1 \ldots x^n) \ , \tag{7.9}$$

the components of the vector in the new frame change accordingly. The new components are given by the chain rule,[2]

$$\vec{v} \sim (\frac{dx^{1'}}{d\lambda} \ldots \frac{dx^{n'}}{d\lambda}) \quad \text{where} \quad \frac{dx^{a'}}{d\lambda} = \frac{\partial x^{a'}}{\partial x^b} \frac{dx^b}{d\lambda} \ . \tag{7.10}$$

In this approach a vector at the point x is defined as an n-tuple of real numbers $(v^1 \ldots v^n)$ that under a change of coordinate frame transform according to

$$v^{a'} = X_b^{a'} v^b \quad \text{where} \quad X_b^{a'} = \frac{\partial x^{a'}}{\partial x^b} \ . \tag{7.11}$$

[2] The notation is standard: the primed frame is indicated by priming the indices, not the quantity — that is $x^{a'}$ rather than x'^a. A derivative with respect to a variable labelled by an upper index transforms as a variable with a lower index and vice versa. For example, $\partial f/\partial x^a = \partial_a f$, and in eq.(7.10) the index b is summed over.

In other words, the vector has different representations in different frames but 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

$$e_a^b = \delta_a^b \ . \tag{7.12}$$

More explicitly,

$$\vec{e}_1 \sim (1, 0 \ldots 0), \ \vec{e}_2 \sim (0, 1, 0 \ldots 0), \ \ldots, \ \vec{e}_1 \sim (0, 0 \ldots 1) \ . \tag{7.13}$$

Any vector $\vec{v}$ can be expressed in terms of the basis vectors,

$$\vec{v} = v^a \, \vec{e}_a \ . \tag{7.14}$$

The basis vectors in the primed frame $\{\vec{e}_{a'}\}$ are defined in the same way

$$e_{a'}^{b'} = \delta_{a'}^{b'} \ . \tag{7.15}$$

so that using eq.(7.11) we have

$$\vec{v} = v^{a'} \, \vec{e}_{a'} = X_b^{a'} v^b \, \vec{e}_{a'} = v^b \ \vec{e}_b \ , \tag{7.16}$$

where

$$\vec{e}_b = X_b^{a'} \ \vec{e}_{a'} \quad \text{or, equivalently,} \quad \vec{e}_{a'} = X_{a'}^b \ \vec{e}_b \ . \tag{7.17}$$

Eq.(7.16) 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.(7.17) shows that the invariance follows from the fact that components and basis vectors transform according to inverse matrices. Explicitly, using the chain rule,

$$X_b^{a'} X_{c'}^b = \frac{\partial x^{a'}}{\partial x^b} \frac{\partial x^b}{\partial x^{c'}} = \frac{\partial x^{a'}}{\partial x^{c'}} = \delta_{c'}^{a'} \ . \tag{7.18}$$

Remark: Eq.(7.16) is the main reason we care about vectors: they are objects that are independent of the accidents of the particular choice of coordinate system. Therefore they are good candidates to represent quantities that carry physical meaning. Conversely, since we can always change coordinates, it is commonly thought that discussions that avoid coordinates and employ methods of analysis that are coordinate-free are somehow deeper or more fundamental. The introduction of coordinates is often regarded as a blemish that is barely tolerated because it often facilitates computation. However, when we come to statistical manifolds the situation is different. In this case coordinates can be parameters in probability distributions (such as, for example, temperatures or chemical potentials) that carry a statistical and physical meaning and therefore have a significance that goes far beyond the mere geometrical role of labelling points. Thus, there is much to gained from recognizing, on one hand, the geometrical freedom to assign coordinates and, on the other hand, that often enough special coordinates are singled out because they carry physically relevant information. The case can therefore be made that when dealing with statistical manifolds coordinate-dependent methods can in many respects be fundamentally superior.

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,

$$\frac{d\phi}{d\lambda}=\frac{\partial\phi}{\partial x^a}\frac{dx^a}{d\lambda}=\frac{\partial\phi}{\partial x^a}v^a \tag{7.19}$$

Note that $d\phi/d\lambda$ is independent of the choice of coordinates. Indeed, using the chain rule

$$\frac{d\phi}{d\lambda}=\frac{\partial\phi}{\partial x^a}v^a=\frac{\partial\phi}{\partial x^{a'}}\frac{\partial x^{a'}}{\partial x^a}v^a=\frac{\partial\phi}{\partial x^{a'}}v^{a'}\ . \tag{7.20}$$

Since ϕ is any generic function we can write

$$\frac{d}{d\lambda}=v^a\frac{\partial}{\partial x^a}\ . \tag{7.21}$$

Note further that the partial derivatives $\partial/\partial x^a$ transform exactly as the basis vectors, eq.(7.17)

$$\frac{\partial}{\partial x^a}=\frac{\partial x^{a'}}{\partial x^a}\frac{\partial}{\partial x^{a'}}=X_a^{a'}\frac{\partial}{\partial x^{a'}}\ , \tag{7.22}$$

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 $x(\lambda)$. In fact, we can use one to represent the other,

$$\vec{v}\sim\frac{d}{d\lambda}\quad\text{and}\quad\vec{e}_a\sim\frac{\partial}{\partial x^a}\ . \tag{7.23}$$

and, since mathematical objects are defined purely through their formal rules of manipulation, it is common practice to ignore the distinction between the two concepts and set

$$\vec{v}=\frac{d}{d\lambda}\quad\text{and}\quad\vec{e}_a=\frac{\partial}{\partial x^a}\ . \tag{7.24}$$

The partial derivative is indeed appropriate because the "vector" $\partial/\partial x^a$ is the derivative along those curves $x^b(\mu)$ parametrized by the parameter $\mu=x^a$ that are defined by keeping the other coordinates constant, $x^b(\mu)=\text{const}$, for $b\neq a$. The vector $\vec{e}_a$ has components

$$e_a^b=\frac{\partial x^b}{\partial x^a}=\delta_a^b\ . \tag{7.25}$$

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 generic manifolds; it has to be supplied as an additional structure — the metric tensor. As we shall see statistical manifolds constitute a most remarkable exception.

To introduce 'distance' one follows the basic intuition 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 $x^{\hat{a}} = f^{\hat{a}}(x^1 \ldots x^n)$ that we *declare* to be locally Cartesian (here denoted with ˆ superscripts). An infinitesimal displacement has components given by

$$dx^{\hat{a}} = X^{\hat{a}}_a \, dx^a \quad \text{where} \quad X^{\hat{a}}_a = \frac{\partial x^{\hat{a}}}{\partial x^a} \ . \tag{7.26}$$

In these locally Cartesian coordinates, also called *normal coordinates*, the corresponding infinitesimal distance can be computed using Pythagoras theorem,

$$d\ell^2 = \delta_{\hat{a}\hat{b}} dx^{\hat{a}} dx^{\hat{b}} \ . \tag{7.27}$$

Remark: The fact that at any given point one can always change to normal coordinates such that Pythagoras' theorem is locally valid is what defines a Riemannian manifold. However, it is important to realize that while one can do this at any single arbitrary point of our choice, in general one cannot find a coordinate frame in which eq.(7.27) is simultaneously valid at all points within an extended region.

Changing back to the original frame

$$d\ell^2 = \delta_{\hat{a}\hat{b}} dx^{\hat{a}} dx^{\hat{b}} = \delta_{\hat{a}\hat{b}} X^{\hat{a}}_a \, X^{\hat{b}}_b \, dx^a dx^b \ . \tag{7.28}$$

Defining the quantities

$$g_{ab} \overset{\text{def}}{=} \delta_{\hat{a}\hat{b}} X^{\hat{a}}_a \, X^{\hat{b}}_b \ , \tag{7.29}$$

we can write the infinitesimal Pythagoras theorem in the original generic frame as

$$d\ell^2 = g_{ab} dx^a dx^b \ . \tag{7.30}$$

The quantities g_{ab} are the components of the metric tensor usually abbreviated to just the 'metric'. One can easily check that under a coordinate transformation g_{ab} transforms according to

$$g_{ab} = X^{a'}_a X^{b'}_a g_{a'b'} \ , \tag{7.31}$$

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,

$$\ell = \int_{\lambda_1}^{\lambda_2} d\ell = \int_{\lambda_1}^{\lambda_2} \left(g_{ab} \frac{dx^a}{d\lambda} \frac{dx^b}{d\lambda} \right)^{1/2} d\lambda \ . \tag{7.32}$$

Once we have defined a measure of distance we can 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 normal coordinates $x^{\hat{a}}$ so that the volume element is simply given by the product

$$dV_n = dx^{\hat{1}}dx^{\hat{2}}\ldots dx^{\hat{n}} \ , \tag{7.33}$$

and then transform back to the original coordinates x^a using eq.(7.26),

$$dV_n = \left|\frac{\partial \hat{x}}{\partial x}\right| dx^1 dx^2 \ldots dx^n = \left|\det X_a^{\hat{a}}\right| d^n x \ . \tag{7.34}$$

This is the volume element written in terms of the coordinates x^a but we still have to calculate the Jacobian of the transformation, $|\partial\hat{x}/\partial x| = \left|\det X_a^{\hat{a}}\right|$. This is done noting that the transformation of the metric from its Euclidean form $\delta_{\hat{a}\hat{b}}$ to g_{ab}, eq.(7.29), is the product of three matrices. Taking the determinant we get

$$g \overset{\text{def}}{=} \det(g_{ab}) = \left[\det X_a^{\hat{a}}\right]^2 \ , \tag{7.35}$$

so that

$$\left|\det\left(X_a^{\hat{a}}\right)\right| = g^{1/2} \ . \tag{7.36}$$

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

$$dV_n = g^{1/2}(x) d^n x \ . \tag{7.37}$$

The volume of any extended region R on the manifold is

$$V_n = \int_R dV_n = \int_R g^{1/2}(x) d^n x \ . \tag{7.38}$$

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

$$d\ell^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2 \ , \tag{7.39}$$

and the corresponding metric tensor is

$$(g_{ab}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2\sin^2\theta \end{pmatrix} \ . \tag{7.40}$$

The volume element is the familiar expression

$$dV = g^{1/2} dr d\theta d\phi = r^2 \sin\theta \, dr d\theta d\phi \ . \tag{7.41}$$

Important example: A uniform distribution over such a curved manifold is one which assigns equal probabilities to equal volumes. Therefore,

$$p(x) d^n x \propto g^{1/2}(x) d^n x \ . \tag{7.42}$$

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 a metric tensor g_{ab}. Our goal is to propose a metric tensor g_{ab} for the statistical manifold of distributions $\{p(x|\theta)\}$. 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. The fact that so many different arguments all lead to the same tensor is significant. It strongly suggests that the information metric is special, and indeed, in Section 7.5 we shall show that up to an overall constant factor reflecting the choice of units of length this metric is unique.

Remark: 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 give up on geometry? Of course not. 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 physical surfaces with rugged geometries detracts from the general value of geometry itself.

7.4.1 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. 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,

$$\Delta = \frac{p(x|\theta+d\theta)-p(x|\theta)}{p(x|\theta)} = \frac{\partial \log p(x|\theta)}{\partial \theta^a}\, d\theta^a . \tag{7.43}$$

The expected value of the relative difference, $\langle\Delta\rangle$, might at first sight seem a good candidate to measure distinguishability, but it does not work because it vanishes identically,

$$\langle\Delta\rangle = \int dx\, p(x|\theta)\, \frac{\partial \log p(x|\theta)}{\partial \theta^a}\, d\theta^a = d\theta^a\, \frac{\partial}{\partial \theta^a} \int dx\, p(x|\theta) = 0 . \tag{7.44}$$

(Depending on the problem the symbol $\int dx$ will be used to represent either discrete sums or integrals over one or more dimensions.) However, the variance does not vanish,

$$d\ell^2 = \langle\Delta^2\rangle = \int dx\, p(x|\theta)\, \frac{\partial \log p(x|\theta)}{\partial \theta^a}\, \frac{\partial \log p(x|\theta)}{\partial \theta^b}\, d\theta^a d\theta^b \ . \tag{7.45}$$

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}

$$g_{ab} \overset{\text{def}}{=} \int dx\, p(x|\theta)\, \frac{\partial \log p(x|\theta)}{\partial \theta^a}\, \frac{\partial \log p(x|\theta)}{\partial \theta^b} \tag{7.46}$$

called the Fisher information *matrix* [Fisher 1925], so that

$$d\ell^2 = g_{ab}\, d\theta^a d\theta^b\ . \tag{7.47}$$

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 the $d\theta^a$ vanish. Thus it is natural to interpret g_{ab} as the metric tensor of a Riemannian space. This is the *information metric.* The realization by C. R. Rao that g_{ab} is a metric in the space of probability distributions [Rao 1945] gave rise to the subject of information geometry [Amari 1985], namely, the application of geometrical methods to problems in inference and in information theory.

Remark: The derivation of (7.46) involved a Taylor expansion, eq.(7.43), to first order in $d\theta$. One might wonder if keeping higher orders in $d\theta$ would lead to a “better” metric. The answer is no. What is being done here is defining the metric tensor, not finding an approximation to it. To illustrate this point consider the following analogy. The trajectory of a moving particle is given by $x^a = x^a(t)$. The position at time $t + \Delta t$ is given by the Taylor expansion

$$x^a(t+\Delta t) = x^a(t) + v^a(t)\Delta t + \frac{1}{2} a^a(t) \Delta t^2 + \ldots \tag{7.48}$$

The velocity of the particle is defined by the term linear in Δt. Despite the Taylor expansion the definition of velocity is exact, no approximation is involved; the higher order terms do not constitute an improvement to the notion of velocity.

Other useful expressions for the information metric are

$$\begin{aligned} g_{ab} &= 4 \int dx\ \frac{\partial p^{1/2}(x|\theta)}{\partial \theta^a}\, \frac{\partial p^{1/2}(x|\theta)}{\partial \theta^b} \\ &= -4 \int dx\ p^{1/2}(x|\theta)\, \frac{\partial^2 p^{1/2}(x|\theta)}{\partial \theta^a \partial \theta^b}\ , \end{aligned} \tag{7.49}$$

and

$$g_{ab} = -\int dx\ p(x|\theta)\, \frac{\partial^2 \log p(x|\theta)}{\partial \theta^a \partial \theta^b} = -\langle \frac{\partial^2 \log p_\theta}{\partial \theta^a \partial \theta^b} \rangle\ . \tag{7.50}$$

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

$$\theta^{a'} = f^{a'}(\theta^1 \ldots \theta^n) \tag{7.51}$$

leads to

$$d\theta^a = \frac{\partial\theta^a}{\partial\theta^{a'}} d\theta^{a'} \quad \text{and} \quad \frac{\partial}{\partial\theta^a} = \frac{\partial\theta^{a'}}{\partial\theta^a}\frac{\partial}{\partial\theta^{a'}} \tag{7.52}$$

so that, substituting into eq.(7.46),

$$g_{ab} = \frac{\partial\theta^{a'}}{\partial\theta^a}\frac{\partial\theta^{b'}}{\partial\theta^b} g_{a'b'} \tag{7.53}$$

7.4.2 From embedding in a Euclidean space

Consider a discrete variable $i = 1 \ldots m$. 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 $(p^1 \ldots p^m)$. The corresponding statistical manifold is the simplex $\mathcal{S}_{m-1} = \{p = (p^1 \ldots p^m) : \sum_i p^i = 1\}$.

Next change to new coordinates $\xi^i = (p^i)^{1/2}$. In these new coordinates the equation for the simplex $\mathcal{S}_{m-1}$ — the normalization condition — reads

$$\sum\nolimits_{i=1}^{m} (\xi^i)^2 = 1 \ , \tag{7.54}$$

which, if we interpret the ξ^i as Cartesian coordinates, is recognized as the equation of an $(m-1)$-sphere embedded in an m-dimensional Euclidean space $\mathbb{R}^m$. This suggests that we assign the simplest possible metric: the distance between the distribution $p(i|\xi)$ and its neighbor $p(i|\xi + d\xi)$ is the Euclidean distance in $\mathbb{R}^m$,

$$d\ell^2 = \sum_i (d\xi^i)^2 = \delta_{ij} d\xi^i d\xi^j \ . \tag{7.55}$$

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 substitute

$$d\xi^i = \frac{1}{2}\frac{dp^i}{p^{1/2}(i)} \tag{7.56}$$

to get

$$d\ell^2 = \frac{1}{4}\sum_i \frac{(dp^i)^2}{p(i)} = g_{ij} dp^i dp^j \quad \text{with} \quad g_{ij} = \frac{1}{4}\frac{\delta_{ij}}{p(i)} \ . \tag{7.57}$$

Remark: In (7.56) and in g_{ij} above we have gone back to the original notation $p(i)$ rather than p^i to emphasize that the repeated index i is not being summed over.

Except for an overall constant (7.57) is the same information metric (7.47) we defined earlier. Indeed, consider an n-dimensional subspace ($n \leq m-1$) of the simplex $\mathcal{S}_{m-1}$ defined by $\xi^i = \xi^i(\theta^1, \ldots, \theta^n)$. 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

$$\begin{aligned} d\ell^2 = \delta_{ij} d\xi^i d\xi^j &= \delta_{ij} \frac{\partial \xi^i}{\partial \theta^a} d\theta^a \frac{\partial \xi^j}{\partial \theta^b} d\theta^b \\ &= \frac{1}{4} \sum_i p^i \frac{\partial \log p^i}{\partial \theta^a} \frac{\partial \log p^i}{\partial \theta^b} d\theta^a d\theta^b \ , \end{aligned} \tag{7.58}$$

which (except for the factor 1/4) we recognize as the discrete version of (7.46) and (7.47).

This interesting result does not constitute a "derivation." There is a priori no reason why the square root coordinates ξ^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 (7.46).

7.4.3 From embedding in a spherically symmetric space

Instead of embedding the simplex $\mathcal{S}_{m-1}$ in a very special flat Euclidean space we can slightly improve the derivation of the previous section by embedding $\mathcal{S}_{m-1}$ in a somewhat more general curved space — a spherically symmetric space. Just as in the previous section the strategy is to use the known geometry of the larger embedding space to find the metric it induces on the embedded simplex.

The generalization we pursue here offers a number of advantages. It allows us to find the metric tensor in its most general form. As we shall later see in section 7.5 it applies even for probability distributions that are not normalized. More importantly, this derivation also has the advantage of emphasizing the central role played by the rotational symmetry, which will be useful in Chapters 10 and 11 where we derive quantum mechanics as an application of entropic and information geometry methods.

As in the previous section we transform to new coordinates $\xi^i = (p^i)^{1/2}$. In these new coordinates the equation for the simplex $\mathcal{S}_{m-1}$ — the normalization condition — reads $\sum (\xi^i)^2 = 1$, which we *declare* to be the equation of an $(m-1)$-sphere embedded in an m-dimensional spherically symmetric space.

What makes an m-dimensional curved space spherically symmetric is that — like an onion — it can be foliated into spheres. This means that every point in the space lies on the $(m-1)$-dimensional "surface" of some sphere,

$$\sum\nolimits_{i=1}^{m} (\xi^i)^2 = r^2 = \text{const.} \tag{7.59}$$

Unlike the flat Euclidean case, in these curved spaces the "radius" r is not to be interpreted as the radial distance to the center; r is just a quantity that labels the different spheres.

To construct the metric of a generic spherically symmetric space consider a short segment that stretches from ξ^i to $\xi^i + d\xi^i$. The length of a segment is given by the Pythagoras form

$$d\ell^2 = d\ell_r^2 + d\ell_t^2 \ , \tag{7.60}$$

where $d\ell_r$ is the length of the segment in the radial direction (*i.e.*, normal to the sphere) and $d\ell_t$ is the length tangent to the sphere. To calculate $d\ell_r$ consider two neighboring spheres of radii r and $r + dr$. Differentiating (7.59) gives,

$$dr = \sum\nolimits_{i=1}^{m} \frac{\xi^i d\xi^i}{r} \ . \tag{7.61}$$

If the sphere were embedded in flat space dr would itself be the radial distance between the two spheres. In our curved space it is not; the best we can do is to claim the actual radial distance is proportional to dr. Therefore

$$d\ell_r^2 = \frac{a(r^2)}{r^2} \left(\sum\nolimits_i \xi^i d\xi^i\right)^2 \tag{7.62}$$

where by spherical symmetry the (positive) function $a(r^2)$ depends only on r^2 so that it is constant over the surface of the sphere. To calculate the tangential length $d\ell_t$ we note that the actual geometry of the sphere is independent of the space in which it is embedded. If it were embedded in flat space then the tangential length would be

$$(\text{tang. length})^2 = \sum\nolimits_i (d\xi^i)^2 - dr^2 \tag{7.63}$$

In our curved space, the actual tangential distance can involve an overall scale factor,

$$d\ell_t^2 = b(r^2) \left(\sum\nolimits_i (d\xi^i)^2 - \left(\frac{1}{r^2} \sum\nolimits_i \xi^i d\xi^i \right)^2 \right) \ . \tag{7.64}$$

By spherical symmetry the (positive) scale factor $b(r^2)$ depends only on r. Substituting into (7.60), we see that the metric of a generic spherically symmetric space involves two arbitrary positive functions of r^2,

$$d\ell^2 = \frac{1}{r^2} \left[a(r^2) - b(r^2) \right] \left(\sum\nolimits_i \xi^i d\xi^i\right)^2 + b(r^2) \sum\nolimits_i (d\xi^i)^2. \tag{7.65}$$

Expressed in terms of the original p^i coordinates the metric of a spherically symmetric space takes the form

$$d\ell^2 = A(|p|) \left(\sum\nolimits_i dp^i\right)^2 + B(|p|) \sum\nolimits_i \frac{1}{2p^i} (dp^i)^2 \ , \tag{7.66}$$

where

$$A(|p|) = \frac{1}{4|p|} \left[a(|p|) - b(|p|) \right] \quad \text{and} \quad B(|p|) = \frac{1}{2} b(|p|) \tag{7.67}$$

and

$$|p| \overset{\text{def}}{=} \sum\nolimits_i p^i = r^2 \ . \tag{7.68}$$

Setting

$$|p| = \sum\nolimits_i p^i = 1 \quad \text{and} \quad \sum\nolimits_i dp^i = 0 \ , \tag{7.69}$$

gives the metric induced on the simplex $\mathcal{S}_{m-1}$,

$$d\ell^2 = B(1) \sum\nolimits_i \frac{1}{p^i} (dp^i)^2 \ . \tag{7.70}$$

Up to an overall scale this result agrees with previous expressions for the information metric.

7.4.4 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 $p = (p_1 \ldots p_n)$. (Here it is slightly more convenient to revert to the notation where indices appear as subscripts.) Suppose the data consists of the numbers $(m_1 \ldots m_m)$ where m_i is the number of times that outcome i occurs. The corresponding frequencies are

$$f_i = \frac{m_i}{N} \quad \text{with} \quad \sum\nolimits_{i=1}^{n} m_i = N \ . \tag{7.71}$$

The probability of a particular frequency distribution $f = (f_1 \ldots f_n)$ is

$$P_N\,(f|p) = \frac{N!}{m_1! \ldots m_n!} p_1^{m_1} \ldots p_n^{m_n} \ . \tag{7.72}$$

For sufficiently large N and m_i we can use Stirling's approximation [see eq.(6.63)], to get

$$P_N\,(f|p) \approx C_N (\prod_i f_i)^{-1/2} \exp(NS[f,p]) \tag{7.73}$$

where C_N is a normalization constant and $S[f,p]$ is the entropy given by eq.(6.12). The Gibbs inequality $S[f,p] \leq 0$, eq.(4.27), 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,

$$P_N\,(p|f) \propto \exp(NS[f,p]) \ , \tag{7.74}$$

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

$$NS[p,p'] = NS[p,p+\delta p] \approx -1 \tag{7.75}$$

so that the probability $P_N\,(p'|f)$ is down by e^{-1} from the maximum. Expanding to second order,

$$S[p,p+\delta p] = -\sum_i p_i \log \frac{p_i}{p_i + \delta p_i} \approx -\frac{1}{2} \sum_i \frac{(\delta p_i)^2}{p_i} \tag{7.76}$$

Thus, the nearest that two neighboring points p and $p + \delta p$ can be while still being distinguishable in N trials is such that

$$\frac{N}{2} \sum_i \frac{(\delta p_i)^2}{p_i} \approx 1 \ . \tag{7.77}$$

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,

$$\text{Statistical length} = \ell = \lim_{N\to\infty} \frac{1}{\sqrt{N/2}} \begin{bmatrix} \text{number of distinguishable (in } N \text{ trials)} \\ \text{distributions that fit along the curve} \end{bmatrix} \tag{7.78}$$

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

$$\frac{N}{2}\sum_i \frac{1}{p_i}(\frac{dp_i}{d\lambda})^2 \delta\lambda^2 \approx 1 \quad \text{or} \quad \delta\lambda \approx \left(\frac{N}{2}\sum_i \frac{1}{p_i}(\frac{dp_i}{d\lambda})^2\right)^{-1/2} . \tag{7.79}$$

The number of distinguishable distributions within the interval $d\lambda$ is $d\lambda/\delta\lambda$ and the corresponding statistical length, eq.(7.78), is

$$d\ell = \left(\sum_i \frac{1}{p_i}(\frac{dp_i}{d\lambda})^2\right)^{1/2} d\lambda = \left(\sum_i \frac{(dp_i)^2}{p_i}\right)^{1/2} \tag{7.80}$$

The length of the curve from λ_0 to λ_1 is

$$\ell = \int_{\lambda_0}^{\lambda_1} d\ell \quad \text{where} \quad d\ell^2 = \sum_i \frac{(dp_i)^2}{p_i} . \tag{7.81}$$

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 hard-wired 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 (7.81) agrees the previous definitions of the information metric. Consider the n-dimensional subspace ($n \le m-1$) 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

$$d\ell^2 = \sum_{i=1}^{m} \frac{(\delta p_i)^2}{p_i} = \sum_{i,j=1}^{m} \frac{1}{p_i}\frac{\partial p_i}{\partial \theta^a} d\theta^a \frac{\partial p_j}{\partial \theta^b} d\theta^b = g_{ab} d\theta^a d\theta^b \tag{7.82}$$

where

$$g_{ab}=\sum_{i=1}^{m}p_i\frac{\partial\log p_i}{\partial\theta^a}\frac{\partial\log p_i}{\partial\theta^b}\ , \tag{7.83}$$

which is the discrete version of (7.46).

7.4.5 From relative entropy

The relation we uncovered above between the information metric and entropy, eq.(7.76), is not restricted to multinomials; it is quite general. Consider the entropy of one distribution $p(x|\theta')$ relative to another $p(x|\theta)$,

$$S(\theta',\theta)=-\int dx\,p(x|\theta')\log\frac{p(x|\theta')}{p(x|\theta)}\ . \tag{7.84}$$

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 $d\theta$

$$S(\theta+d\theta,\theta)=\frac{1}{2}\frac{\partial^2S(\theta',\theta)}{\partial\theta'^a\partial\theta'^b}\bigg|_{\theta'=\theta}d\theta^ad\theta^b+\ldots\leq 0\ , \tag{7.85}$$

which suggests defining the distance $d\ell$ by

$$S(\theta+d\theta,\theta)=-\frac{1}{2}d\ell^2\ . \tag{7.86}$$

The second derivative is

$$\begin{aligned}-\frac{\partial^2S(\theta',\theta)}{\partial\theta'^a\partial\theta'^b}&=\frac{\partial}{\partial\theta'^a}\int dx\left(\log\frac{p(x|\theta')}{p(x|\theta)}+1\right)\frac{\partial p(x|\theta')}{\partial\theta'^b}\\&=\int dx\left(\frac{\partial\log p(x|\theta')}{\partial\theta'^a}\frac{\partial p(x|\theta')}{\partial\theta'^b}+[\log\frac{p(x|\theta')}{p(x|\theta)}+1]\frac{\partial^2p(x|\theta')}{\partial\theta'^a\partial\theta'^b}\right)\ .\end{aligned}$$

Evaluating at $\theta'=\theta$ and using the fact that the $p(x|\theta')$ are normalized gives the desired result,

$$-\frac{\partial S^2(\theta',\theta)}{\partial\theta'^a\partial\theta'^b}\bigg|_{\theta'=\theta}=\int dx\,p(x|\theta)\frac{\partial\log p(x|\theta)}{\partial\theta^a}\frac{\partial\log p(x|\theta)}{\partial\theta^b}=g_{ab}\ . \tag{7.87}$$

7.5 Uniqueness of the information metric

The most remarkable fact about the information metric is that it is essentially unique: except for a 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 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 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

$$\Pr(A) = P^A = \sum\nolimits_{i \in A} p^i \ . \tag{7.88}$$

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.(7.88). 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

$$q_A^i = \begin{cases} \Pr(i|A) & \text{if} \quad i \in A \\ 0 & \text{if} \quad i \notin A \end{cases} \tag{7.89}$$

with

$$\sum\nolimits_i q_A^i = \sum\nolimits_{i \in A} \Pr(i|A) = 1 \ . \tag{7.90}$$

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

$$p^i = q_A^i P^A \ . \tag{7.91}$$

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,

$$\sum\nolimits_i p^i = \sum\nolimits_i q_A^i P^A = \sum\nolimits_A P^A \ . \tag{7.92}$$

Example: A coarse graining map G for the case of $R_3^+ \to R_2^+$ is

$$G(p^1, p^2, p^3) = (p^1, p^2 + p^3) = (P^1, P^2) \ . \tag{7.93}$$

One Markov map Q running in the opposite direction $R_2^+ \to R_3^+$ could be

$$Q(P^1,P^2) = (P^1, \frac{1}{3}P^2, \frac{2}{3}P^2) = (p^1,p^2,p^3) \ . \tag{7.94}$$

This particular map is defined by setting all $q_A^i = 0$ except $q_1^1 = 1$, $q_3^2 = 1/3$, and $q_2^3 = 2/3$.

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 $\{q_A^i\}$, 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},

$$Q(S_{M-1}) = S_{M-1} \ , \tag{7.95}$$

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 a larger and larger number of constraints. It could very well have happened that no Riemannian metric managed to survive such restrictive conditions; it is quite remarkable that some do and it is even more remarkable that (up to an uninteresting scale factor which amounts to a choice of the unit of distance) 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

$$\vec{V} = V^A \frac{\partial}{\partial P^A} = V^A \vec{E}_A \ , \tag{7.96}$$

where $\{\vec{E}_A\}$ is a coordinate basis. The inner product of two such vectors is

$$(\vec{V}, \vec{U})_M = g^{(M)}_{AB} V^A U^B \tag{7.97}$$

where the metric is

$$g^{(M)}_{AB} \overset{\text{def}}{=} (\vec{E}_A, \vec{E}_B)_M \ . \tag{7.98}$$

Similarly, vectors tangent to R^+_n are denoted by

$$\vec{v} = v^i \frac{\partial}{\partial p^i} = v^i \vec{e}_i \ , \tag{7.99}$$

and the inner product of two such vectors is

$$(\vec{v}, \vec{u})_n = g^{(n)}_{ij} v^i u^j \tag{7.100}$$

where

$$g^{(n)}_{ij} \overset{\text{def}}{=} (\vec{e}_i, \vec{e}_j)_n \ . \tag{7.101}$$

The images of vectors $\vec{V}$ tangent to R^+_m under Q are obtained from eq.(7.91)

$$Q_* \frac{\partial}{\partial P^A} = \frac{\partial p^i}{\partial P^A} \frac{\partial}{\partial p^i} = q^i_A \frac{\partial}{\partial p^i} \quad \text{or} \quad Q_* \vec{E}_A = q^i_A \vec{e}_i \ , \tag{7.102}$$

which leads to

$$Q_* \vec{V} = \vec{v} \quad \text{with} \quad v^i = q^i_A V^A \ . \tag{7.103}$$

Therefore, the invariance or isometry we want to impose is expressed as

$$(\vec{V}, \vec{U})_M = (Q_* \vec{V}, Q_* \vec{U})_n = (\vec{v}, \vec{u})_n \ . \tag{7.104}$$

The Theorem

Let $(\ , \ )_M$ 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

$$g^{(M)}_{AB} = (\bar{e}_A, \bar{e}_B)_M = \alpha(|P|) + |P| \beta(|P|) \frac{\delta_{AB}}{P^A} \ , \tag{7.105}$$

where $|P| \overset{\text{def}}{=} \sum_A P^A$, and α and β are smooth (C^∞) functions with $\beta > 0$ and $\alpha + \beta > 0$. The proof is given in the next section.

An important by-product of this theorem is that (7.105) has turned out to be the metric of a generic spherically symmetric space, eq.(7.66). In other words,

> *Invariance under Markovian embeddings implies that the statistical manifold of unnormalized probabilities is spherically symmetric.*

As we shall see in Chapters 10 and 11 this fact will turn out to be important in the derivation of quantum mechanics.

The metric above refers to the positive cone R^+_M but ultimately we are interested in the metric induced on the simplex $\mathcal{S}_{M-1}$ defined by $|P|=1$. In order to find the induced metric we first show that vectors that are tangent to the simplex $\mathcal{S}_{M-1}$ are such that

$$|V| \overset{\text{def}}{=} \sum_A V^A = 0\ . \tag{7.106}$$

Indeed, consider the derivative of any function $f=f(|P|)$ defined on R^+_M along the direction defined by $\vec{V}$,

$$0 = V^A \frac{\partial f}{\partial P^A} = V^A \frac{df}{d|P|}\frac{\partial |P|}{\partial P^A} = |V|\frac{df}{d|P|}\ , \tag{7.107}$$

where we used $\partial|P|/\partial P^A = 1$. Therefore $|V|=0$.

Next consider the inner product of any two vectors $\vec{V}$ and $\vec{U}$,

$$\begin{aligned}(\vec{V},\vec{U})_M &= \sum_{AB} V^A U^B \left(\alpha(|P|) + |P|\beta(|P|)\frac{\delta_{AB}}{P^A}\right)\\ &= \alpha(|P|)|V||U| + |P|\beta(|P|)\sum_A \frac{V^A U^A}{P^A}\ .\end{aligned} \tag{7.108}$$

For vectors tangent to the simplex S_{M-1} this simplifies to

$$(\vec{V},\vec{U})_M = \beta(1)\sum_A \frac{V^A U^A}{P^A}\ . \tag{7.109}$$

Therefore the choice of the function $\alpha(|P|)$ is irrelevant and the corresponding metric on $\mathcal{S}_{M-1}$ is determined up to a multiplicative constant $\beta(1)=\beta$

$$g_{AB} = \beta\frac{\delta_{AB}}{P^A}\ , \tag{7.110}$$

which is the information metric that was heuristically suggested earlier, eqs.(7.57) and (7.58). Indeed, transforming to a generic coordinate frame $P^A = P^A(\theta^1,\ldots,\theta^M)$ yields

$$d\ell^2 = g_{AB}\delta P^A \delta P^B = g_{ab}d\theta^a d\theta^b \tag{7.111}$$

with

$$g_{ab} = \beta{\textstyle\sum_A} P^A \frac{\partial \log P^A}{\partial\theta^a}\frac{\partial \log P^A}{\partial\theta^b}\ . \tag{7.112}$$

The Proof

The strategy is to consider special cases of Markov embeddings to determine what kind of constraints they 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

$$\begin{aligned}(p^1, \ldots p^a, \ldots p^b, \ldots p^M) &= Q'(P^1, \ldots P^M) \\ &= (P^1, \ldots P^b, \ldots P^a, \ldots P^m) \end{aligned} \tag{7.113}$$

The required q_A^i are

$$q_A^a = \delta_A^b\,, \quad q_A^b = \delta_A^a \quad \text{and} \quad q_A^i = \delta_A^i \quad \text{for} \quad A \neq a, b\ , \tag{7.114}$$

so that eq.(7.102), $Q'_* \vec{E}_A = q_A^i \vec{e}_i$, gives

$$Q'_* \vec{E}_a = \vec{e}_b\,, \quad Q'_* \vec{E}_b = \vec{e}_a \quad \text{and} \quad Q'_* \vec{E}_A = \vec{e}_A \quad \text{for} \quad A \neq a, b\ . \tag{7.115}$$

The invariance

$$(\vec{E}_A, \vec{E}_B)_M = (Q'_* \vec{E}_A, Q'_* \vec{E}_B)_M \tag{7.116}$$

yields,

$$\begin{aligned} & g_{aA}^{(M)}(P) = g_{bA}^{(M)}(p) \quad \text{and} \quad g_{bA}^{(M)}(P) = g_{aA}^{(M)}(p) \quad \text{for} \quad A \neq a, b & (7.117) \\ & g_{aa}^{(M)}(P) = g_{bb}^{(M)}(p) \quad \text{and} \quad g_{bb}^{(M)}(P) = g_{aa}^{(M)}(p) & (7.118) \\ & g_{AB}^{(M)}(P) = g_{AB}^{(M)}(p) \quad \text{for} \quad A, B \neq a, b\ . \end{aligned}$$

These conditions are useful for points along the line through 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 $p_c = Q'(P_c) = P_c$. Imposing eqs.(7.117) and (7.118) for all choices of the pairs (a, b) implies

$$\begin{aligned} g_{AA}^{(M)}(P_c) &= F_M(c) \\ g_{AB}^{(M)}(P_c) &= G_M(c) \quad \text{for} \quad A \neq B\ , \end{aligned} \tag{7.119}$$

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 $k \geq 2$

$$\begin{aligned} Q''(P^1, \ldots P^M) &= (p^1, \ldots p^{kM}) \\ &= (\underbrace{\frac{P^1}{k}, \ldots \frac{P^1}{k}}_{k \text{ times}}, \underbrace{\frac{P^2}{k}, \ldots \frac{P^2}{k}}_{k \text{ times}}, \ldots, \underbrace{\frac{P^M}{k}, \ldots \frac{P^M}{k}}_{k \text{ times}})\ . \end{aligned} \tag{7.120}$$

Q'' is implemented by choosing

$$q_A^i = \begin{cases} 1/k & \text{if} \quad i \in \{k(A-1)+1, \ldots kA\} \\ 0 & \text{if} \quad i \notin \{k(A-1)+1, \ldots kA\} \end{cases} \tag{7.121}$$

Under the action of Q'' vectors are transformed according to eq.(7.102),

$$Q''_* \vec{E}_A = q_A^i \vec{e}_i = \frac{1}{k} \left(\vec{e}_{k(A-1)+1} + \ldots + \ \vec{e}_{kA} \right) \tag{7.122}$$

so that the invariance

$$(\vec{E}_A, \vec{E}_B)_M = (Q''_* \vec{E}_A, Q''_* \vec{E}_B)_{kM} \tag{7.123}$$

yields,

$$g^{(M)}_{AB}(P) = \frac{1}{k^2} \sum_{i,\ j=k(A-1)+1}^{kA} g^{(kM)}_{ij}(p) \ . \tag{7.124}$$

Along the center lines, $P_c = (c/M, \ldots, c/M)$ and $p_c = (c/kM, \ldots, c/kM)$, equations (7.119) and (7.124) give

$$F_M(c) = \frac{1}{k} F_{kM}(c) + \frac{k-1}{k} G_{kM}(c) \tag{7.125}$$

and

$$G_M(c) = G_{kM}(c) \ . \tag{7.126}$$

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.(7.125) can be rewritten as

$$\frac{1}{M} \left[F_M(c) - \alpha(c) \right] = \frac{1}{kM} \left[F_{kM}(c) - \alpha(c) \right] = \beta(c) \ , \tag{7.127}$$

where $\beta(c)$ is a function independent of the integer M. Indeed, for any two integers M_1 and M_2 we have

$$\frac{1}{M_1} \left[F_{M_1}(c) - \alpha(c) \right] = \frac{1}{M_1 M_2} \left[F_{M_1 M_2}(c) - \alpha(c) \right] = \frac{1}{M_2} \left[F_{M_2}(c) - \alpha(c) \right] \ . \tag{7.128}$$

Therefore,

$$F_M(c) = \alpha(c) + M\beta(c) \ , \tag{7.129}$$

and for points along the center line,

$$g^{(M)}_{AB}(P_c) = \alpha(c) + M\beta(c)\delta_{AB} \ . \tag{7.130}$$

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

$$Q'''(P^1, \ldots P^M) = (\underbrace{\frac{P^1}{k_1}, \ldots \frac{P^1}{k_1}}_{k_1 \text{ times}}, \underbrace{\frac{P^2}{k_2}, \ldots \frac{P^2}{k_2}}_{k_2 \text{ times}}, \ldots, \underbrace{\frac{P^M}{k_M}, \ldots \frac{P^M}{k_M}}_{k_m \text{ times}}) \ . \tag{7.131}$$

Q''' is implemented by choosing

$$q^i_A = \begin{cases} 1/k_A & \text{if} \quad i \in \{(k_1 + \ldots + k_{A-1} + 1), (k_1 + \ldots + k_{A-1} + 2), \\ & \qquad\qquad \ldots, (k_1 + \ldots + k_A)\} \\ 0 & \text{if} \quad i \notin \{(k_1 + \ldots + k_{A-1} + 1), (k_1 + \ldots + k_{A-1} + 2), \\ & \qquad\qquad \ldots, (k_1 + \ldots + k_A)\} \end{cases} \tag{7.132}$$

Next note that any point P in R^+_M can be arbitrarily well approximated by points of the "rational" form

$$P = \left(\frac{ck_1}{n}, \frac{ck_2}{n}, \ldots \frac{ck_M}{n} \right) \ , \tag{7.133}$$

where the ks are positive integers and $\sum k_A = n$ and $|P| = c$. For these rational points the action of Q''' is

$$Q'''(P^1, \ldots P^M) = q^i_A P^A = \left(\frac{c}{n}, \frac{c}{n}, \ldots \frac{c}{n} \right) = p_c \tag{7.134}$$

which lies along the center line of R^+_n where the metric is known, eq.(7.130).

The action of Q''' on vectors, eq.(7.102), gives

$$Q'''_* \vec{E}_A = q^i_A \vec{e}_i = \frac{1}{k_A} \left(\vec{e}_{k_1 + \ldots + k_{A-1} + 1} + \ldots + \ \vec{e}_{k_1 + \ldots + k_A} \right) \ . \tag{7.135}$$

Using eq.(7.130) the invariance

$$(\vec{E}_A, \vec{E}_B)_M = (Q'''_* \vec{E}_A, Q'''_* \vec{E}_B)_n \tag{7.136}$$

yields, for $A = B$,

$$\begin{aligned} g^{(M)}_{AA}(P) &= \frac{1}{(k_A)^2} \sum_{i,\ j = k_1 + \ldots + k_{A-1} + 1}^{k_1 + \ldots + k_A} g^{(n)}_{ij}(p_c) \\ &= \frac{1}{(k_A)^2} \sum_{i,\ j = k_1 + \ldots + k_{A-1} + 1}^{k_1 + \ldots + k_A} [\alpha(c) + n\beta(c)\delta_{ij}] \\ &= \frac{1}{(k_A)^2} \left[(k_A)^2 \alpha(c) + k_A n \beta(c) \right] \\ &= \alpha(c) + \frac{n}{k_A} \beta(c) = \alpha(c) + \frac{c\beta(c)}{P^A} \ , \end{aligned} \tag{7.137}$$

where we used eq.(7.133), $P^A = ck_A/n$. Similarly, for $A \neq B$,

$$g^{(M)}_{AB}(P) = \frac{1}{k_A k_B} \sum_{i = k_1 + \ldots + k_{A-1} + 1}^{k_1 + \ldots + k_A} \ \sum_{j = k_1 + \ldots + k_{B-1} + 1}^{k_1 + \ldots + k_B} g^{(n)}_{ij}(p_c) \tag{7.138}$$

$$= \frac{1}{k_A k_B} k_A k_B \alpha(c) = \alpha(c) \ . \tag{7.139}$$

Therefore,

$$g_{AB}^{(M)} = \langle \vec{E}_A, \vec{E}_B \rangle_M = \alpha(c) + c\beta(c)\frac{\delta_{AB}}{P^A} \ , \tag{7.140}$$

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.(7.108),

$$(\vec{V}, \vec{V}) = \alpha|V|^2 + |P|\beta \sum_A \frac{\left(V^A\right)^2}{P^A}, \tag{7.141}$$

we see that for vectors with $|V| = 0$, $(\vec{V}, \vec{V}) \geq 0$ implies that $\beta > 0$, while for vectors with $V^A = KP^A$, where K is any constant we have

$$(\vec{V}, \vec{V}) = K^2|P|^2 (\alpha + \beta) > 0 \Rightarrow \alpha + \beta > 0 \ . \tag{7.142}$$

Conversely, we show that if these sign conditions are satisfied then $(\vec{V}, \vec{V}) \geq 0$ for all vectors. Using Cauchy's inequality,

$$\left(\sum_i x_i^2\right)\left(\sum_i y_i^2\right) \geq \left(\sum_i \|x_i y_i\|\right)^2 , \tag{7.143}$$

where $\|.\|$ denotes the modulus, we have

$$\left(\sum_A P^A\right)\left(\sum_B \frac{\left(V^B\right)^2}{P^B}\right) \geq \left(\sum_A \left\|V^A\right\|\right)^2 \geq \left(\sum_A V^A\right)^2 \ . \tag{7.144}$$

Therefore,

$$(\vec{V}, \vec{V}) = \alpha|V|^2 + |P|\beta \sum_A \frac{\left(V^A\right)^2}{P^A} \geq |V|^2(\alpha + \beta) \geq 0 \ , \tag{7.145}$$

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 (7.140). It remains to prove the converse, that this condition is sufficient. This is much easier. Indeed,

$$\begin{aligned}(Q_*\vec{E}_A, Q_*\vec{E}_B)_n &= q_A^i q_B^j (\bar{e}_i, \bar{e}_j)_n \\ &= \sum_{ij} q_A^i q_B^j \left[\alpha(|p|) + |p|\beta(|p|)\frac{\delta_{ij}}{p^i}\right] \ .\end{aligned} \tag{7.146}$$

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

$$(Q_*\vec{E}_A, Q_*\vec{E}_B)_n = \alpha(|P|) + |P|\beta(|P|)\sum_i \frac{q_A^i q_B^i}{p^i} \ . \tag{7.147}$$

Furthermore, since $q_A^i = 0$ unless $i \in A$,

$$\sum_i \frac{q_A^i q_B^i}{p^i} = \delta_{AB} \sum_i \frac{q_A^i}{P^A} = \frac{\delta_{AB}}{P^A} \ . \tag{7.148}$$

which finally leads to

$$(Q_* \vec{E}_A, Q_* \vec{E}_B)_n = \alpha(|P|) + |P|\beta(|P|)\frac{\delta_{AB}}{P^A} = (\bar{e}_A, \bar{e}_B)_M \tag{7.149}$$

which concludes the proof.

7.6 The metric for some common distributions

Multinomial distributions

The statistical manifold of multinomials,

$$P_N\left(n|\theta\right) = \frac{N!}{n_1! \ldots n_m!}\theta_1^{n_1} \ldots \theta_m^{n_m} \ , \tag{7.150}$$

where

$$n = (n_1 \ldots n_m) \quad \text{with} \quad \sum_{i=1}^{m} n_i = N \quad \text{and} \quad \sum_{i=1}^{m} \theta_i = 1 \, , \tag{7.151}$$

is the simplex $\mathcal{S}_{m-1}$. The metric is given by eq.(7.83),

$$g_{ij} = \sum_n P_N \frac{\partial \log P_N}{\partial \theta_i} \frac{\partial \log P_N}{\partial \theta_j} \quad \text{where} \quad 1 \leq i, j \leq m-1 \ . \tag{7.152}$$

The result is

$$g_{ij} = \left\langle (\frac{n_i}{\theta_i} - \frac{n_m}{\theta_m})(\frac{n_j}{\theta_j} - \frac{n_m}{\theta_m}) \right\rangle \ , \tag{7.153}$$

which, on computing the various correlations, gives

$$g_{ij} = \frac{N}{\theta_i}\delta_{ij} + \frac{N}{\theta_m} \quad \text{where} \quad 1 \leq i, j \leq m-1 \ . \tag{7.154}$$

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

$$d\ell^2 = \sum_{i,j=1}^{m-1} (\frac{N}{\theta_i}\delta_{ij} + \frac{N}{\theta_m}) d\theta_i d\theta_j \ . \tag{7.155}$$

Using

$$\sum_{i=1}^{m} \theta_i = 1 \implies \sum_{i=1}^{m} d\theta_i = 0 \ . \tag{7.156}$$

the second sum can be written as

$$\frac{N}{\theta_m}\sum_{i,j=1}^{m-1} d\theta_i \sum_{i,j=1}^{m-1} d\theta_j = \frac{N}{\theta_m}(d\theta_m)^2 \ . \tag{7.157}$$

Therefore,

$$d\ell^2 = \sum_{i,j=1}^{m} g_{ij} d\theta_i d\theta_j \quad \text{with} \quad g_{ij} = \frac{N}{\theta_i}\delta_{ij} \ . \tag{7.158}$$

Remark: As we saw in the previous section, eq.(7.112), 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 $\mathcal{S}_{m-1}^{(N)}$ is one which assigns equal probabilities to equal volumes,

$$P(\theta)d^{m-1}\theta \propto g^{1/2} d^{m-1}\theta \quad \text{with} \quad g = \frac{N^{m-1}}{\theta_1\theta_2\ldots\theta_m} \tag{7.159}$$

In the particular case of binomial distributions $m = 2$ with $\theta_1 = \theta$ and $\theta_2 = 1-\theta$ the results above become

$$g = g_{11} = \frac{N}{\theta(1-\theta)} \tag{7.160}$$

so that the uniform distribution over θ (with $0 < \theta < 1$) is

$$P(\theta)d\theta \propto d\ell = [\frac{N}{\theta(1-\theta)}]^{1/2} d\theta \ . \tag{7.161}$$

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 4.10 and 5.4) obtained by maximizing entropy $S[p,m]$ subject to n constraints $\langle f^a\rangle = F^a$ for $a = 1\ldots n$, plus normalization,

$$p(z|F) = \frac{1}{Z(\lambda)} m(z) e^{-\lambda_a f^a(z)} \quad \text{where} \quad Z(\lambda) = \int dz\, m(z) e^{-\lambda_a f^a(z)} \ . \tag{7.162}$$

The set of numbers $F = (F^1\ldots F^n)$ 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

$$\langle f^a\rangle = F^a = -\frac{\partial \log Z}{\partial \lambda_a} \ , \tag{7.163}$$

and it is straightforward to show that a further derivative with respect to λ_b yields the covariance matrix. Indeed,

$$\frac{\partial F^a}{\partial \lambda_b} = \frac{\partial}{\partial \lambda_b}(-\frac{1}{Z}\frac{\partial Z}{\partial \lambda_a}) = \frac{1}{Z^2}\frac{\partial Z}{\partial \lambda_a}\frac{\partial Z}{\partial \lambda_b} - \frac{1}{Z}\frac{\partial^2 Z}{\partial \lambda_a \partial \lambda_b} \tag{7.164}$$

$$= F^a F^b - \langle f^a f^b \rangle \ . \tag{7.165}$$

Therefore

$$C^{ab} \overset{\text{def}}{=} \langle (f^a - F^a)(f^b - F^b) \rangle = -\frac{\partial F^a}{\partial \lambda_b} \ . \tag{7.166}$$

Next, using the chain rule

$$\delta_a^c = \frac{\partial \lambda_a}{\partial \lambda_c} = \frac{\partial \lambda_a}{\partial F^b}\frac{\partial F^b}{\partial \lambda_c} \ , \tag{7.167}$$

we see that the matrix

$$C_{ab} = -\frac{\partial \lambda_a}{\partial F^b} \tag{7.168}$$

is the inverse of the covariance matrix,

$$C_{ab}C^{bc} = \delta_a^c \ ,$$

Furthermore, using eq.(4.92), we have

$$C_{ab} = -\frac{\partial^2 S(F)}{\partial F^a \partial F^b} \ . \tag{7.169}$$

The information metric is

$$g_{ab} = \int dz\, p(z|F)\, \frac{\partial \log p(z|F)}{\partial F^a}\, \frac{\partial \log p(z|F)}{\partial F^b}$$

$$= \frac{\partial \lambda_c}{\partial F^a}\frac{\partial \lambda_d}{\partial F^b}\int dz\, p\, \frac{\partial \log p}{\partial \lambda_c}\, \frac{\partial \log p}{\partial \lambda_d} \ . \tag{7.170}$$

Using eqs.(7.162) and (7.163),

$$\frac{\partial \log p(z|F)}{\partial \lambda_c} = F^c - f^c(z) \tag{7.171}$$

therefore,

$$g_{ab} = C_{ca}C_{db}C^{cd} \implies g_{ab} = C_{ab} \ , \tag{7.172}$$

so that the metric tensor g_{ab} is the inverse of the covariance matrix C^{ab} which, by eq.(7.169), is the Hessian of the entropy.

Instead of F^a we could use the Lagrange multipliers λ_a themselves as coordinates. Then the information metric is the covariance matrix,

$$g^{ab} = \int dz\, p(z|\lambda)\, \frac{\partial \log p(z|\lambda)}{\partial \lambda_a}\, \frac{\partial \log p(z|\lambda)}{\partial \lambda_b} = C^{ab} \ . \tag{7.173}$$

The distance $d\ell$ between neighboring distributions can then be written in either of two equivalent forms,

$$d\ell^2 = g_{ab}dF^a dF^b = g^{ab}d\lambda_a d\lambda_b \; . \tag{7.174}$$

The uniform distribution over the space of macrostates assigns equal probabilities to equal volumes,

$$P(F)d^nF \propto C^{1/2}d^nF \quad \text{or} \quad P'(\lambda)d^n\lambda \propto C^{-1/2}d^n\lambda \; , \tag{7.175}$$

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,

$$p(x|\mu, C) = \frac{c^{1/2}}{(2\pi)^{D/2}} \exp\left[-\frac{1}{2}C_{ij}(x^i - \mu^i)(x^j - \mu^j)\right] \; , \tag{7.176}$$

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, 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

$$d\ell^2 = g_{ij}d\mu^i d\mu^j + g_k^{ij}dC_{ij}d\mu^k + g^{ij\,kl}dC_{ij}dC_{kl} \; , \tag{7.177}$$

where

$$g_{ij} = C_{ij} \, , \quad g_k^{ij} = 0 \, , \quad \text{and} \quad g^{ij\,kl} = \frac{1}{4}(C^{ik}C^{jl} + C^{il}C^{jk}) \; , \tag{7.178}$$

where C^{ik} is the correlation matrix, that is, $C^{ik}C_{kj} = \delta^i_j$. Therefore,

$$d\ell^2 = C_{ij}d\mu^i d\mu^j + \frac{1}{2}C^{ik}C^{jl}dC_{ij}dC_{kl} \; . \tag{7.179}$$

To conclude we consider a few interesting 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,

$$d\ell^2 = C_{ij}d\mu^i d\mu^j \; , \tag{7.180}$$

which is an instance of eq.(7.172).

Another interesting special case is that of spherically symmetric Gaussians,

$$p(x|\mu,\sigma)=\frac{1}{(2\pi\sigma^2)^{D/2}}\exp\left[-\frac{1}{2\sigma^2}\delta_{ij}(x^i-\mu^i)(x^j-\mu^j)\right] \ . \tag{7.181}$$

The covariance matrix and its inverse are both diagonal and proportional to the unit matrix,

$$C_{ij}=\frac{1}{\sigma^2}\delta_{ij} \ , \quad C^{ij}=\sigma^2\delta^{ij} \ , \quad \text{and} \quad c=\sigma^{-2D} \ . \tag{7.182}$$

Using

$$dC_{ij}=d\frac{1}{\sigma^2}\delta_{ij}=-\frac{2\delta_{ij}}{\sigma^3}d\sigma \tag{7.183}$$

in eq.(7.179), the induced information metric is

$$d\ell^2=\frac{1}{\sigma^2}\delta_{ij}d\mu^i d\mu^j+\frac{1}{2}\sigma^4\delta^{ik}\delta^{jl}\frac{2\delta_{ij}}{\sigma^3}d\sigma\frac{2\delta_{kl}}{\sigma^3}d\sigma \tag{7.184}$$

which, using

$$\delta^{ik}\delta^{jl}\delta_{ij}\delta_{kl}=\delta^k_j\delta^j_k=\delta^k_k=D \ , \tag{7.185}$$

simplifies to

$$d\ell^2=\frac{\delta_{ij}}{\sigma^2}d\mu^i d\mu^j+\frac{2D}{\sigma^2}(d\sigma)^2 \ . \tag{7.186}$$

For Gaussians in one dimension, $D=1$, the statistical manifold is two dimensional with coordinates (μ,σ). The metric is

$$d\ell^2=\frac{1}{\sigma^2}(d\mu)^2+\frac{2}{\sigma^2}(d\sigma)^2 \ . \tag{7.187}$$

This space is a pseudo-sphere — a two-dimensional manifold of constant negative curvature.

7.7 Dimensionless distance?

There is one feature of the information distance $d\ell$ in eq.(7.47) that turns out to be very peculiar: $d\ell$ is dimensionless. (See [Caticha 2015b].) Indeed, we can easily verify from eq.(7.46) that if $d\theta^i$ has units of length, then $p(x|\theta)$, being a density, has units of inverse volume, and g_{ij} has units of inverse length squared. Distances are supposed to be measured in some units, perhaps of length; what sort of distance is this dimensionless $d\ell$?

A simple example will help clarify this issue. Consider two neighboring spherically symmetric Gaussian distributions, eq.(7.181) with some fixed variance. The distance between $p(x|\mu,\sigma)$ and $p(x|\mu+d\mu,\sigma)$ is given by eq.(7.186) with $d\sigma=0$,

$$d\ell^2=\frac{\delta_{ij}}{\sigma^2}d\mu^i d\mu^j \ . \tag{7.188}$$

This is the Euclidean metric δ_{ij} rescaled by σ^2. Therefore the dimensionless $d\ell$ represents a distance measured in units of the uncertainty σ.

The result is not restricted to Gaussian distributions. For example, for canonical distributions, eq.(7.166) shows that the information metric $g_{ab} = C_{ab}$ is an inverse measure of the fluctuations as given by the variance-covariance matrix C^{ab}. Therefore, *the information metric $g_{ab}(\theta)$ measures distinguishability in units of the local uncertainty implied by the distribution $p(x|\theta)$.*

A perhaps unexpected consequence of the notion of a dimensionless information distance is the following. As we saw in section (7.4.4) we can measure the length of a curve on a statistical manifold by "counting" the number of points along it. Counting points depends on a decision what we mean by a point and, in particular, on what we mean by two different points. If we agree that two points ought to be counted separately only when we can distinguish them, then one can assert that the number of distinguishable points in any finite segment is finite.

The same idea can used to measure areas and volumes: just count the number of distinguishable points within the region of interest. Remarkably this counting argument allows us to compare the sizes of regions of different dimensionality: it is meaningful to assert that a surface and a volume are of the same "information size" whenever they contain the same number of distinguishable points.

The conclusion is that statistical manifolds are peculiar objects that have properties that one would normally associate to a continuum but also have properties that one would normally only associate with discrete structures such as lattices.

Chapter 8

Entropy IV: Entropic Inference

There is one last issue that must be addressed before one can claim that the design of the inference method of maximum entropy (ME) is more or less complete. The goal is to rank probability distributions in order to select a distribution that, according to some desirability criteria, is preferred over all others. The ranking tool is entropy; higher entropy represents higher preference. But there is nothing in our 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 the quantity 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 joint distributions $p_J(x,\theta)$?

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_{ab} 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. 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 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 Chapter, 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 11, 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 Avoiding pitfalls – III

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.

8.3.1 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 (see fig.8.1). 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

$$S(\theta) = -\sum_i \theta_i \log \theta_i \tag{8.8}$$

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 might suggest that the three situations are epistemically identical — which they obviously are not. This is a source of concern since failing to see a distinction where one actually exists will inevitably give rise to 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 uncertainties in the vicinity of θ_C are all different. Indeed, the fact that the maximum of the entropy $S(\theta)$

[1] Other objections raised by these authors, such as the compatibility of Bayesian and entropic methods, have been addressed elsewhere in these lectures.

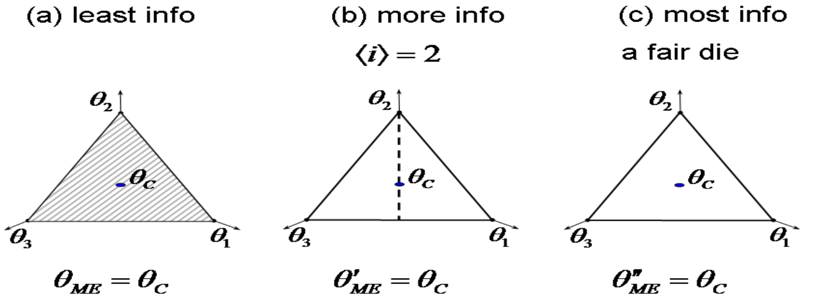

Figure 8.1: Three different states of information concerning a three-sided die. **(a)** Absolute ignorance: the distribution assigned by MaxEnt is $\theta_C = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. **(b)** We know that $r = \langle i \rangle = 2$: the MaxEnt distribution is also θ_C. **(c)** The die 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 uncertainties about θ_C are different.

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),

$$P_a(\theta)d\theta_1 d\theta_2 \propto \; e^{S(\theta)} g^{1/2}(\theta) d\theta_1 d\theta_2 \quad \text{with} \quad g(\theta) = \frac{1}{\theta_1\theta_2(1-\theta_1-\theta_2)} . \tag{8.9}$$

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),

$$P_b(\theta_2)d\theta_2 \propto \; e^{S(\theta)} g^{1/2}(\theta_2) d\theta_2 \quad \text{with} \quad g(\theta_2) = \frac{1}{\theta_2(1-\theta_2)} . \tag{8.10}$$

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.

Finally, in case (c) the distribution is concentrated at the single central point θ_C,

$$P_c(\theta) = \delta(\theta - \theta_C) \; , \tag{8.11}$$

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 variables we are being ignorant about.

8.3.2 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.

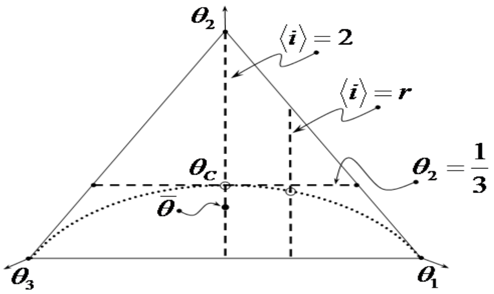

Figure 8.2: The MaxEnt solution for the constraint $\langle i \rangle = r$ for different values of r leads to the dotted line. If r is unknown averaging over r should lead to the distribution at the point θ marked by $\bar{\theta}$.

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 $\theta_{ME}(r)$.

[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.

But r is itself unknown so a more honest θ assignment is an average over r,

$$\bar{\theta} = \int dr\, p(r)\theta_{ME}(r) \ , \tag{8.12}$$

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,

$$f(1) = 2, \quad f(2) = 3, \quad \text{and} \quad f(3) = 1 \ , \tag{8.13}$$

then repeating the analysis above would lead us to conclude that $\bar{\theta}_1 < 1/3$, which represents a die biased against $i = 1$. 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 4.11 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.(8.12) 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 nothing about a die is not the same as knowing that the die is biased against a particular phase but not knowing by how much.*

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,

$$p(i|\beta)=\frac{e^{-\beta\varepsilon_i}}{Z(\beta)}\quad\text{where}\quad Z(\beta)=\sum_i e^{-\beta\varepsilon_i}\ . \tag{8.14}$$

Since the inverse temperature $\beta=\beta(E)$ is itself unknown we must average over β,

$$p_t(i)=\int d\beta\, p(\beta)p(i|\beta)\ . \tag{8.15}$$

To the extent that both distributions reflect complete ignorance we must have

$$p_u(i)=p_t(i) \qquad\qquad \text{(wrong!)}$$

which can only happen provided

$$p(\beta)=\delta(\beta)\quad\text{or}\quad\beta=0\ . \tag{8.16}$$

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 knowing that it is in thermal equilibrium at a temperature that happens not be unknown.*

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

$$p(i) = \int d\beta \, p(i,\beta) \quad \text{where} \quad p(i,\beta) = p(\beta)p(i|\beta) \ . \tag{8.17}$$

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

Topics in Statistical Mechanics*

9.1 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

$$p(z|F) = \frac{1}{Z(\lambda)}\, m(z)\, e^{-\lambda_k f^k(z)} \quad \text{with} \quad Z(\lambda) = \textstyle\int dz\, m(z)\, e^{-\lambda_k f^k(z)} \ , \tag{9.1}$$

and the corresponding entropy is

$$S(F) = \log Z(\lambda) + \lambda_k F^k \ . \tag{9.2}$$

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.(4.92), 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

$$p_0(z, z') = \frac{1}{Z_0}\, m(z) m'(z')\, e^{-\lambda_{0\alpha}\left(f^k(z) + f'^k(z')\right)}. \tag{9.3}$$

But distributions of lower entropy are not totally ruled out; to explore the possibility that the quantities F^k_T 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

$$p_J(z, z', F) = p(F) p(z|F) p(z'|F_T - F), \tag{9.4}$$

where $p(z|F)$ is the canonical distribution in eq.(9.1), its entropy is eq.(9.2) 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

$$P(F)dF = \frac{1}{\zeta}\, e^{S_T(F)} g^{1/2}(F) dF, \tag{9.5}$$

where ζ is a normalization constant and the entropy $S_T(F)$ of the system plus the bath is

$$S_T(F) = S(F) + S'(F_T - F). \tag{9.6}$$

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

$$S'(F_T - F) = \log Z'(\lambda_0) + \lambda_{0k}\left(F^k_T - F^k\right) = \text{const} - \lambda_{0k} F^k. \tag{9.7}$$

It remains to calculate the determinant $g(F)$ of the information metric given by eq.(7.87),

$$g_{ij} = -\frac{\partial^2 S_T(\dot{F}, F)}{\partial \dot{F}^i \partial \dot{F}^j} = -\frac{\partial^2}{\partial \dot{F}^i \partial \dot{F}^j}\left[S(\dot{F}, F) + S'(F_T - \dot{F}, F_T - F)\right] \tag{9.8}$$

where the dot indicates that the derivatives act on the first argument. The first term on the right is

$$\begin{aligned}\frac{\partial^2 S(\dot{F}, F)}{\partial \dot{F}^i \partial \dot{F}^j} &= -\frac{\partial^2}{\partial \dot{F}^i \partial \dot{F}^j}\int dz\, p(z|\dot{F}) \log \frac{p(z|\dot{F})}{m(z)}\frac{m(z)}{p(z|F)} \\ &= \frac{\partial^2 S(F)}{\partial F^i \partial F^j} + \int dz\, \frac{\partial^2 p(z|F)}{\partial F^i \partial F^j} \log \frac{p(z|F)}{m(z)} \ .\end{aligned} \tag{9.9}$$

To calculate the integral on the right use eq.(9.1) written in the form

$$\log \frac{p(z|F)}{m(z)} = -\log Z(\lambda) - \lambda_k f^k(z) \ , \tag{9.10}$$

so that the integral vanishes,

$$-\log Z(\lambda)\frac{\partial^2}{\partial F^i \partial F^j}\int dz\, p(z|F)\ - \lambda_k \frac{\partial^2}{\partial F^i \partial F^j}\int dz\, p(z|F) f^k(z) = 0 \ . \tag{9.11}$$

Similarly,

$$\begin{aligned}\frac{\partial^2}{\partial \dot{F}^i \partial \dot{F}^j} S'(F_T - \dot{F}, F_T - F) &= \frac{\partial^2 S'(F_T - F)}{\partial F^i \partial F^j} \\ &\quad + \int dz'\, \frac{\partial^2 p(z'|F_T - F)}{\partial F^i \partial F^j} \log \frac{p(z'|F_T - F)}{m'(z')}\end{aligned} \tag{9.12}$$

and here, using eq.(9.7), both terms vanish. Therefore

$$g_{ij} = -\frac{\partial^2 S(F)}{\partial F^i \partial F^j} \ . \tag{9.13}$$

We conclude that the probability that the value of F fluctuates into a small volume $g^{1/2}(F)dF$ becomes

$$p(F)dF = \frac{1}{\zeta}\, e^{S(F) - \lambda_{0k} F^k} g^{1/2}(F) dF \ . \tag{9.14}$$

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.(9.4) 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

$$\left\langle \delta\lambda_i \delta f^j \right\rangle = \left\langle \delta\lambda_i \delta F^j \right\rangle = -\frac{\partial \langle\lambda_i\rangle}{\partial\lambda_{0j}} + (\lambda_{0i} - \langle\lambda_i\rangle)\left(F_0^j - \left\langle F^j \right\rangle\right). \qquad (9.15)$$

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,

$$\left\langle \delta\lambda_i \delta f^j \right\rangle \approx -\delta_i^j \ . \qquad (9.16)$$

9.2 Variational approximation methods – I*

9.2.1 Mean field Theory*

9.2.2 Classical density functional theory*

[Yousefi Caticha 2021]

Chapter 10

A Prelude to Dynamics: Kinematics

In this and the following chapters our main concern will be to deploy the concepts of probability, entropy, and information geometry to formulate quantum mechanics (QM) as a dynamical model that describes the evolution of probabilities in time. The fact that the dynamical variables are probability distributions turns out to be highly significant because all changes of probabilities — including their time evolution — must be compatible with the basic principles for updating probabilities. In other words, the kinds of dynamics we seek are those driven by the maximization of an entropy subject to constraints that carry the information that is relevant to the particular system at hand.

The goal of *entropic dynamics* is to generate a trajectory in a space of probability distributions. As we saw in Chapter 7, these spaces are statistical manifolds and have an intrinsic metric structure given by the information metric. Furthermore, our interest in trajectories naturally leads us to consider both their tangent vectors and their dual covectors because it is these objects that will be used to represent velocities and momenta respectively. It turns out that, just as the statistical manifold has a natural metric structure, the statistical manifold plus all the spaces of tangent covectors is itself a manifold — the cotangent bundle — that can be endowed with a natural structure called *symplectic*.[1] Our goal will be to formulate an entropic dynamics that naturally reflects these metric and symplectic structures.

But not every curve is a trajectory and not every parameter that labels points along a curve is time. In order to develop a true dynamics we will have to construct a concept of time — a problem that will be addressed in the next chapter. This chapter is devoted to kinematics. As a prelude to a true dynamics

[1] The term symplectic was invented by Weyl in 1946. The old name for the family of groups of transformations that preserve certain antisymmetric bilinear forms had been *complex* groups. Since the term was already in use for complex variables, Weyl thought this was unnecessarily confusing. So he invented a new term by literally translating 'complex' from its Latin roots *com-plexus*, which means "together-braided," to its Greek roots $\sigma\upsilon\mu$-$\pi\lambda\epsilon\kappa\tau\iota\kappa\acute{o}\varsigma$.

we shall develop some of the tools needed to study families of curves that are closely associated with the symplectic and metric structures.[2]

To simplify the discussion in this chapter we shall consider the special case of a statistical manifold of finite dimension. Back in Chapter 7 we studied the information geometry of the manifold associated with a parametric family of probability distributions. The uncertain variable x can be either discrete or continuous and the distributions $\rho_\theta(x) = \rho(x|\theta)$ are labeled by parameters θ^i $(i = 1 \ldots n)$ which will be used as coordinates on the manifold.[3] First, to introduce the main ideas, we shall consider the simpler example in which the θ^i are generic parameters of no particular significance. Then, we shall address the example of a simplex — a statistical manifold for which the uncertain variables are discrete, $x = i = 1 \ldots n$, and the probabilities themselves are used as coordinates, $\theta^i = \rho(i)$. The result is a formalism in which the linearity of the evolution equations, the emergence of a complex structure, Hilbert spaces, and a Born rule, are derived rather than postulated.

10.1 Gradients and covectors

Just as we chose displacements as the prototype vectors, we choose the prototype covectors (also known as *covariant vectors* and as *1-forms*) to be the gradients of functions. To see how this comes about we note that the derivative $df/d\lambda$ of the function $f(\theta)$ along the curve $\theta^i = \theta^i(\lambda)$ parametrized by λ can be analyzed in two ways.

The first way consists of interpreting $df/d\lambda$ as the action of an operator $\bar{V}$ on f. Recall from Section 7.2 that the vector $\bar{V}$ tangent to the curve $\theta^i(\lambda)$ can be "identified" with a directional derivative. Indeed, if $f(\theta)$ is a scalar function, then the derivative along the curve is

$$\frac{df}{d\lambda} = V^i \frac{\partial f}{\partial \theta^i} \quad \text{where} \quad V^i(\theta) = \frac{d\theta^i}{d\lambda} \ . \tag{10.1}$$

Since there is a strict 1-1 correspondence between

$$\frac{d}{d\lambda} = V^i \frac{\partial}{\partial \theta^i} \quad \text{and} \quad \bar{V} = V^i \bar{e}_i \ , \tag{10.2}$$

we shall define the action of $\bar{V}$ on f by

$$\bar{V}(f) \overset{\text{def}}{=} \frac{d}{d\lambda} f = (V^i \partial_i) f \ , \tag{10.3}$$

[2] The material of this chapter is adapted from [Caticha 2019, 2021b] which itself builds and expands on previous work on the geometric and symplectic structure of quantum mechanics [Kibble 1979; Heslot 1985; Anandan and Aharonov 1990; Cirelli et al. 1990; Abe 1992; Hughston 1995; Ashekar and Schilling 1998; de Gosson, Hiley 2011; Elze 2012; Reginatto and Hall 2011, 2012].

[3] We will continue to adopt the standard notation of using upper indices to label coordinates and components of vectors (*e.g.* θ^i and $\vec{V} = V^i \vec{e}_i$) and lower indices to denote components of covectors (e.g. $\partial F/\partial \theta^i = \partial_i F$). We also adopt the Einstein summation convention: a sum over an index is understood whenever it appears repeated as an upper and a lower index.

so that

$$\frac{d}{d\lambda} = \bar{V} \ . \tag{10.4}$$

The vectors

$$\bar{e}_i = \frac{\partial}{\partial\theta^i} = \partial_i \tag{10.5}$$

constitute the "coordinate" basis — a basis of vectors that is adapted to the coordinate grid in the sense that the vectors $\{\bar{e}_i\}$ are tangent to the grid lines. More explicitly, the vector $\bar{e}_i$ is tangent to the coordinate curve defined by holding constant all θ^js with $j \neq i$, and using θ^i as the parameter along the curve.

The second way to think about the directional derivative $df/d\lambda$ is to write

$$\nabla f[\bar{V}] \overset{\text{def}}{=} (\partial_i f)V^i = \frac{d}{d\lambda} f \tag{10.6}$$

and interpret $df/d\lambda$ as the scalar that results from the action of the linear functional ∇f on the vector $\bar{V}$. Indeed, using linearity the action of ∇f on the vector $\bar{V}$ is

$$\nabla f[\bar{V}] = V^i\,\nabla f[\bar{e}_i] \quad \text{so that} \quad \nabla f[\bar{e}_i] = \partial_i f \ . \tag{10.7}$$

When the function f is one of the coordinates, $f(\theta) = \theta^j$, we obtain

$$\nabla\theta^j[\bar{e}_i] = \frac{\partial\theta^j}{\partial\theta^i} = \delta^j_i \ . \tag{10.8}$$

Furthermore, using the chain rule

$$\nabla f(\theta) = \frac{\partial f}{\partial\theta^i}\nabla\theta^i = \partial_i f\,\nabla\theta^i \ , \tag{10.9}$$

we see that $\{\partial_i f\}$ are the components of the covector ∇f, and that $\{\nabla\theta^i\}$ constitute a covector basis which is *dual* or *reciprocal* to the vector basis $\{\bar{e}_i\}$.

The transformation of vector components and of basis vectors under a change of coordinates (see eqs.(7.11) and (7.17)) is such that the vectors $\bar{V} = V^i\bar{e}_i$ are invariant. Generic covectors $\omega = \omega_i\nabla\theta^i$ can also be defined as invariant objects whose components ω_i transform as $\partial_i f$. Using the chain rule the transformation to primed coordinates, $\theta^i \to \theta^{i'}$, is

$$\frac{\partial f}{\partial\theta^{i'}} = \frac{\partial\theta^j}{\partial\theta^{i'}}\frac{\partial f}{\partial\theta^j} \quad \text{or} \quad \partial_{i'} f = \frac{\partial\theta^j}{\partial\theta^{i'}}\partial_j f \ . \tag{10.10}$$

Using

$$\nabla\theta^{i'} = \frac{\partial\theta^{i'}}{\partial\theta^j}\nabla\theta^j \quad \text{and} \quad \omega_{i'} = \frac{\partial\theta^j}{\partial\theta^{i'}}\omega_j \tag{10.11}$$

we can check that ω is indeed invariant,

$$\omega = \omega_i\nabla\theta^i = \omega_{i'}\nabla\theta^{i'} \ . \tag{10.12}$$

Alternatively, we can define generic covectors as linear functionals of vectors. Indeed, using linearity and (10.8) the action of ω on the vector $\bar{V}$,

$$\omega[\bar{V}] = \omega_i\nabla\theta^i[\bar{V}] = \omega_i V^j\,\nabla\theta^i[\bar{e}_j] = \omega_i V^i \ , \tag{10.13}$$

is invariant.

10.2 Lie derivatives

The task of defining the derivative of a vector field (and more generally of tensor fields) in a curved manifold amounts to comparing vectors in two neighboring tangent spaces (see e.g., [Schutz 1980]). The problem is to provide a criterion to decide which vector in one tangent space is considered as being the "same" as another vector in a neighboring and therefore *different* tangent space. Such a criterion would allow us to give meaning to the statement that a particular vector field is "constant" or has a "vanishing derivative." Solving this problem requires introducing additional structure. One solution is to introduce a *connection* field and this leads to the concept of a covariant derivative. Another solution, due to Sophus Lie, is to define the derivative relative to a "reference" *vector* field. This approach leads to the concept of the Lie derivative — the derivative of a tensor field with respect to a vector field. (See e.g., [Schutz 1980].) Lie derivatives are introduced as follows.

Since vectors are defined as tangents to curves, if we are given a space-filling congruence of curves, then we can define the associated vector *field*: to every point θ we associate the vector $\bar{V}(\theta)$ that happens to be tangent to the particular curve that passes through θ. Conversely, if we are given a vector field, then we can define the corresponding congruence of curves $\theta^i = \theta^i(\lambda)$ that are tangent to $\bar{V}(\theta)$ at every point θ.

The vector field $\bar{V}(\theta)$ can be used to define a diffeomorphism V_λ the action of which is to map the point θ to the point θ_λ displaced by a parameter "distance" λ along the congruence,

$$\text{if} \quad \theta = \theta(\lambda_0) \quad \text{then} \quad V_\lambda(\theta) = \theta(\lambda_0 + \lambda) = \theta_\lambda \ . \tag{10.14}$$

We shall assume that the map V_λ is sufficiently smooth and invertible.

To define the Lie derivative of a scalar function $f(\theta)$ along (the congruence defined by) the field $\bar{V}(\theta)$ we first introduce the notion of Lie-dragging. Given the function f define a new function f_λ called the *pull-back* of f under the action of the map V_λ,

$$f_\lambda(\theta) = f(\theta_\lambda) \ . \tag{10.15}$$

Then the Lie derivative of f along $\bar{V}$ is defined by

$$\pounds_V f \stackrel{\text{def}}{=} \lim_{\lambda \to 0} \frac{1}{\lambda}[f_\lambda(\theta) - f(\theta)] \ . \tag{10.16}$$

The important point here is that both functions f and f_λ are evaluated at the same point θ. The idea is that when Lie-dragging is applied to a vector field, its Lie derivative would involve subtracting vectors located at the *same* tangent space. As $\lambda \to 0$,

$$\begin{aligned} f_\lambda(\theta) &= f(\theta_\lambda) = f(\theta^i + \frac{d\theta^i}{d\lambda}\lambda) \\ &= f(\theta) + \frac{\partial f}{\partial \theta^i}\frac{d\theta^i}{d\lambda}\lambda = f(\theta) + \lambda\frac{df}{d\lambda} \end{aligned} \tag{10.17}$$

so that

$$\pounds_V f = \frac{df}{d\lambda} = \bar{V}[f] \ . \tag{10.18}$$

This result is not particularly surprising: *the Lie derivative of f along $\bar{V}$ is just the derivative of f along $\bar{V}$.* It gets more interesting when we apply the same idea to the Lie derivative of a vector field $\bar{U}$ along the congruence defined by $\bar{V}$. We note, in particular, that the Lie derivative of a scalar function is a scalar function too or, in other words, the Lie derivative is a scalar differential operator and, therefore, its action on a vector $\bar{U}$ yields a vector, $\pounds_V \bar{U}$, that can itself act on functions, $\left(\pounds_V \bar{U}\right)[f]$, to yield other scalar functions.

The definition of the Lie derivative can be extended to vectors and tensors by imposing the natural additional requirement that the Lie derivative be a derivative, that is, it must obey a Leibniz rule. For example, the Lie derivative $\pounds_V \bar{U}$ of a vector $\bar{U}$ is *defined* so that for any scalar function f the Lie derivatives satisfy

$$\pounds_V \left(\bar{U}[f]\right) \overset{\text{def}}{=} \left(\pounds_V \bar{U}\right)[f] + \bar{U}[\pounds_V f] \ , \tag{10.19}$$

while the derivative $\pounds_V T$ of a generic tensor T acting on a collection $a, b, ...$ of vectors or covectors satisfies

$$\pounds_V \left[T(a,b,...)\right] \overset{\text{def}}{=} \left[\pounds_V T\right](a,b,...) + T(\pounds_V a, b, ...) + T(a, \pounds_V b, ...) + ... \tag{10.20}$$

Next we compute the Lie derivatives of vectors, covectors, and tensors in terms of their components.

10.2.1 Lie derivative of vectors

We wish to calculate the Lie derivative of $\bar{U}$ along $\bar{V}$. First we note that if $f = f(\theta)$, then

$$\bar{U}[f] = \frac{df}{d\mu} = \frac{\partial f}{\partial \theta^i}\frac{d\theta^i}{d\mu} \tag{10.21}$$

is just a scalar function of θ so that

$$\pounds_V \left(\bar{U}[f]\right) = \bar{V}\left(\bar{U}[f]\right) \ . \tag{10.22}$$

On the other hand, imposing the Leibniz rule (10.19), gives

$$\pounds_V \left(\bar{U}[f]\right) = \left(\pounds_V \bar{U}\right)[f] + \bar{U}\bar{V}[f] \ . \tag{10.23}$$

Therefore,

$$\pounds_V \bar{U} = \bar{V}\bar{U} - \bar{U}\bar{V} \overset{\text{def}}{=} [\bar{V}, \bar{U}] \ . \tag{10.24}$$

where we introduced the Lie bracket notation on the right. Since the Lie bracket is antisymmetric, so is the Lie derivative,

$$\pounds_V \bar{U} = -\pounds_U \bar{V} \ . \tag{10.25}$$

To evaluate the Lie derivative in terms of components one calculates the derivatives in (10.24),

$$\bar{V}\bar{U}[f] = V^i\partial_i\left(U^j\partial_j f\right) = V^i\partial_i U^j\partial_j f + V^i U^j\partial_i\partial_j f \ , \tag{10.26}$$

$$\bar{U}\bar{V}[f] = U^i\partial_i\left(V^j\partial_j f\right) = U^i\partial_i V^j\partial_j f + U^i V^j\partial_i\partial_j f \ . \tag{10.27}$$

The result is

$$\pounds_V\bar{U} = [\bar{V},\bar{U}] = \left(V^i\partial_i U^j - U^i\partial_i V^j\right)\partial_j \ , \tag{10.28}$$

which explicitly shows that $\pounds_V\bar{U}$ is a vector with components

$$(\pounds_V\bar{U})^j = [\bar{V},\bar{U}]^j = V^i\partial_i U^j - U^i\partial_i V^j \ . \tag{10.29}$$

Side remark: Equation (10.29) shows an important difference between the Lie derivative $\pounds_V\bar{U}$ and the covariant derivative $\nabla_V\bar{U}$. The latter depends on $\bar{V}(\theta)$ only at the point θ. Indeed, if $f(\theta)$ is some scalar function, then $\nabla_{fV}\bar{U} = f\nabla_V\bar{U}$. In contrast, $\pounds_V\bar{U}$ also depends on the derivatives of $\bar{V}(\theta)$ at θ.

10.2.2 Lie derivative of covectors

To calculate the Lie derivative $\pounds_V\omega$ of ω along the congruence defined by $\bar{V}$ we first recall that $\pounds_V$ is a scalar operator so that $\pounds_V\omega$ is itself a covector,

$$\pounds_V\omega = [\pounds_V\omega]_i\nabla\theta^i \ . \tag{10.30}$$

Next we consider a generic vector field $\bar{U}$ and use the fact that $\omega(\bar{U}) = \omega_i U^i$ is a scalar function. The Lie derivative of the right hand side is

$$\begin{aligned}\pounds_V(\omega_i U^i) = \bar{V}(\omega_i U^i) &= V^j\partial_j(\omega_i U^j)\\ &= V^j(\partial_j\omega_i)U^i + V^j\omega_i(\partial_j U^i) \ .\end{aligned} \tag{10.31}$$

Since $\pounds_V$ obeys the Leibniz rule, the Lie derivative of the left hand side can also be written as

$$\pounds_V[\omega(\bar{U})] = [\pounds_V\omega](\bar{U}) + \omega(\pounds_V\bar{U}) = [\pounds_V\omega]_i U^i + \omega_i(\pounds_V\bar{U})^i \ . \tag{10.32}$$

Equating (10.31) and (10.32), and using (10.29) we get

$$V^j(\partial_j\omega_i)U^i + V^j\omega_i(\partial_j U^i) = [\pounds_V\omega]_i U^i + \omega_i(V^j\partial_j U^i - U^j\partial_j V^i) \ , \tag{10.33}$$

$$V^j(\partial_j\omega_i)U^i = [\pounds_V\omega]_i U^i - \omega_j U^i\partial_i V^j \ . \tag{10.34}$$

Since this must hold for any arbitrary vector U^i we get

$$(\pounds_V\omega)_i = V^j\partial_j\omega_i + \omega_j\partial_i V^j \ . \tag{10.35}$$

10.2.3 Lie derivative of the metric

To find $\pounds_V G$ where G is the metric tensor we deploy the same trick as in the previous sections: use the fact that the action of G on two generic vector fields $\bar{A}$ and $\bar{B}$, $G(\bar{A},\bar{B}) = G_{ij}A^iB^j$, is a scalar function and take the Lie derivative of both sides. The right hand side gives

$$\pounds_V(G_{ij}A^iB^j) = V^k\partial_k(G_{ij}A^iB^j) \tag{10.36}$$

and, using the Leibniz rule on the left hand side, we get

$$\begin{aligned}\pounds_V[G(\bar{A},\bar{B})] &= [\pounds_V G](\bar{A},\bar{B})] + G(\pounds_V\bar{A},\bar{B}) + G(\bar{A},\pounds_V\bar{B}) \\ &= [\pounds_V G]_{ij}A^iB^j + G_{ij}(\pounds_V\bar{A})^iB^j + G_{ij}A^i(\pounds_V\bar{B})^j \ .\end{aligned} \tag{10.37}$$

Setting (10.36) equal to (10.37) and using (10.29) leads to the desired expression,

$$[\pounds_V G]_{ij} = V^k\partial_k G_{ij} + G_{ik}\partial_j V^k + G_{kj}\partial_i V^k \ . \tag{10.38}$$

Notice that in the derivation above we have not used the symmetry or any other properties of G beyond the fact that $G(\cdot,\cdot)$ is a tensor so that $G(\bar{A},\bar{B})$ is a scalar function. This means that the expression (10.38) gives the Lie derivative of any covariant tensor with components T_{ij},

$$[\pounds_V T]_{ij} = V^k\partial_k T_{ij} + T_{ik}\partial_j V^k + T_{kj}\partial_i V^k \ . \tag{10.39}$$

10.3 The cotangent bundle

The construction of an entropic dynamics in the next chapter requires that we deal with several distinct spaces. One is the space of microstates — the variables that we are trying to predict. This space will be called the *ontic configuration space.*[4] In this chapter we deal with discrete microstates labelled by $x = 1\ldots N$ such as might describe an N-sided (possibly "quantum") die. A second space of interest is the statistical manifold of normalized distributions,

$$\mathcal{P} = \left\{\rho(x|\theta)|\ \textstyle\sum_{x=1}^N \rho(x|\theta) = 1\right\} \tag{10.40}$$

labelled by coordinates θ^i, $i = 1\ldots n$. This space will be called the *epistemic configuration space* which we abbreviate to the *e-configuration space*. The n-dimensional space $\mathcal{P}$ is a subspace of the $(N-1)$-dimensional simplex of normalized distributions $\mathcal{S} = \{\rho(x)|\sum_{x=1}^N \rho(x) = 1\}$.

[4] These ontic variables are meant to represent something real only within the context of a particular model. One may consider models where the positions of particles are assumed ontic. In other models one might assume that it is the field variables that are ontic, while the particles are quantum excitations of the fields. It is even possible to conceive of hybrid models in which some ontic variables represent the particles we call "matter" (e.g., the fermions) while some other ontic variables represent the gauge fields we call "forces" (e.g., the electromagnetic field).

Given any manifold such as $\mathcal{P}$ we can construct two other manifolds that turn out to be useful. One of these manifolds is denoted by $T\mathcal{P}$ and is called the *tangent bundle.* The idea is the following. Consider all the curves passing through a point $\theta = (\theta^1 \dots \theta^n)$. The set of all the vectors that are tangent to those curves is a vector space called the tangent space at θ and is denoted $T\mathcal{P}_\theta$. The space $T\mathcal{P}$ composed of $\mathcal{P}$ plus all its tangent spaces $T\mathcal{P}_\theta$ turns out to be a manifold of a special type generically called a *fiber bundle*; $\mathcal{P}$ is called the *base manifold* and $T\mathcal{P}_\theta$ is called the *fiber* at the point θ. Thus, $T\mathcal{P}$ is appropriately called the *tangent bundle.*

We can also consider the space of all covectors at a point θ. Such a space is denoted $T^*\mathcal{P}_\theta$ and is called the cotangent space at θ. The second special manifold we can construct is the fiber bundle composed of $\mathcal{P}$ plus all its cotangent spaces $T^*\mathcal{P}_\theta$. This fiber bundle is denoted by $T^*\mathcal{P}$ and is called the *cotangent bundle.*

The reason we care about vectors and covectors is that these are the objects that will eventually be used to represent velocities and momenta. Indeed, if $\mathcal{P}$ is the e-configuration space, its associated cotangent bundle $T^*\mathcal{P}$, which we will call the *e-phase space*, will play a central role. But that is for later; for now all we need is that the tangent and cotangent bundles are geometric objects that are always available to us independently of any physical considerations.

10.3.1 Vectors, covectors, etc.

A point $X \in T^*\mathcal{P}$ will be represented as $X = (\theta, \phi)$, where $\theta = (\theta^1 \dots \theta^n)$ are coordinates on the base manifold $\mathcal{P}$ and $\phi = (\phi_1 \dots \phi_n)$ are some generic coordinates on the space $T^*\mathcal{P}_\theta$ that is cotangent to $\mathcal{P}$ at the point θ. Curves on $T^*\mathcal{P}$ allow us to define vectors on the tangent spaces $T(T^*\mathcal{P})_X$. Let $X = X(\lambda)$ be a curve parametrized by λ, then the vector $\bar{V}$ tangent to the curve at $X = (\theta, \phi)$ has components $d\theta^i/d\lambda$ and $d\phi_i/d\lambda$, and is written

$$\bar{V} = \frac{d}{d\lambda} = \frac{d\theta^i}{d\lambda}\frac{\partial}{\partial\theta^i} + \frac{d\phi_i}{d\lambda}\frac{\partial}{\partial\phi_i} \ , \tag{10.41}$$

where $\partial/\partial\theta^i$ and $\partial/\partial\phi_i$ are the basis vectors and the index $i = 1 \dots n$ is summed over. The directional derivative of a function $F(X)$ along the curve $X(\lambda)$ is

$$\frac{dF}{d\lambda} = \frac{\partial F}{\partial\theta^i}\frac{d\theta^i}{d\lambda} + \frac{\partial F}{\partial\phi_i}\frac{d\phi_i}{d\lambda} = \tilde{\nabla}F[\bar{V}] \ , \tag{10.42}$$

where $\tilde{\nabla}$ is the gradient in $T^*\mathcal{P}$, that is, the gradient of a generic function $F(X) = F(\theta, \phi)$ is

$$\tilde{\nabla}F = \frac{\partial F}{\partial\theta^i}\tilde{\nabla}\theta^i + \frac{\partial F}{\partial\phi_i}\tilde{\nabla}\phi_i \ . \tag{10.43}$$

The tilde '~' serves to remind us that this is the gradient $\tilde{\nabla}$ on the bundle $T^*\mathcal{P}$.

To simplify the notation further instead of keeping separate track of the θ^i and ϕ_i coordinates it is more convenient to combine them into a single X. A point $X=(\theta,\phi)$ will then be labelled by its coordinates

$$X^{\alpha i}=(X^{1i},X^{2i})=\left(\theta^i,\phi_i\right) \ , \tag{10.44}$$

where αi is a composite index. The first index α (chosen from the beginning of the Greek alphabet) takes two values, $\alpha=1,2$. It is used to keep track of whether i is an upper θ^i index ($\alpha=1$) or a lower ϕ_i index ($\alpha=2$).[5] Then eqs.(10.41) and (10.43) are written as

$$\bar{V}=\frac{d}{d\lambda}=V^{\alpha i}\frac{\partial}{\partial X^{\alpha i}} \ , \quad \text{with} \quad V^{\alpha i}=\frac{dX^{\alpha i}}{d\lambda}=\begin{bmatrix} d\theta^i/d\lambda \\ d\phi_i/d\lambda \end{bmatrix} \ , \tag{10.45}$$

and

$$\tilde{\nabla}F=\frac{\partial F}{\partial X^{\alpha i}}\tilde{\nabla}X^{\alpha i} \ . \tag{10.46}$$

The repeated indices indicate a double summation over α and i. The action of the linear functional $\tilde{\nabla}F[\cdot]$ on a vector $\bar{V}$ is defined by the action of the basis covectors $\tilde{\nabla}X^{\alpha i}$ on the basis vectors, $\partial/\partial X^{\beta j}=\partial_{\beta j}$,

$$\tilde{\nabla}X^{\alpha i}[\partial_{\beta j}]=\frac{\partial X^{\alpha i}}{\partial X^{\beta j}}=\delta^{\alpha i}_{\beta j} \ . \tag{10.47}$$

Using linearity we find

$$\tilde{\nabla}F[\bar{V}]=\frac{\partial F}{\partial X^{\alpha i}}V^{\alpha i}=\frac{dF}{d\lambda} \ . \tag{10.48}$$

10.4 Hamiltonian flows

We have seen that vectors that are tangent to a space-filling congruence of curves $X^{\alpha i}=X^{\alpha i}(\lambda)$ define a vector field — a vector $\bar{V}(X)$ at each point $X\in T^*\mathcal{P}$. Conversely, a vector field $\bar{V}(X)$ defines the congruence of curves $X^{\alpha i}=X^{\alpha i}(\lambda)$ that are tangent to the field $\bar{V}(X)$ at every point X. We are interested in those special congruences or *flows* that reflect the structure of the symmetries of the manifold.

10.4.1 The symplectic form

Once a manifold is supplied with the symmetric bilinear form that we call the metric tensor a number of remarkable properties follow. The metric tensor gives the manifold a fairly rigid structure that is described as the *geometry* of the manifold. It also induces a natural map from vectors to covectors and a special significance is given to those transformations that preserve the bilinear form and to the associated group of isometries.

Something similar occurs when the manifold happens to be a cotangent bundle. Then it is possible to endow it with an *antisymmetric* bilinear form,

[5] This allows us the freedom to switch from θ^i to θ_i as convenience dictates; occasionally we shall write $\theta_i=\theta^i$.

called the *symplectic form*, and the manifold acquires a certain floppy structure that is somewhat less rigid than that provided by a metric. This structure is described as the *symplectic geometry* of the manifold [Arnold 1997][Souriau 1997][Schutz 1980]. As was the case for the metric tensor, the symplectic form also induces a map from vectors to covectors and the group of transformations that preserve it is particularly important. It is called the symplectic group which in Hamiltonian mechanics has long been known as the group of canonical transformations.

Once local coordinates (θ^i, ϕ_i) on $T^*\mathcal{P}$ have been established there is a natural choice of symplectic form

$$\Omega[\cdot,\cdot] = \tilde{\nabla}\theta^i[\cdot] \otimes \tilde{\nabla}\phi_i[\cdot] - \tilde{\nabla}\phi_i[\cdot] \otimes \tilde{\nabla}\theta^i[\cdot] \; . \tag{10.49}$$

The action of $\Omega[\cdot,\cdot]$ on two vectors $\bar{V} = d/d\lambda$ and $\bar{U} = d/d\mu$ is obtained using (10.47),

$$\tilde{\nabla}\theta^i(\bar{V}) = V^{1i} \quad \text{and} \quad \tilde{\nabla}\phi_i(\bar{V}) = V^{2i} \; . \tag{10.50}$$

The result is

$$\Omega(\bar{V},\bar{U}) = V^{1i}U^{2i} - V^{2i}U^{1i} = \Omega_{\alpha i,\beta j}V^{\alpha i}U^{\beta j} \; , \tag{10.51}$$

so that the components of Ω are

$$\Omega_{\alpha i,\beta j} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \delta_{ij} \; . \tag{10.52}$$

The form Ω is non-degenerate, that is, for every vector $\bar{V}$ there exists some vector $\bar{U}$ such that $\Omega[\bar{V},\bar{U}] \neq 0$.

An aside: In the language of exterior calculus the symplectic form Ω can be derived by first introducing the Poincare 1-form

$$\omega = \phi_i \tilde{\mathsf{d}}\theta^i \; , \tag{10.53}$$

where $\tilde{\mathsf{d}}$ is the exterior derivative on $T^*\mathcal{P}$ and the corresponding symplectic 2-form is

$$\Omega = -\tilde{\mathsf{d}}\omega = \tilde{\mathsf{d}}\theta^i \wedge \tilde{\mathsf{d}}\phi_i \; . \tag{10.54}$$

By construction Ω is locally exact ($\Omega = -\tilde{\mathsf{d}}\omega$) and closed ($\tilde{\mathsf{d}}\Omega = 0$).

Remark: A generic $2n$-dimensional symplectic manifold is a manifold with a differential two-form $\omega(\cdot,\cdot)$ that is closed ($\tilde{\mathsf{d}}\omega = 0$) and non-degenerate (that is, if the vector field $\bar{V}$ is nowhere vanishing, then the one-form $\omega(\bar{V},\cdot)$ is nowhere vanishing too). The Darboux theorem [Guillemin Sternberg 1984] states that one can choose coordinates (q^i, p_i) so that at any point the two-form ω can be written as

$$\omega = \tilde{\mathsf{d}}q^i \wedge \tilde{\mathsf{d}}p_i \; . \tag{10.55}$$

In general this can only be done locally; there is no choice of coordinates that will accomplish this diagonalization globally. The point of this remark is to emphasize that our construction of Ω in (10.49) follows a very different logic: we do not start with a $2n$-dimensional manifold with a pre-assigned two-form $\omega(\cdot,\cdot)$ which we then proceed to locally diagonalize. We start with a $2n$-dimensional manifold and a given set of privileged coordinates (ρ,ϕ) which we then use to construct the globally diagonal symplectic form Ω.

10.4.2 Hamilton's equations and Poisson brackets

Next we derive the $2n$-dimensional $T^*\mathcal{P}$ analogues of results that are standard in classical mechanics [Arnold 1997][Souriau 1997][Schutz 1980]. Given a vector field $\bar V(X)$ on $T^*\mathcal{P}$ we can integrate $V^{\alpha i}(X) = dX^{\alpha i}/d\lambda$ to find its integral curves $X^{\alpha i} = X^{\alpha i}(\lambda)$. We are particularly interested in those vector fields $\bar V(X)$ that generate flows that preserve the symplectic structure in the sense that

$$\pounds_V\Omega = 0\ , \tag{10.56}$$

where the Lie derivative is given by (10.39),

$$(\pounds_V\Omega)_{\alpha i,\beta j} = V^{\gamma k}\partial_{\gamma k}\Omega_{\alpha i,\beta j} + \Omega_{\gamma k,\beta j}\partial_{\alpha i}V^{\gamma k} + \Omega_{\alpha i,\gamma k}\partial_{\beta j}V^{\gamma k}\ . \tag{10.57}$$

Since by eq.(10.52) the components $\Omega_{\alpha i,\beta j}$ are constant, $\partial_{\gamma k}\Omega_{\alpha i,\beta j} = 0$, we can rewrite $\pounds_V\Omega$ as

$$(\pounds_V\Omega)_{\alpha i,\beta j} = \partial_{\alpha i}(\Omega_{\gamma k,\beta j}V^{\gamma k}) - \partial_{\beta j}(\Omega_{\alpha i,\gamma k}V^{\gamma k})\ , \tag{10.58}$$

which is the exterior derivative (roughly, the curl) of the covector $\Omega_{\gamma k,\alpha i}V^{\gamma k}$. By Poincare's lemma, requiring $\pounds_V\Omega = 0$ (a vanishing curl) implies that $\Omega_{\gamma k,\alpha i}V^{\gamma k}$ is the gradient of a scalar function, which we will denote $\tilde V(X)$,

$$\Omega_{\gamma k,\alpha i}V^{\gamma k} = \partial_{\alpha i}\tilde V \quad \text{or} \quad \Omega(\bar V,\cdot) = \tilde\nabla\tilde V(\cdot)\ . \tag{10.59}$$

In the opposite direction we can easily check that (10.59) implies $\pounds_V\Omega = 0$. Indeed,

$$(\pounds_V\Omega)_{\alpha i,\beta j} = \partial_{\alpha i}(\partial_{\beta j}\tilde V) - \partial_{\beta j}(\partial_{\alpha i}\tilde V)\ = 0\ . \tag{10.60}$$

Using (10.52), eq.(10.59) is more explicitly written as

$$\frac{d\theta^i}{d\lambda}\tilde\nabla\phi_i - \frac{d\phi_i}{d\lambda}\tilde\nabla\theta^i = \frac{\partial\tilde V}{\partial\theta^i}\tilde\nabla\theta^i + \frac{\partial\tilde V}{\partial\phi_i}\tilde\nabla\phi_i\ , \tag{10.61}$$

or

$$\frac{d\theta^i}{d\lambda} = \frac{\partial\tilde V}{\partial\phi_i} \quad \text{and} \quad \frac{d\phi_i}{d\lambda} = -\frac{\partial\tilde V}{\partial\theta^i}\ , \tag{10.62}$$

which we recognize as Hamilton's equations for a Hamiltonian function $\tilde V$. This justifies calling $\bar V$ the *Hamiltonian vector field* associated to the *Hamiltonian function* $\tilde V$. This is how Hamiltonians enter physics — as a way to generate vector fields that preserve Ω.

From (10.51), the action of the symplectic form Ω on two Hamiltonian vector fields $\bar V = d/d\lambda$ and $\bar U = d/d\mu$ generated respectively by $\tilde V$ and $\tilde U$ is

$$\Omega(\bar V,\bar U) = \frac{d\theta^i}{d\lambda}\frac{d\phi_i}{d\mu} - \frac{d\phi_i}{d\lambda}\frac{d\theta^i}{d\mu}\ , \tag{10.63}$$

which, using (10.62), gives

$$\Omega(\bar{V},\bar{U}) = \frac{\partial \tilde{V}}{\partial \theta^i}\frac{\partial \tilde{U}}{\partial \phi_i} - \frac{\partial \tilde{V}}{\partial \phi_i}\frac{\partial \tilde{U}}{\partial \theta^i} \overset{\text{def}}{=} \{\tilde{V},\tilde{U}\} \ , \tag{10.64}$$

where, on the right hand side, we have introduced the Poisson bracket notation. It is easy to check that the derivative of an arbitrary function $F(X)$ along the congruence defined by the vector field $\bar{V} = d/d\lambda$, which is given by (10.42) or (10.48),

$$\frac{dF}{d\lambda} = \frac{\partial F}{\partial X^{\alpha i}}\frac{dX^{\alpha i}}{d\lambda} = \frac{\partial F}{\partial \theta^i}\frac{d\theta^i}{d\lambda} + \frac{\partial F}{\partial \phi_i}\frac{d\phi_i}{d\lambda} \ , \tag{10.65}$$

can be expressed in terms of Poisson brackets,

$$\frac{dF}{d\lambda} = \{F,\tilde{V}\} \ . \tag{10.66}$$

These results are summarized as follows:
(1) The flows that preserve the symplectic structure, $\pounds_V\Omega = 0$, are generated by Hamiltonian vector fields $\bar{V}$ associated to Hamiltonian functions $\tilde{V}$, eq.(10.62),

$$V^{\alpha i} = \frac{dX^{\alpha i}}{d\lambda} = \{X^{\alpha i},\tilde{V}\} \ . \tag{10.67}$$

(2) The action of Ω on two Hamiltonian vector fields is the Poisson bracket of the associated Hamiltonian functions,

$$\Omega(\bar{V},\bar{U}) = \Omega_{\alpha i,\beta j}V^{\alpha i}U^{\beta j} = \{\tilde{V},\tilde{U}\} \ . \tag{10.68}$$

We end this section with a word of caution. We have uncovered a mathematical formalism that resembles classical mechanics. We could arbitrarily choose one particular Hamiltonian function $\tilde{V}$ and call it $\tilde{H}$, and we could rename the parameter λ and call it time, but this does not mean that we have thereby constructed a dynamical theory. One facet of the problem is that the choice of symplectic form depends on the choice of local coordinates (θ^i,ϕ_i). It may be natural to assign a privileged statistical or physical significance to the parameters θ^i but how do we choose the corresponding conjugate momentum ϕ_i? In classical mechanics this question is settled by appealing to a Lagrangian $L(q,\dot{q})$ which allows us to define the conjugate momentum $p_i = \partial L/\partial \dot{q}^i$, but here we do not have a Lagrangian. Another facet of this same problem is that time is not merely just another parameter labeling points along a curve. What makes time special? What choices of Hamiltonian functions qualify as being the generators of time evolution? Having raised these issues — which will be addressed in the next chapter — it is nevertheless nothing less than astonishing to see that the familiar Hamiltonian formalism emerges from purely geometrical considerations.

Although we have not yet constructed a proper dynamical theory it is desirable to adopt a more suggestive notation that anticipates the dynamics to be derived in the next chapter: The flow generated by a Hamiltonian function $\tilde{H}(X)$ and parametrized by τ is given by Hamilton's equations

$$\frac{d\theta^i}{d\tau}=\frac{\partial\tilde{H}}{\partial\phi_i}\quad\text{and}\quad\frac{d\phi_i}{d\tau}=-\frac{\partial\tilde{H}}{\partial\theta^i}\ ,\tag{10.69}$$

and the τ evolution of any function $f(X)$ given by the Hamiltonian vector $\bar{H}(X)$ is

$$\frac{df}{d\tau}=\bar{H}(f)=\{f,\tilde{H}\}\quad\text{with}\quad\bar{H}=\frac{\partial\tilde{H}}{\partial\phi_i}\frac{\partial}{\partial\theta^i}-\frac{\partial\tilde{H}}{\partial\theta^i}\frac{\partial}{\partial\phi_i}\ .\tag{10.70}$$

10.5 The information geometry of e-phase space

As a prelude to a true dynamics we wish to characterize those special flows that reflect the structures intrinsic to the e-phase space $T^*\mathcal{P}$. We have already discussed flows that preserve the symplectic structure. Next we consider the other natural structure present in a statistical manifold, namely, its metric structure. The immediate obstacle here is that although the space $\mathcal{P}$ is a statistical manifold and is automatically endowed with a unique information metric, the cotangent bundle $T^*\mathcal{P}$ is not a statistical manifold. Thus, our next goal is to endow $T^*\mathcal{P}$ with a metric that is compatible with the metric of $\mathcal{P}$.

Once a metric structure is in place we can ask: does the distance between two neighboring points — the extent to which we can *distinguish* them — grow or decrease with the flow? Or does it stay the same? There are many possibilities but for pragmatic (and esthetic) reasons we are led to consider the simplest form of flow — one that preserves the metric. This will lead us to study the Hamilton flows (those that preserve the symplectic structure) that are also Killing flows (those that preserve the metric structure).

10.5.1 The metric of e-phase space $T^*\mathcal{P}$

The present goal is to extend the metric of the statistical manifold $\mathcal{P}$ — given by information geometry — to the full e-phase space, $T^*\mathcal{P}$. The extension can be carried out in many ways; here we focus on a particular extension that will turn out to be useful for quantum mechanics. The virtue of the formulation below is that the number of input assumptions is kept to a minimum.

The central idea is that *the only metric structure at our disposal* is that of the statistical manifold $\mathcal{P}$. As we saw in Chapter 7,

$$\delta\ell^2=g_{ij}\delta\theta^i\delta\theta^j\ ,\tag{10.71}$$

where

$$g_{ij}(\theta)=\sum\nolimits_x\rho(x|\theta)\frac{\partial\log\rho(x|\theta)}{\partial\theta^i}\frac{\partial\log\rho(x|\theta)}{\partial\theta^j}\ .\tag{10.72}$$

Since the only available tensor is g_{ij} the length element of $T^*\mathcal{P}$,

$$\delta\tilde{\ell}^2=G_{\alpha i,\beta j}\delta X^{\alpha i}\delta X^{\beta j}=G_{1i,1j}\delta\theta^i\delta\theta^j+2G_{1i,2j}\delta\theta^i\delta\phi_j+G_{2i,2j}\delta\phi_i\delta\phi_j\ ,\tag{10.73}$$

must be of the form

$$\delta\tilde{\ell}^2 = \alpha g_{ij}\delta\theta^i\delta\theta^j + \beta g_i^j\delta\theta^i\delta\phi_j + \gamma g^{ij}\delta\phi_i\delta\phi_j \ , \tag{10.74}$$

where α, β, and γ are constants to be determined next.

To fix the value of α we recall that the information metric g_{ij} is unique up to an overall multiplicative constant which is ultimately irrelevant; its role is to set the units of ℓ and of $\tilde{\ell}$ relative to those of θ. We can rewrite (10.74) as

$$\delta\tilde{\ell}^2 = \alpha\left[g_{ij}\delta\theta^i\delta\theta^j + \beta' g_i^j\delta\theta^i\delta\phi_j + \gamma' g^{ij}\delta\phi_i\delta\phi_j\right] \ , \tag{10.75}$$

with new constants β' and γ' and we can either keep or drop α as convenience or convention dictates. In contrast, the values of β' and γ' are significant; they are not a matter of convention. For future convenience we shall write $\gamma' = 1/h^2$ in terms of a new constant h, and choose the irrelevant α as $\alpha = h$.[6] This allows us to absorb h into g_{ij} and write

$$g_{ij}(\theta) = h\sum\nolimits_x \rho(x|\theta)\frac{\partial\log\rho(x|\theta)}{\partial\theta^i}\frac{\partial\log\rho(x|\theta)}{\partial\theta^j} \ . \tag{10.76}$$

To fix the value of β' we impose an additional requirement that is motivated by its eventual relevance to physics. Consider a curve $[\theta(\tau),\phi(\tau)]$ on $T^*\mathcal{P}$ and its flow-reversed curve — or τ-reversed curve — is given by

$$\theta(\tau) \to \theta'(\tau) = \theta(-\tau) \quad \text{and} \quad \phi(\tau) \to \phi'(\tau) = -\phi(-\tau) \ . \tag{10.77}$$

When projected to $\mathcal{P}$ the flow-reversed curve coincides with the original curve, but it is now traversed in the opposite direction. We shall require that the speed $|d\tilde{\ell}/d\tau|$ remains invariant under flow-reversal. Since under flow-reversal the mixed $\theta\phi$ terms in (10.74) change sign, it follows that invariance implies that $\beta' = 0$.

The net result is that the line element, *which has been designed to be fully determined by information geometry*, takes a particularly simple form,

$$\delta\tilde{\ell}^2 = g_{ij}\delta\theta^i\delta\theta^j + g^{ij}\delta\phi_i\delta\phi_j \ . \tag{10.78}$$

Remark: We emphasize that assuming that e-phase space is symmetric under flow-reversal does not amount to imposing that the dynamics itself be *time*-reversal invariant. Eventually we will want to construct dynamical models which exhibit time-reversal symmetry for some interactions and violate it for others. This requires an e-phase space that *allows* the symmetry, and any potential violations will then be due to specific interaction terms in the Hamiltonian. In other words, we shall restrict ourselves to models in which time-reversal violations are induced at the dynamical level of the Hamiltonian and not at the kinematical level of the geometry of e-phase space.

[6] In the next chapter we shall find it useful to assign conventional units to the momenta ϕ that are conjugate to the probability densities ρ of the positions of particles. There we shall rewrite γ' as $\gamma' = 1/\hbar^2$, and the (irrelevant) constant α as $\alpha = \hbar$. In conventional units the value of $\hbar$ is fixed by experiment but one can always choose units so that $\hbar = 1$.

10.5.2 A complex structure for $T^*\mathcal{P}$

The metric tensor $G_{\alpha i,\beta j}$ and its inverse $G^{\alpha i,\beta j}$ can be used to lower and raise indices. In particular, we can raise the first index of the symplectic form $\Omega_{\alpha i,\beta j}$ in eq.(10.52)

$$G^{\alpha i,\gamma k}\Omega_{\gamma k,\beta j} = -J^{\alpha i}{}_{\beta j} \ . \tag{10.79}$$

(The convenience of introducing a minus sign will become clear later. See eqs.(10.158) and (10.159)) We note that both $G_{\alpha i,\beta j}$ and the symplectic form $\Omega_{\alpha i,\beta j}$ in eq.(10.52) map vectors to covectors while the tensor $J^{\alpha i}{}_{\beta j}$ maps vectors to vectors. Indeed, the action of J is such that Ω maps a vector to a covector which is then mapped by the inverse G^{-1} back to a vector.

The tensor J has an important property that is most easily derived by writing G and Ω in block matrix form,

$$G = \begin{bmatrix} g & 0 \\ 0 & g^{-1} \end{bmatrix} , \ G^{-1} = \begin{bmatrix} g^{-1} & 0 \\ 0 & g \end{bmatrix} , \ \Omega = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \ . \tag{10.80}$$

Then eq.(10.79) is

$$J = -G^{-1}\Omega = \begin{bmatrix} 0 & -g^{-1} \\ g & 0 \end{bmatrix} \ . \tag{10.81}$$

We can immediately check that

$$JJ = -\mathbf{1} \quad \text{or} \quad J^{\alpha i}{}_{\gamma k}J^{\gamma k}{}_{\beta j} = -\delta^{\alpha i}_{\beta j} \ , \tag{10.82}$$

which shows that J is a square root of the negative unit matrix. This fact is expressed by saying that J endows $T^*\mathcal{P}$ with a complex structure.

Furthermore, we can check that the action of J on any two vectors $\bar{U}$ and $\bar{V}$ is an isometry, that is

$$G(J\bar{U}, J\bar{V}) = G(\bar{U}, \bar{V}) \ . \tag{10.83}$$

Proof: In matrix form the LHS is

$$(J\bar{U})^T G(J\bar{V}) = \bar{U}^T J^T GJ\bar{V} \tag{10.84}$$

(the superscript T stands for transpose). But, from (10.80) and (10.81), we have

$$J^T GJ = G \ . \tag{10.85}$$

Therefore

$$(J\bar{U})^T G(J\bar{V}) = \bar{U}^T G\bar{V} \ , \tag{10.86}$$

which is (10.83).

To summarize, in addition to the symplectic Ω and metric G structures the cotangent bundle $T^*\mathcal{P}$ is also endowed with a complex structure J. Such highly structured spaces are generically known as Kähler manifolds. Here we deal with a curved Kähler manifold that is special in that it inherits its metric from information geometry.

10.6 Quantum kinematics: symplectic and metric structures

After considering generic statistical manifolds we shall now specialize to the type of e-configuration space that is relevant to quantum mechanics. In the next chapter the uncertain variable x will be a continuous variable that labels the positions of particles. Here, to simplify the discussion, we shall assume that x is a discrete variable, $x = i = (1 \ldots n)$, such as one might use to describe an n-sided quantum die. The e-configuration space is the $(n-1)$-dimensional simplex,

$$\mathcal{S} = \{\rho |\ \rho(i) \geq 0\,;\ \textstyle\sum_{i=1}^{n} \rho(i) = 1\}\ , \tag{10.87}$$

and as coordinates we shall use the probabilities themselves, $\rho^i = \rho(i)$. Except for the inconvenience that the space $\mathcal{S}$ is constrained to normalized probabilities so that the coordinates ρ^i are not independent, a mere substitution $\theta^i \to \rho^i$ allows the results of the previous sections to carry through essentially unchanged. This technical problem can, however, be handled by embedding the $(n-1)$-dimensional manifold $\mathcal{S}$ into a manifold of one dimension higher, the so-called positive-cone, denoted $\mathcal{S}^+$, where the coordinates ρ^i are unconstrained. Thus, a point $X = (\rho, \phi)$ in the $2n$-dimensional $T^*\mathcal{S}^+$ will be labelled by its coordinates $X^{\alpha i} = (X^{1i}, X^{2i}) = (\rho^i, \phi_i)$, and with the substitution $\theta^i \to \rho^i$ our previous results for $T^*\mathcal{P}$ can be directly imported to $T^*\mathcal{S}^+$.

The issue of normalization has, however, important consequences which we address next.

10.6.1 The normalization constraint

Since our actual interest is not in flows on the extended $T^*\mathcal{S}^+$ but on the constrained $T^*\mathcal{S}$ of normalized probabilities we shall consider flows that preserve the normalization of probabilities. Let

$$|\rho| \overset{\text{def}}{=} \sum_{i=1}^{n} \rho^i \quad \text{and} \quad \tilde{N} \overset{\text{def}}{=} 1 - |\rho|\ . \tag{10.88}$$

The Hamiltonians $\tilde{H}$ that are relevant to quantum mechanics are such that the initial condition

$$\tilde{N} = 0 \tag{10.89}$$

is preserved by the flow. However, as we shall see in the next chapter, the actual quantum Hamiltonians will also preserve the constraint $\tilde{N} = \text{const}$ even when the constant does not vanish.[7] Therefore, we have

$$\partial_\tau \tilde{N} = \{\tilde{N}, \tilde{H}\} = 0 \quad \text{or} \quad \textstyle\sum_i \dfrac{\partial \tilde{H}}{\partial \phi_i} = \textstyle\sum_i \dfrac{d\rho^i}{d\tau} = 0\ . \tag{10.90}$$

[7] As we shall see in the next chapter the quantum evolution of probabilities $\rho(x)$ takes the form of a local conservation equation, eq.(11.49). This means that the Hamiltonian will preserve the constraint $\tilde{N} = \text{const}$ whether the constant vanishes or not.

Since the probabilities ρ^i must remain positive we shall further require that $d\rho^i/d\tau \geq 0$ at the border of the simplex where $\rho^i = 0$.

In addition to the flow generated by $\tilde{H}$ we can also consider the flow generated by $\tilde{N}$ and parametrized by ν. From eq.(10.62) the corresponding Hamiltonian vector field $\bar{N}$ is given by

$$\bar{N} = N^{\alpha i}\frac{\partial}{\partial X^{\alpha i}} \quad \text{with} \quad N^{\alpha i} = \frac{dX^{\alpha i}}{d\nu} = \{X^{\alpha i}, \tilde{N}\} \ , \tag{10.91}$$

or, more explicitly,

$$N^{1i} = \frac{d\rho^i}{d\nu} = 0 \ , \quad N^{2i} = \frac{d\phi_i}{d\nu} = 1 \ , \quad \text{or} \quad \bar{N} = \sum_i \frac{\partial}{\partial \phi_i} \ . \tag{10.92}$$

The congruence of curves generated by $\tilde{N}$ is found by integrating (10.92). The resulting curves are

$$\rho^i(\nu) = \rho^i(0) \quad \text{and} \quad \phi_i(\nu) = \phi_i(0) + \nu \ , \tag{10.93}$$

which amounts to shifting all momenta by the i-independent parameter ν.

A Global Gauge Symmetry — We can also see that if $\tilde{N}$ is conserved along $\bar{H}$, then $\tilde{H}$ is conserved along $\bar{N}$,

$$\frac{d\tilde{H}}{d\nu} = \{\tilde{H}, \tilde{N}\} = 0 \ , \tag{10.94}$$

which implies that the conserved quantity $\tilde{N}$ is the generator of a symmetry transformation.

The phase space of interest is the $2(n-1)$-dimensional $T^*\mathcal{S}$ but the description is simplified by using the n unnormalized coordinates ρ of the larger embedding space $T^*\mathcal{S}^+$. The introduction of one superfluous ρ coordinate forces us to also introduce one superfluous ϕ momentum. We eliminate the extra coordinate by imposing the constraint $\tilde{N} = 0$. We eliminate the extra momentum by declaring it unphysical: the shifted point $(\rho', \phi') = (\rho, \phi + \nu)$ is declared to be equivalent to (ρ, ϕ), which we describe by saying that (ρ, ϕ) and $(\rho, \phi + \nu)$ lie on the same "ray". This equivalence is described as a global "gauge" symmetry which, as we shall later see, is the reason why quantum mechanical states are represented by rays rather than vectors in a Hilbert space.

10.6.2 The embedding space $T^*\mathcal{S}^+$

As we saw in section 7.4.3 the metric of a generic embedding space $\mathcal{S}^+$ turns out to be spherically symmetric. This fact turns out to be significant for quantum mechanics. The length element is given by eqs.(7.66) and (7.105)

$$\delta\ell^2 = g_{ij}\delta\rho^i\delta\rho^j \quad \text{with} \quad g_{ij} = A\, n_i n_j + \frac{B}{2\rho^i}\delta_{ij} \ , \tag{10.95}$$

where n is a covector with components $n_i = 1$ for all $i = 1 \ldots n$,[8] and $A = A(|\rho|)$ and $B = B(|\rho|)$ are smooth scalar functions of $|\rho| = \sum \rho^i$. These expressions can be simplified by a suitable change of coordinates.

The important term here is 'suitable'. The point is that coordinates are often chosen because they receive a particularly useful interpretation; the coordinate might be a temperature, an angle or, in our case, a probability. In such cases the advantages of the freedom to change coordinates might be severely outweighed by the loss of a clear physical interpretation. Thus, we seek a change of coordinates $\rho^i \to \rho'^i$ that will preserve the interpretation of ρ as unnormalized or relative probabilities. Such transformations are of the form,

$$\rho^i = \rho^i(\rho') = \alpha(|\rho'|)\rho'^i \ , \tag{10.96}$$

where the scale α is a positive function of $|\rho'|$. Substituting

$$\delta\rho^i = \dot{\alpha}\, |\delta\rho'|\rho'^i + \alpha\delta\rho'^i \tag{10.97}$$

where $\dot{\alpha} = d\alpha/d|\rho'|$ and $|\delta\rho| = \sum \delta\rho^i$ into (10.95) we find

$$\delta\ell^2 = \sum_{ij} \left(A + \frac{B}{2\alpha\rho'^i}\delta_{ij} \right) \left(\dot{\alpha}\, |\delta\rho'|\rho'^i + \alpha\delta\rho'^i \right) \left(\dot{\alpha}\, |\delta\rho'|\rho'^j + \alpha\delta\rho'^j \right) \tag{10.98}$$

where we used $n_i = 1$ and the sum over ij is kept explicit. Then,

$$\delta\ell^2 = g'_{ij}\delta\rho'^i\delta\rho'^j \quad \text{with} \quad g'_{ij} = A'\, n_i n_j + B\alpha\frac{1}{2\rho^i}\delta_{ij} \ , \tag{10.99}$$

where we introduced a new function $A' = A'(|\rho'|)$,

$$A' = A\, \left(\dot{\alpha}|\rho'| + \alpha\right)^2 + \frac{B\dot{\alpha}^2}{2\alpha}|\rho'| + B\dot{\alpha} \ . \tag{10.100}$$

We can now take advantage of the freedom to choose α: set $\alpha(|\rho|) = \hbar/B(|\rho|)$ where $\hbar$ is a constant.[9] Dropping the primes, the length element is

$$\delta\ell^2 = g_{ij}\delta\rho^i\delta\rho^j \quad \text{with} \quad g_{ij} = A(|\rho|)\, n_i n_j + \frac{\hbar}{2\rho^i}\delta_{ij} \ , \tag{10.101}$$

or,

$$\delta\ell^2 = A(|\rho|)\, |\delta\rho|^2 + \sum_i \frac{\hbar}{2\rho^i}(\delta\rho^i)^2 \ . \tag{10.102}$$

[8] The only reason to introduce the peculiar covector n is to maintain the Einstein convention of summing over repeated indices,

$$A n_i n_j \delta\rho^i\delta\rho^j = A{\textstyle\sum_{ij}}\delta\rho^i\delta\rho^j = A|\delta\rho|^2 \ .$$

[9] The constant $\hbar$ plays the same role as h in (10.76). Of course, $\hbar$ will eventually be identified with Planck's constant divided by 2π, and one could choose units so that $\hbar = 1$.

We can check that the inverse tensor g^{ij} is

$$g^{ij} = \frac{2\rho^i}{\hbar}\delta^{ij} + C\rho^i\rho^j \quad \text{where} \quad C(|\rho|) = \frac{-2A}{\hbar A|\rho| + \hbar^2/2} \ . \tag{10.103}$$

We are now ready to write down the metric for T^*S^+. We follow the same argument that led to eq.(10.78) and impose invariance under flow reversal. Since the only tensors at our disposal are g_{ij} and g^{ij}, the length element of $T^*\mathcal{S}^+$ must be of the form,

$$\delta\tilde{\ell}^2 = G_{\alpha i,\beta j}\delta X^{\alpha i}\delta X^{\beta j} = g_{ij}\delta\rho^i\delta\rho^j + g^{ij}\delta\phi_i\delta\phi_j \ . \tag{10.104}$$

Therefore, substituting (10.95) and (10.103), $\delta\tilde{\ell}^2$ can be more explicitly written as

$$\delta\tilde{\ell}^2 = A\left(\sum_{i=1}^{n}\delta\rho_i\right)^2 + C\left(\sum_{i=1}^{n}\rho_i\delta\phi_i\right)^2 + \sum_{i=1}^{n}\left(\frac{\hbar}{2\rho_i}\delta\rho_i^2 + \frac{2\rho_i}{\hbar}\delta\phi_i^2\right) \ , \tag{10.105}$$

or

$$\delta\tilde{\ell}^2 = A|\delta\rho|^2 + C|\rho|^2\langle\delta\phi\rangle^2 + \sum_{i=1}^{n}\left(\frac{\hbar}{2\rho_i}\delta\rho_i^2 + \frac{2\rho_i}{\hbar}\delta\phi_i^2\right) \ . \tag{10.106}$$

From (10.104), writing the $\rho\phi$ indices as a 2×2 matrix, the metric tensors are

$$G = \begin{bmatrix} g & 0 \\ 0 & g^{-1} \end{bmatrix} \ , \ G^{-1} = \begin{bmatrix} g^{-1} & 0 \\ 0 & g \end{bmatrix} \ . \tag{10.107}$$

As before, the tensor G and its inverse G^{-1} can be used to lower and raise indices. Using G^{-1} to raise the first index of the symplectic form $\Omega_{\alpha i,\beta j}$ as we did in eq.(10.79), we see that eqs.(10.80) and (10.81),

$$\Omega = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad \text{and} \quad J = -G^{-1}\Omega = \begin{bmatrix} 0 & -g^{-1} \\ g & 0 \end{bmatrix} \ , \tag{10.108}$$

remain valid for T^*S^+. But ultimately the geometry of $T^*\mathcal{S}^+$ is only of marginal interest; what matters is the geometry it induces on the e-phase space $T^*\mathcal{S}$ of normalized probabilities to which we turn next.

10.6.3 The metric induced on the e-phase space $T^*\mathcal{S}$

We saw that the e-phase space $T^*\mathcal{S}$ can be obtained from the space $T^*\mathcal{S}^+$ by the restriction $|\rho| = 1$ and by identifying the gauge equivalent points (ρ^i, ϕ_i) and $(\rho^i, \phi_i + n_i\nu)$. Consider two neighboring points (ρ^i, ϕ_i) and (ρ'^i, ϕ'_i) with $|\rho| = |\rho'| = 1$. The metric induced on $T^*\mathcal{S}$ will be defined as the shortest $T^*\mathcal{S}^+$ distance between (ρ^i, ϕ_i) and points on the ray defined by (ρ'^i, ϕ'_i). Since the $T^*\mathcal{S}^+$ distance between (ρ^i, ϕ_i) and $(\rho^i + \delta\rho^i, \phi_i + \delta\phi_i + n_i\nu)$ is

$$\delta\tilde{\ell}^2(\nu) = g_{ij}\delta\rho^i\delta\rho^j + g^{ij}(\delta\phi_i + n_i\nu)(\delta\phi_j + n_j\nu) \ , \tag{10.109}$$

the metric on $T^*\mathcal{S}$ will be defined by

$$\delta\tilde{s}^2 = \min_{\nu} \delta\tilde{\ell}^2 \Big|_{|\rho|\ =1} \ . \tag{10.110}$$

The value of ν that minimizes (10.109) is

$$\nu_{\min} = -\langle\delta\phi\rangle = -\sum_i \rho^i \delta\phi_i \ . \tag{10.111}$$

Therefore, setting $|\delta\rho|\ = 0$, the metric on $T^*\mathcal{S}$, which measures the distance between neighboring rays, is

$$\delta\tilde{s}^2 = \sum_{i=1}^{n} \left[\frac{\hbar}{2\rho^i} (\delta\rho^i)^2 + \frac{2\rho^i}{\hbar} (\delta\phi_i - \langle\delta\phi\rangle)^2 \right] \ . \tag{10.112}$$

Although the metric (10.112) is expressed in a notation that may be unfamiliar, it turns out to be equivalent to the well-known *Fubini-Study metric.*[10] The recognition that the e-phase space is the cotangent bundle of a statistical manifold has led us to a derivation of the Fubini-Study metric that emphasizes a natural connection to information geometry.

Remark: The physical meaning of $\hbar$ has, ever since Planck, been a matter of interest and wonder. The interpretations range from the deeply metaphysical "quantum of action", to the "scale" that defines the boundary between quantum and classical regimes, to a mere constant that fixes the units of energy relative to those of frequency ($E = \hbar\omega$) and which can always be chosen equal to one. In entropic dynamics the role of $\hbar$ can be characterized in yet another more geometric way. When assigning the information geometry length to an e-phase space vector $(\delta\rho^i, \delta\phi_i)$ there are independent contributions from the coordinate $\delta\rho^i$ and the momentum components $\delta\phi_i$. The constant $\hbar$ determines the relative weights of the two contributions.

Back to $T^*\mathcal{S}^+$ — Even as we have succeeded in assigning a metric to the e-phase space $T^*\mathcal{S}$ it is still true that the normalization constraint is an inconvenience and that to proceed further in our study of Hamiltonian flows we are forced to return to the larger embedding space $T^*\mathcal{S}^+$. To this end we note an important feature of the $T^*\mathcal{S}$ metric (10.112) that we can exploit to our advantage: the metric is independent of the choice of the *function* $A(|\rho|)$ in eq.(10.105) that defines the particular embedding geometry. Therefore, without any loss of generality, we can impose $A(|\rho|) = 0$.[11] With these choices we assign the simplest possible geometries to the embedding spaces $\mathcal{S}^+$ and $T^*\mathcal{S}^+$, namely, they are both flat. The $T^*\mathcal{S}^+$ metric, eq.(10.104), then becomes

$$\delta\tilde{\ell}^2 = \sum_{i=1}^{n} \left[\frac{\hbar}{2\rho^i} \delta\rho_i^2 + \frac{2\rho^i}{\hbar} \delta\phi_i^2 \right] = G_{\alpha i,\beta j} \delta X^{\alpha i} \delta X^{\beta j} \ . \tag{10.113}$$

[10] This metric was introduced years before the invention of quantum mechanics by G. Fubini (1904) and E. Study (1905) in their studies of shortest paths on complex projective spaces. The latter include the projective Hilbert spaces used in quantum mechanics.

[11] Later we shall explicitly show that choosing $A \neq 0$ has no effect on the Hamiltonian flows that are relevant to quantum mechanics.

Writing the $\alpha\beta$ indices in 2×2 as a matrix, we have

$$[G_{ij}]=\begin{bmatrix}\frac{\hbar}{2\rho_i}\delta_{ij} & 0\\ 0 & \frac{2\rho_i}{\hbar}\delta_{ij}\end{bmatrix}\ , \tag{10.114}$$

and the tensor J, eq.(10.108), which defines the complex structure, becomes

$$J^{\alpha i}{}_{\beta j}=-G^{\alpha i,\gamma k}\Omega_{\gamma k,\beta j}\quad\text{or}\quad [J^i{}_j]=\begin{bmatrix}0 & -\frac{2\rho_i}{\hbar}\delta^i_j\\ \frac{\hbar}{2\rho_i}\delta^i_j & 0\end{bmatrix}\ . \tag{10.115}$$

10.6.4 Refining the choice of cotangent space

Having endowed the e-phase space $T^*\mathcal{S}^+$ with both metric and complex structures we can now revisit and refine our choice of cotangent spaces. So far we had assumed the cotangent space $T^*\mathcal{S}^+_\rho$ at ρ to be the flat n-dimensional Euclidean space $\mathbb{R}^n$. It turns out that the cotangent space that is relevant to quantum mechanics requires a further restriction. To see what this is we argue that the fact that $T^*\mathcal{S}^+$ is endowed with a complex structure suggests a canonical transformation from (ρ,ϕ) to complex coordinates $(\psi, i\hbar\psi^*)$,

$$\psi_j=\rho_j^{1/2}e^{i\phi_j/\hbar}\quad\text{and}\quad i\hbar\psi_j^*=i\hbar\rho_j^{1/2}e^{-i\phi_j/\hbar}\ , \tag{10.116}$$

Thus, a point $\Psi\in T^*\mathcal{S}^+$ has coordinates

$$\Psi^{\mu j}=\begin{pmatrix}\Psi^{1j}\\ \Psi^{2j}\end{pmatrix}=\begin{pmatrix}\psi_j\\ i\hbar\psi_j^*\end{pmatrix}\ , \tag{10.117}$$

where the index $\mu=1,2$ takes two values (with $\mu,\nu,\ldots$ chosen from the middle of the Greek alphabet).

Since changing the phase $\phi_j\to\phi_j+2\pi\hbar$ in (10.116) yields the same point ψ we see that the cotangent space $T^*\mathcal{S}_\rho$ is a flat n-dimensional "hypercube" (its edges have *coordinate* length $2\pi\hbar$) with the opposite faces identified, something like periodic boundary conditions.[12] Thus, the new $T^*\mathcal{S}_\rho$ is still locally isomorphic to the old $\mathbb{R}^n$, which makes it a legitimate choice of cotangent space.

Remark: The choice of cotangent space is central to the derivation of quantum mechanics and some additional justification might be desirable. Here we take the easy way out and argue that identifying the relevant e-phase space in which the quantum dynamics is played out — see chapter 11 — represents significant progress even when its physical origin remains unexplained. Nevertheless, a more illuminating justification has in fact been proposed by Selman Ipek (see section 4.5 in [Ipek 2021]) in the context of the *relativistic* dynamics of fields, a subject that lies outside the scope of the non-relativistic physics discussed in this book.

[12] Strictly, $T^*\mathcal{S}_\rho$ is a parallelepiped; from (10.113) we see that the lengths of its edges are $\tilde{\ell}_j=2\pi(2\hbar\rho_j)^{1/2}$ which vanish at the boundaries of the simplex.

We can check that the transformation from real (ρ,ϕ) to complex coordinates $(\psi, i\hbar\psi^*)$ is canonical, that is, $i\hbar\psi^*$ is the momentum conjugate to ψ. The transformation to the $\psi=\rho^{1/2}e^{i\phi/\hbar}$ coordinates proceeds as follows,

$$\delta\psi_i = \frac{\delta\rho_i}{2\rho_i^{1/2}}e^{i\phi/\hbar} + i\frac{\delta\phi_i}{\hbar}\rho^{1/2}e^{i\phi/\hbar} = \left(\frac{\delta\rho_i}{2\rho_i} + i\frac{\delta\phi_i}{\hbar}\right)\psi_i \tag{10.118}$$

so that

$$\frac{\delta\psi_i}{\psi_i} = \frac{\delta\rho_i}{2\rho_i} + i\frac{\delta\phi_i}{\hbar} \quad \text{and} \quad \frac{\delta\psi_i^*}{\psi_i^*} = \frac{\delta\rho_i}{2\rho_i} - i\frac{\delta\phi_i}{\hbar} \ . \tag{10.119}$$

Adding and subtracting these equations we find

$$\delta\rho_j = \psi_j^*\delta\psi_j + \psi_j\delta\psi_j^* \quad \text{and} \quad \delta\phi_j = \frac{\hbar}{2i\rho_j}\left(\psi_j^*\delta\psi_j - \psi_j\delta\psi_j^*\right) \ . \tag{10.120}$$

The action of Ω on two generic vectors $\bar{V}=d/d\lambda$ and $\bar{U}=d/d\mu$ is

$$\Omega(\bar{V},\bar{U}) = \Omega_{\alpha j,\beta k}\frac{dX^{\alpha j}}{d\lambda}\frac{dX^{\beta k}}{d\mu} = \frac{d\rho^j}{d\lambda}\frac{d\phi_j}{d\mu} - \frac{d\phi_j}{d\lambda}\frac{d\rho^j}{d\mu} \ . \tag{10.121}$$

In ψ coordinates this becomes

$$\Omega(\bar{V},\bar{U}) = \Omega_{\mu j,\nu k}\frac{d\Psi^{\mu j}}{d\lambda}\frac{d\Psi^{\nu k}}{d\mu} = \frac{d\psi_j}{d\lambda}\frac{di\hbar\psi_j^*}{d\mu} - \frac{di\hbar\psi_j^*}{d\lambda}\frac{d\psi_j}{d\mu} \ , \tag{10.122}$$

which shows that the symplectic form Ω,

$$[\Omega_{jk}] = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\delta_{jk} \ , \tag{10.123}$$

retains the same form as (10.108).[13]

Similarly, the metric G on $T^*\mathcal{S}^+$, eq.(10.113), becomes

$$\delta\tilde{\ell}^2 = -2i\sum_{j=1}^{n}\delta\psi_j\delta i\hbar\psi_j^* = G_{\mu j,\nu k}\,\delta\Psi^{\mu j}\delta\Psi^{\nu k} \ , \tag{10.124}$$

and the metric tensor and its inverse take a particularly simple form,

$$[G_{jk}] = -i\delta_{jk}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad [G^{jk}] = i\delta^{jk}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \ . \tag{10.125}$$

[13] The canonical transformation is generated by the function

$$F(\rho,\psi) = \frac{-i\hbar}{2}\sum_k \rho_k\left(1+\log\frac{\psi_k^2}{\rho_k}\right)$$

according to

$$\frac{\partial F}{\partial\rho_k} = \phi_k \ , \quad \frac{\partial F}{\partial\psi_k} = -i\hbar\psi_k^* \ .$$

Note, in particular, that the choice $A(|\rho|)=0$ for the embedding space has led to a metric tensor that is independent of the coordinates ψ which corroborates that $T^*\mathcal{S}^+$ is indeed flat.

Finally, using $G^{\mu j,\lambda k}$ to raise the first index of $\Omega_{\lambda k,\nu l}$ gives the ψ components of the tensor J

$$J^{\mu j}{}_{\nu l} \overset{\text{def}}{=} -G^{\mu j,\lambda k}\Omega_{\lambda k,\nu l} \quad \text{or} \quad [J^j{}_l] = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} \delta^j_l \ . \tag{10.126}$$

10.7 Quantum kinematics: Hamilton-Killing flows

We have seen that Hamiltonian flows preserve the symplectic form. Our next goal is to find those Hamiltonian flows $\bar{H}$ that also happen to preserve the metric G of $T^*\mathcal{S}$, that is, we also want $\bar{H}$ to be a Killing vector.

Remark: It may be worthwhile to emphasize that while we adopt the usual H notation associated with time evolution, the vector field $\bar{H}$ refers to any flow that preserves the normalization $\tilde{N}$, the symplectic form Ω, and the metric G.

The condition for $H^{\mu i}$ is

$$\pounds_H G = 0 \ , \tag{10.127}$$

or, using (10.38),

$$(\pounds_H G)_{\mu j,\nu k} = H^{\lambda l}\partial_{\lambda l}G_{\mu j,\nu k} + G_{\lambda l,\nu k}\partial_{\mu j}H^{\lambda l} + G_{\mu j,\lambda l}\partial_{\nu k}H^{\lambda l} = 0 \ . \tag{10.128}$$

The metric G, eq.(10.125), gives $\partial_{\lambda l}G_{\mu j,\nu k}=0$, and the Killing equation simplifies to

$$(\pounds_H G)_{\mu j,\nu k} = G_{\lambda l,\nu k}\partial_{\mu j}H^{\lambda l} + G_{\mu j,\lambda l}\partial_{\nu k}H^{\lambda l} = 0 \ . \tag{10.129}$$

More explicitly,

$$[(\pounds_H G)_{jk}] = -i \begin{bmatrix} \frac{\partial H^{2k}}{\partial \psi_j} + \frac{\partial H^{2j}}{\partial \psi_k} \ ; & \frac{\partial H^{1k}}{\partial \psi_j} + \frac{\partial H^{2j}}{\partial i\hbar\psi^*_k} \\ \frac{\partial H^{2k}}{\partial i\hbar\psi^*_j} + \frac{\partial H^{1j}}{\partial \psi_k} \ ; & \frac{\partial H^{1k}}{\partial i\hbar\psi^*_j} + \frac{\partial H^{1j}}{\partial i\hbar\psi^*_k} \end{bmatrix} = 0 \ . \tag{10.130}$$

If we further require that $\bar{H}$ be a Hamiltonian flow, $\pounds_H\Omega=0$, then $H^{\mu j}$ satisfies Hamilton's equations,

$$H^{1j} = \frac{\partial \tilde{H}}{\partial i\hbar\psi^*_j} \quad \text{and} \quad H^{2j} = -\frac{\partial \tilde{H}}{\partial \psi_j} \ , \tag{10.131}$$

and we find

$$[(\pounds_H G)_{jk}] = 2i \begin{bmatrix} \frac{\partial^2 \tilde{H}}{\partial \psi_j \partial \psi_k} & 0 \\ 0 & \frac{1}{\hbar^2}\frac{\partial^2 \tilde{H}}{\partial \psi^*_j \partial \psi^*_k} \end{bmatrix} = 0 \ , \tag{10.132}$$

so that

$$\frac{\partial^2 \tilde{H}}{\partial \psi_j \partial \psi_k} = 0 \quad \text{and} \quad \frac{\partial^2 \tilde{H}}{\partial \psi^*_j \partial \psi^*_k} = 0 \ . \tag{10.133}$$

Therefore, in order to generate a flow that preserves both G and Ω, the function $\tilde{H}(\psi,\psi^*)$ must be *linear in* ψ *and linear in* ψ^*,

$$\tilde{H}(\psi,\psi^*)=\sum_{j,k=1}^{n}\psi_j^*\hat{H}_{jk}\psi_k+\sum_{j=1}^{n}\left(\psi_j^*\hat{L}_j+\hat{M}_j\psi_j\right)+\text{const} \;, \tag{10.134}$$

where the kernels $\hat{H}_{jk}$, $\hat{L}_j$, and $\hat{M}_j$ are independent of ψ and ψ^*, and the additive constant can be dropped because it has no effect on the flow. Imposing that the flow preserves the normalization constraint $\tilde{N}=\text{const}$, eq.(10.90), implies that $\tilde{H}$ must be invariant under the phase shift $\psi\to\psi e^{i\nu}$. Therefore, $\hat{L}_j=\hat{M}_j=0$, and we conclude that

$$\tilde{H}(\psi,\psi^*)=\sum_{j,k=1}^{n}\psi_j^*\hat{H}_{jk}\psi_k \;. \tag{10.135}$$

The corresponding HK flow is given by Hamilton's equations,

$$\frac{d\psi_j}{d\tau}=H^{1j}=\frac{\partial\tilde{H}}{\partial i\hbar\psi_j^*}=\frac{1}{i\hbar}\sum_{k=1}^{n}\hat{H}_{jk}\psi_k \;, \tag{10.136}$$

$$\frac{di\hbar\psi_j^*}{d\tau}=H^{2j}=-\frac{\partial\tilde{H}}{\partial\psi_j}=-\sum_{k=1}^{n}\psi_k^*\hat{H}_{kj} \;. \tag{10.137}$$

Taking the complex conjugate of (10.136) and comparing with (10.137), shows that the kernel $\hat{H}_{ij}$ is Hermitian, and that the Hamiltonian function $\tilde{H}$ is real,

$$\hat{H}_{jk}^*=\hat{H}_{kj}\quad\text{and}\quad\tilde{H}(\psi,\psi^*)^*=\tilde{H}(\psi,\psi^*) \;. \tag{10.138}$$

To summarize: *the preservation of the symplectic structure, the metric structure, and the normalization constraint leads to Hamiltonian functions* $\tilde{H}$ *that are bilinear in* ψ *and* ψ^*, eq.(10.135). The flow generated by the bilinear Hamiltonian (10.135) is given by the Poisson bracket or its corresponding Hamilton equation,

$$\frac{d\psi_j}{d\tau}=\{\psi_j,\tilde{H}\}\quad\text{or}\quad i\hbar\frac{d\psi_j}{d\tau}=\sum_{k=1}^{n}\hat{H}_{jk}\psi_k \;, \tag{10.139}$$

and the latter is recognized as the Schrödinger equation. Beyond being Hermitian, the actual form of the kernel $\hat{H}_{jk}$ remains undetermined. These are the main results of this chapter.

Linearity — The central feature of Hamilton's equations (10.136), or of the Schrödinger equation (10.139), is that they are linear. Given two solutions $\psi^{(1)}$ and $\psi^{(2)}$ and arbitrary constants c_1 and c_2, the linear combination

$$\psi^{(3)}=c_1\psi^{(1)}+c_2\psi^{(2)} \tag{10.140}$$

is a solution too and this is extremely useful in calculations. Unfortunately, these are Hamilton-Killing flows on the embedding space $T^*\mathcal{S}^+$ and when the flow is projected onto the e-phase space $T^*\mathcal{S}$ the linearity is severely restricted.

If $\psi^{(1)}$ and $\psi^{(2)}$ are normalized the superposition $\psi^{(3)}$ will not in general be normalized except for appropriately chosen constants. More importantly, the gauge-transformed states

$$\psi'^{(1)} = \psi^{(1)} e^{i\nu_1} \quad \text{and} \quad \psi'^{(2)} = \psi^{(2)} e^{i\nu_2} \tag{10.141}$$

are supposed to be "physically" equivalent to the original $\psi^{(1)}$ and $\psi^{(2)}$ but in general the superposition

$$\psi'^{(3)} = c_1 \psi'^{(1)} + c_2 \psi'^{(2)} \tag{10.142}$$

is not equivalent to $\psi^{(3)}$. In other words, *the mathematical linearity of (10.136) or (10.139) does not extend to a full blown Superposition Principle for physically equivalent states.*

On the other hand, any point ψ deserves to be called a "state" in the limited sense that it may serve as the initial condition for a curve in $T^*\mathcal{S}^+$. Since given two states $\psi^{(1)}$ and $\psi^{(2)}$ their superposition $\psi^{(3)}$ is a state too, we see that the set of states $\{\psi\}$ forms a linear vector space. This is a structure that we can further exploit.

The effect of curvature on the HK flows — The previous analysis was based on the fact that the curvature of the embedding space $T^*\mathcal{S}^+$, that is the choice of the function $A(|\rho|)$, has no effect on the metric of the e-phase space $T^*\mathcal{S}$ and therefore should have no effect on the HK flows. This allowed us to simplify the analysis by imposing $A = 0$ which makes $T^*\mathcal{S}^+$ flat. Before we proceed further it is advisable to verify explicitly that choosing $A \neq 0$ still leads to the same bilinear form for the Hamiltonian (10.135).

We recall eqs.(10.105) and (10.107) and use (10.120). Then, for a generic function $A(|\rho|)$, the metric of $T^*\mathcal{S}^+$ in complex coordinates is

$$[G_{ij}] = \begin{bmatrix} A_- \psi_i^* \psi_j^* & -i\delta_{ij} - \frac{i}{\hbar} A_+ \psi_i^* \psi_j \\ -i\delta_{ij} - \frac{i}{\hbar} A_+ \psi_i \psi_j^* & -\frac{1}{\hbar^2} A_{\_} \psi_i \psi_j \end{bmatrix} . \tag{10.143}$$

where

$$A_\pm(|\rho|) = A \pm \frac{C|\rho|^2}{4} \quad \text{with} \quad C(|\rho|) = \frac{-2A}{A\hbar|\rho| + \hbar^2/2} . \tag{10.144}$$

The condition for $\bar{H}$ to generate a Hamilton-Killing flow is given by eqs.(10.127), (10.128), and (10.131). In block matrix form this reads

$$[(\pounds_H G)_{ij}] = \begin{bmatrix} (\pounds_H G)_{1i,1j} & (\pounds_H G)_{1i,2j} \\ (\pounds_H G)_{2i,1j} & (\pounds_H G)_{2i,2j} \end{bmatrix} = 0 . \tag{10.145}$$

The argument involves some straightforward but lengthy algebra. Substitute (10.143) into (10.128), and impose the Hamilton flow condition, eq.(10.131).

Then, the 11 and 12 matrix elements are found to be

$$\begin{aligned}0 = (\pounds_H G)_{1i,1j} &= 2i\frac{\partial^2 \tilde{H}}{\partial\psi_i \partial\psi_j} + iA_-\psi_i^*\left(1-\psi_k^*\frac{\partial}{\partial\psi_k^*}\right)\frac{\partial\tilde{H}}{\partial\psi_j} \\ &+ iA_-\psi_j^*\left(1-\psi_k^*\frac{\partial}{\partial\psi_k^*}\right)\frac{\partial\tilde{H}}{\partial\psi_i} + \frac{i}{\hbar}A_+\psi_i^*\psi_k\frac{\partial^2\tilde{H}}{\partial\psi_k\partial\psi_j} + \frac{i}{\hbar}A_+\psi_j^*\psi_k\frac{\partial^2\tilde{H}}{\partial\psi_k\partial\psi_i}\ ,\end{aligned} \tag{10.146}$$

and

$$\begin{aligned}0 = (\pounds_H G)_{1i,2j} &= -\frac{1}{\hbar}A_+\psi_i^*\left(1-\psi_k\frac{\partial}{\partial\psi_k}\right)\frac{\partial\tilde{H}}{\partial\psi_j^*} + \frac{1}{\hbar}A_+\psi_j\left(1-\psi_k^*\frac{\partial}{\partial\psi_k^*}\right)\frac{\partial\tilde{H}}{\partial\psi_i} \\ &- A_-\psi_i^*\psi_k^*\frac{\partial^2\tilde{H}}{\partial\psi_k^*\partial\psi_j^*} + A\_\psi_j\psi_k\frac{\partial^2\tilde{H}}{\partial\psi_k\partial\psi_i}\ .\end{aligned} \tag{10.147}$$

Similarly, the other two matrix elements, 21 and 22, are given by

$$(\pounds_H G)_{2i,1j} = (\pounds_H G)_{1j,2i} \quad \text{and} \quad (\pounds_H G)_{2i,2j} = -\frac{1}{\hbar^2}(\pounds_H G)_{1i,1j}^*\ , \tag{10.148}$$

which shows that they provide no additional information about the form of $\tilde{H}$ beyond that already provided by eqs.(10.146) and (10.147).

It is easy to verify that the family of bilinear Hamiltonians, eq.(10.135), provides the desired solution. Indeed, we can easily check that a bilinear $\tilde{H}$ implies that the quantities

$$\frac{\partial^2\tilde{H}}{\partial\psi_i\partial\psi_j}\ , \quad \left(1-\psi_k^*\frac{\partial}{\partial\psi_k^*}\right)\frac{\partial\tilde{H}}{\partial\psi_j}\ , \tag{10.149}$$

and their complex conjugates all vanish. Then, both eqs.(10.146) and (10.147) are satisfied identically.

To see that there are no other solutions we argue as follows. Consider (10.146) as a system of linear equations in the unknown 2nd derivatives, $\frac{\partial^2\tilde{H}}{\partial\psi_i\partial\psi_j}$. The coordinates ψ_i and ψ_i^*, the first derivatives $\frac{\partial\tilde{H}}{\partial\psi_i}$, $\frac{\partial\tilde{H}}{\partial\psi_i^*}$, and the mixed derivatives $\frac{\partial^2\tilde{H}}{\partial\psi_i\partial\psi_j^*}$ are independent quantities that define the constant coefficients in the linear system. The number of unknowns is $n(n+1)/2$ and, since

$$(\pounds_H G)_{1j,1i} = (\pounds_H G)_{1i,1j}\ , \tag{10.150}$$

we see that the number of equations matches the number of unknowns. Thus, since the determinant of the system does not vanish, except possibly for special values of the ψs, we conclude that the solution

$$\frac{\partial^2\tilde{H}}{\partial\psi_i\partial\psi_j} = 0 \tag{10.151}$$

is unique. The observation that the remaining eqs.(10.147) are also satisfied too concludes the proof.

In conclusion: whether the embedding space $T^*\mathcal{S}^+$ is flat ($A=0$) or not ($A\neq 0$) the HK flows are described by the linear Schrödinger equation (10.139).

10.8 Hilbert space

We just saw that the possible initial conditions for an HK flow, the *points* $\Psi \in T^*\mathcal{S}^+$, form a linear space. In Section 10.6.2 we had seen that the geometry of the embedding space $T^*\mathcal{S}^+$ was not fully determined. This is a freedom we exploited by setting $A(|\rho|) = 0$ in eq.(10.113) and (10.114) so that $T^*\mathcal{S}^+$ is flat. To take full advantage of linearity we would like to further endow the flat e-phase space with the additional structure of an inner product and thus transform it into a Hilbert space.[14]

The metric tensor defined by (10.125) is supposed to act on *vectors* on this space; its action on the *points* Ψ is not defined — we have a notion of length for vectors but not a notion of length for points Ψ. But in a flat space a point is also a vector. Indeed, since the vector tangent to a curve is (up to a scalar factor) just the difference $\delta\Psi$ of two Ψs, we see that points on the manifold and vectors tangent to the manifold are objects of the same kind. In other words, the tangent spaces $T[T^*\mathcal{S}^+]_\psi$ are identical to the space $T^*\mathcal{S}^+$ itself. Thus, the linear space in which the Ψs are both points and vectors can be identified with the flat embedding e-phase space $T^*\mathcal{S}^+$.

Remark: We have shown that whether the embedding space $T^*\mathcal{S}^+$ is flat ($A = 0$) or curved ($A \neq 0$) the HK flows are described by the very same linear Schrödinger equation (10.139). But it makes no sense to introduce an inner product between *points* in a curved space. It is only when the points happen to also be vectors that inner products and Hilbert spaces make sense. It is, therefore, important to emphasize that the whole additional structure of a Hilbert space is neither necessary nor fundamental. It is a merely useful tool designed for the specific purpose of exploiting the full calculational advantages of linearity.

The inner product — The choice of an inner product for the points Ψ is now "natural" in the sense that the necessary ingredients are already available. The Hamilton-Killing flows followed from imposing that the symplectic form Ω, eq.(10.123), and the flat space tensor G, eq.(10.125), be preserved. In order that the inner product also be preserved it is natural to choose an inner product defined in terms of those two tensors. We adopt the familiar Dirac notation to represent the states Ψ as vectors $|\psi\rangle$. The inner product $\langle\psi|\phi\rangle$ is defined in terms of the tensors G and Ω,

$$\langle\psi|\phi\rangle = a\left(G_{\mu i,\nu j} + b\Omega_{\mu i,\nu j}\right)\Psi^{\mu i}\Phi^{\nu j} \ , \tag{10.152}$$

[14] We use the term Hilbert space loosely to describe any complex vector space with a Hermitian inner product. In this chapter we deal with complex vector spaces of finite dimensionality. The term Hilbert space is more commonly applied to infinite dimensional vector spaces of square-integrable functions that can be spanned by a countable basis. In infinite dimensions all sorts of questions arise concerning the the convergence of sums with an infinite numbers of terms and various other limiting procedures. It can be rigourously shown that the conclusions we draw here for finite dimensions also hold in the infinite dimensional case dimensions.

where a and b are constants. Using eq.(10.123) and (10.125) we get

$$\langle\psi|\phi\rangle = a\left(\psi_j, i\hbar\psi_j^*\right)[G_{jk}+b\Omega_{jk}]\begin{pmatrix}\phi_k\\ i\hbar\phi_k^*\end{pmatrix} = a\hbar\sum_{j=1}^{n}\left((1-ib)\psi_j^*\phi_j+(1+ib)\phi_j^*\psi_j\right)\ . \tag{10.153}$$

We shall adopt the standard definitions and conventions. Requiring that $\langle\psi|\phi\rangle^* = \langle\phi|\psi\rangle$ implies that $a^* = a$ and $b = \pm i|b|$. Furthermore, the standard convention that the inner product $\langle\psi|\phi\rangle$ be anti-linear in its first factor and linear in the second leads us to choose $b = +1$. Finally, we adopt the standard normalization and set $a = 1/2\hbar$. The result is the familiar expression for the positive definite inner product,

$$\langle\psi|\phi\rangle \overset{\text{def}}{=} \frac{1}{2\hbar}\left(G_{\mu j,\nu k}+i\Omega_{\mu j,\nu k}\right)\Psi^{\mu j}\Phi^{\nu k} = \sum_{j=1}^{n}\psi_j^*\phi_j\ . \tag{10.154}$$

We can check that this inner product is positive-definite,

$$\langle\psi|\psi\rangle \geq 0 \quad\text{with}\quad \langle\psi|\psi\rangle = 0 \quad\text{only if}\quad |\psi\rangle = 0\ . \tag{10.155}$$

The map between points and vectors, $\psi \leftrightarrow |\psi\rangle$, is defined by

$$|\psi\rangle = \sum_{j=1}^{n}|j\rangle\psi_j \quad\text{where}\quad \psi_j = \langle j|\psi\rangle\ , \tag{10.156}$$

where, to be explicit about the interpretation, we emphasize that j and $|j\rangle$ are completely different objects: j represents an ontic state while $|j\rangle$ represents an epistemic state — j is one of the faces of the quantum die, and $|j\rangle$ represents the state of certainty that the actual face is j, that is $\rho(j) = 1$. In this "j" representation, the vectors $\{|j\rangle\}$ form a basis that is orthogonal and complete,

$$\langle k|j\rangle = \delta_{jk} \quad\text{and}\quad \sum_{j=1}^{n}|j\rangle\langle j| \ = \hat{1}\ . \tag{10.157}$$

Complex structure — The tensors Ω and G were originally meant to act on tangent vectors but now they can also act on all points $\psi \in T^*\mathcal{S}^+$. For example, the action of the mixed tensor $J = -G^{-1}\Omega$, eq.(10.126), on a point ψ is

$$[(J\Psi)^j] = \begin{bmatrix} i & 0\\ 0 & -i\end{bmatrix}\begin{pmatrix}\psi_j\\ i\hbar\psi_j^*\end{pmatrix} = \begin{pmatrix} i\psi_i\\ i(i\hbar\psi_i)^*\end{pmatrix}\ , \tag{10.158}$$

which shows that J plays the role of multiplication by i, that is, when acting on a point ψ the action of J is represented by an operator $\hat{J}$,

$$\psi \xrightarrow{J} i\psi \quad\text{is}\quad |\psi\rangle \xrightarrow{J} \hat{J}|\psi\rangle = i|\psi\rangle\ . \tag{10.159}$$

Hermitian and unitary operators — There is another insight to be derived from the embedding of e-phase space $T^*\mathcal{S}$ into a flat $T^*\mathcal{S}^+$. From eq.(10.154) we see that the real part is the inner product of the space $\mathbb{R}^{2n}$. The group of transformations that preserve this inner product is the orthogonal group $\mathrm{O}(2n,\mathbb{R})$. Similarly, the imaginary part of (10.154) is the symplectic form which is preserved by transformations of the symplectic group $\mathrm{Sp}(2n,\mathbb{R})$. The transformations that preserve the inner product on the left of (10.154) are the unitary transformations $\mathrm{U}(n,\mathbb{C})$.

We conclude that the subset of symplectic or canonical transformations that are also rotations in e-phase space turn out to be unitary transformations or, in terms of the corresponding groups, the unitary group arises as the intersection of the symplectic and orthogonal groups [Arnold 1997],

$$\mathrm{Sp}(2n,\mathbb{R}) \cap \mathrm{O}(2n,\mathbb{R}) = \mathrm{U}(n,\mathbb{C}) \ . \tag{10.160}$$

The bilinear Hamilton functions $\tilde{H}(\psi,\psi^*)$ with kernel $\hat{H}_{ij}$ in eq.(11.146) can now be written in terms of a Hermitian operator $\hat{H}$ and its matrix elements,

$$\tilde{H}(\psi,\psi^*) = \langle\psi|\hat{H}|\psi\rangle \quad \text{and} \quad \hat{H}_{jk} = \langle j|\hat{H}|k\rangle \ . \tag{10.161}$$

In words: up to normalization the Hamiltonian function $\tilde{H}$ is the expected value of the Hamiltonian operator $\hat{H}$. The corresponding Hamilton-Killing flows are given by

$$i\hbar\frac{d}{d\tau}\langle j|\psi\rangle = \langle j|\hat{H}|\psi\rangle \quad \text{or} \quad i\hbar\frac{d}{d\tau}|\psi\rangle = \hat{H}|\psi\rangle \ . \tag{10.162}$$

These flows are described by unitary transformations

$$|\psi(\lambda)\rangle = \hat{U}_H(\tau)|\psi(0)\rangle \quad \text{where} \quad \hat{U}_H(\tau) = \exp(-i\hat{H}\tau/\hbar) \ . \tag{10.163}$$

Commutators — The Poisson bracket of two Hamiltonian functions $\tilde{U}[\psi,\psi^*]$ and $\tilde{V}[\psi,\psi^*]$,

$$\{\tilde{U},\tilde{V}\} = \sum_{j=1}^{n}\left(\frac{\delta\tilde{U}}{\delta\psi_j}\frac{\delta\tilde{V}}{\delta i\hbar\psi_j^*} - \frac{\delta\tilde{U}}{\delta i\hbar\psi_j^*}\frac{\delta\tilde{V}}{\delta\psi_j}\right) \ ,$$

can be written in terms of the commutator of the associated operators,

$$\{\tilde{U},\tilde{V}\} = -i\hbar\langle\psi|[\hat{U},\hat{V}]|\psi\rangle \ . \tag{10.164}$$

Thus the Poisson bracket is the expectation of the commutator. This *identity* is much sharper than Dirac's pioneering discovery that the quantum commutator of two quantum variables is merely analogous to the Poisson bracket of the corresponding classical variables.

10.9 Assessment: is this all there is to Quantum Mechanics?

The framework above takes us a long way towards justifying the mathematical formalism that underlies quantum mechanics. It clarifies how complex numbers, the Born rule $\rho^i = |\psi_i|^2$, and the linearity of the Schrödinger equation are a consequence of the symplectic structure and the metric structure associated to information geometry. The normalization constraint leads to the equivalence of states along rays in a very convenient Hilbert space. Is there anything else to explain?

There have been numerous attempts to derive or construct the mathematical formalism of quantum mechanics by adapting the symplectic geometry of classical mechanics. Such phase-space methods invariably start from a classical phase space of positions and momenta (q^i, p_i) and through some series of "quantization rules" posit a correspondence to self-adjoint operators $(\hat{Q}^i, \hat{P}_i)$ which no longer constitute a phase space. The connection to a classical mechanics is lost. The interpretation of $\hat{Q}^i$ and $\hat{P}_i$ and even the answer to the question of what, if anything, is real or *ontic* in such a theory all become highly controversial. Probabilities play a secondary role in the formulation; they are introduced almost as an afterthought, as part of phenomenological rules for how to handle those mysterious processes called measurements.

In this chapter we have taken a different starting point that places probabilities at the very foundation. We have discussed special families of curves — the Hamiltonian-Killing flows — that promise to be useful for the study of quantum mechanics. We have shown that the Hamilton-Killing flows that preserve the symplectic and the metric structures of the e-phase space — the cotangent bundle of probabilities ρ^i and their conjugate momenta ϕ_i — reproduce much of the mathematical formalism of quantum theory.

But many questions are immediately raised: When we refer to the probability $\rho^i = \rho(i)$ what is this i that we are uncertain about? It is presumably meant to represent something real, but are there other observables? Are those other observables real, or are they created by the measurement process? Is there a Born rule that applies to them? What are these processes called measurements and how do we model them? Where is classical mechanics in all this?

The Hamilton-Killing flows will be used to describe the evolution of probabilities in time and this raises further questions. As variables go probabilities are very peculiar because they carry a purpose. They are meant to guide us as to what we ought to believe. This means that all their changes — including their evolution in time — must be compatible with the basic entropic principles for updating probabilities. Are Hamilton-Killing flows compatible with entropic updating? And then there is the issue of time itself. Time is not just another parameter along a curve: what makes time special? These and other questions will be addressed in the following chapters.

Chapter 11

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]

"... but it is important to note that the whole content of the theory depends critically on just what we mean by 'probability'."

E. T Jaynes[2]

11.1 Mechanics without mechanism

The drive to explain nature has always led us to seek the mechanisms hidden behind the phenomena. Descartes, for example, claimed to explain the motion of planets as being swept along in the flow of some vortices. The model did not work very well but at least it gave the illusion of a mechanical explanation and thereby satisfied a deep psychological need. Newton's theory fared much better. He took the important step of postulating that gravity was a universal force acting at a distance but he abstained from offering any mechanical explanations

[1] [Wheeler Zurek 1983, p. 203 and 210]

[2] [Jaynes 1957c]

— a stroke of genius immortalized in his famous "hypotheses non fingo." At first there were objections. Huygens, for instance, recognized the undeniable value of Newton's achievement but was nevertheless deeply disappointed: the theory *works* but it does not *explain.* And Newton agreed. In 1693 he wrote that any action at a distance would represent "so great an absurdity... that no man who has in philosophical matters a competent faculty of thinking can ever fall into it." [Newton 1693]

Over the following 18th century, however, impressed by the many successes of Newtonian mechanics, people started to downplay and then even forget their qualms about the absurdity of an action at a distance. Mechanical explanations were, of course, still desired but the very meaning of what counted as "mechanical" suffered a gradual but irreversible shift. It no longer meant "caused by contact forces" but rather "described according to Newton's laws." Over time Newtonian forces, including those mysterious actions at a distance, became "real" which qualified them to count as the causes behind the phenomena.

But this did not last too long. With Lagrange, Hamilton, and the principle of least action, the notion of force started to lose some of its recently acquired fundamental status. Later, after Maxwell succeeded in extending the principles of dynamics to include the electromagnetic field, the meaning of 'mechanical explanation' changed once again. It no longer meant identifying the Newtonian forces, but rather finding the right equations of evolution — which is done by identifying the right Lagrangian or the right Hamiltonian for the theory. Thus, today gravity is no longer explained through a force but through the curvature of space-time. And the concept of force finds no place in quantum mechanics where interactions are described as the evolution of vectors in an abstract Hilbert space.

The goal of this chapter[3] is to derive non-relativistic quantum theory without invoking an underlying mechanism — there is no ontic dynamics operating at a sub-quantum level. This does not mean that such mechanisms do not exist [4,5] It is just that useful models can be constructed without having to go through the trouble of keeping track of a myriad of microscopic ontic details that often turn out to be ultimately irrelevant. The idea can be illustrated by contrasting the two very different ways in which the theory of Brownian motion was originally derived by Smoluchowski and by Einstein. In Smoluchowski's approach one keeps track of the microscopic details of molecular collisions through a stochastic Langevin equation and a macroscopic effective theory is then derived by taking suitable averages. In Einstein's approach, on the other hand, one focuses directly on those pieces of information that turn out to be relevant for the prediction of macroscopic effects. The advantages of Einstein's approach are twofold. On

[3] The contents of this chapter is directly taken from [Caticha 2019 and 2021] which collect material that evolved gradually in a series of previous publications [Caticha 2009a, 2010a, 2010b], [Caticha et al 2014], and [Caticha 2012c, 2014b, 2015a, 2017a, 2017b].

[4] At present this possibility appears unlikely, but we should not underestimate the cleverness of future scholars.

[5] Non-relativistic quantum mechanics can, of course, be derived from an underlying relativistic quantum field theory, but no ontic dynamics is assumed to underwrite the latter (see [Ipek et al 2014, 2018, 2020]).

one hand there is the simplicity that arises from not having to keep track of irrelevant details that are eventually washed out when taking the averages and, on the other hand, it allows the intriguing possibility that there is no ontic sub-quantum dynamics at all.

Quantum mechanics involves probabilities and, therefore, it is a theory of inference. But this has not always been clear. The center of the controversy has been 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][6]

The ontic interpretations have been fairly common. At the very beginning, Schrödinger's original waves were meant to be real material waves — although the need to formulate the theory in configuration space immediately made that interpretation quite problematic.

Then the Copenhagen interpretation — an umbrella designation for the not always overlapping views of Bohr, Heisenberg, Pauli, and Born [Stapp 1972; Jammer 1966, 1974] — took over and became the orthodoxy (see, however, [Howard 2004]). On the question of quantum reality and the epistemic vs. ontic nature of the quantum state it is deliberately vague. As crystallized in the standard textbooks, including the classics by [Dirac 1948], [von Neumann 1955] and [Landau Lifshitz 1977], it regards the quantum state as an objective and complete specification of the properties of the system but only after they become actualized through the act of measurement. According to Bohr the connection between the wave function and the world is indirect. The wave function does not represent the world itself, but is a mere tool to compute probabilities for the outcomes of measurements that we are forced to describe using a classical language that is ultimately inadequate [Bohr 1933, 1958, 1963]. Heisenberg's position is somewhat more ambiguous. While he fully agrees with Bohr on the inadequacy of our classical language to describe a quantum reality, his take on the wave function is that it describes something more ontic, an objective tendency or potentiality for events to occur. And then there is also Einstein's ensemble or statistical interpretation which is more explicitly epistemic. In his words, "the ψ-function is to be understood as the description not of a single system but of an ensemble of systems" [Einstein 1949b, p. 671]. Whether he meant a virtual ensemble in the sense of Gibbs is not so clear. (See also [Ballentine 1970, Fine 1996].)

[6] An important point to be emphasized here is that the distinction ontic/epistemic is not the same as the distinction objective/subjective. (See section 1.1.3.) To be explicit, probabilities are fully epistemic — they are tools for reasoning with incomplete information — but they can lie anywhere in the spectrum from being completely subjective (two different agents can hold different beliefs) to being completely objective. In QM, for example, probabilities are both epistemic and fully objective. Indeed, at the current state of development, anyone who computes probabilities that disagree with QM will be led to experimental predictions that are demonstrably wrong.

Bohr, Heisenberg, Einstein and other founders of quantum theory were all keenly aware of the epistemological and pragmatic elements at the foundation of quantum mechanics (see e.g., [Stapp 1972] on Bohr, and [Fine 1996] on Einstein) but, unfortunately, they wrote at a time when the language and the tools of a quantitative epistemology — the Bayesian and entropic methods that are the subject of this book — had not yet been sufficiently developed.

The conceptual problems that plagued the orthodox interpretation motivated the creation of ontic alternatives such as the de Broglie-Bohm pilot wave theory [Bohm Hiley 1993, Holland 1993], Everett's many worlds interpretation [Everett 1957, Zeh 2016], and the spontaneous collapse theories [Ghirardi et al 1986, Bassi et al 2013]. In these theories the wave function is ontic, it represents a real state of affairs. On the other side, the epistemic interpretations have had a growing number of advocates including, for example, [Ballentine 1970, 1998; Caves et al 2007; Harrigan Spekkens 2010; Friedrich 2011; Leifer 2014].[7] The end result is that the conceptual struggles with quantum theory have 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.* [Schlosshauer 2004; Jaeger 2009; Leifer 2014].

Faced with all this controversy, Jaynes also understood where one might 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 proceeds to explain "is often called Bayesian inference" and is "supplemented by the notion of information entropy".

The Entropic Dynamics (ED) developed below achieves ontological clarity by sharply separating the ontic elements from the epistemic elements — positions of particles (or distributions of fields) on one side and probabilities and their conjugate momenta on the other. In this regard ED is in agreement with Einstein's view that "... on one supposition we should in my opinion hold absolutely fast: The real factual situation of the system S_2 is independent of what is done with system S_1 which is spatially separated from the former." [Einstein 1949a, p.85] (See also [Howard 1985].) ED is also in broad agreement with Bell's views on the desirability of formulating physics in terms of local "beables"[8] (See Bell's papers reproduced in [Bell 2004].)

ED is an epistemic dynamics of probabilities and not an ontic dynamics of particles (or fields). Of course, if probabilities at one instant are large in one place and at a later time they are large in some other place one infers that the particles must have moved — but nothing in ED assumes the existence of something that has pushed the particles around. ED is a *mechanics without*

[7] For criticism of the epistemic view see *e.g.* [Zeh 2002; Ferrero et al 2004; Marchildon 2004].

[8] In contast to mere *observ*ables, the *be*ables are supposed to represent something that is ontic. In the Bohmian approach and in ED particle positions (and fields) are local beables. According to the Bohmian and the many worlds interpretations the wave function is a nonlocal beable.

a mechanism. We avoid the temptation to share Newton's belief that a mechanics without a mechanism is "so great an absurdity..." and assert that *the laws of quantum mechanics are not laws of nature; they are rules for updating probabilities about nature.* The challenge, of course, is to identify those pieces of information that happen to induce the correct updating of probabilities.

In the entropic approach one does not merely postulate a mathematical formalism and then *append* 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 the beliefs of an (ideally rational) agent. It is also necessary to show that the particular ways in which quantum states are handled stand in complete agreement with the tightly constrained ways in which probabilities are to be manipulated, computed, and updated. We can 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 and, in particular, we must show that changes or updates of the epistemic ψ — which include both unitary time evolution according to the Schrödinger equation and its "collapse" during measurement — are nothing but instances of entropic and Bayesian updating (see also Chapter 13). The universal applicability of probability theory including the entropic and Bayesian updating methods leaves no room for alternative "quantum" probabilities obeying alternative forms of Bayesian inference.

There is a large literature on reconstructions of quantum mechanics[9] and there are several approaches based on information theory.[10] Before we proceed with our subject it may be worthwhile to preview some of the features that set ED apart. As mentioned above, one such feature is a strict adherence to Bayesian and entropic methods. Another is a clear ontological commitment: over and above all other observables, positions are assigned the privileged role of being the only ontic variables.[11] As one might expect, ED shows some formal similarities with other position-based models such as the de Broglie-Bohm pilot wave theory [Bohm 1952, Bohm Hiley 1993, Holland 1993] and Nelson's stochastic mechanics [Nelson 1966, 1967, 1985].[12] Indeed, as we shall see, ED allows both the Brownian trajectories of stochastic mechanics and the smooth Bohmian trajectories as special cases. The conceptual differences are otherwise enormous: both stochastic and Bohmian mechanics operate totally at the ontological level while ED operates almost completely at the epistemological level. Bohm's interpretation leads to a deterministic causal theory. Nelson, on the other hand, seeks a realistic interpretation of quantum theory as arising from a deeper, possibly non-local, but essentially stochastic and classical reality. For

[9] See *e.g.*, [Nelson 1985; Adler 2004; Smolin 2006; de la Peña Cetto 2014; Groessing 2008, 2009; 't Hooft 2016] and references therein.

[10] For a very incomplete list where more references can be found see *e.g.*, [Wootters 1981; Rovelli 1996; Caticha 1998, 2006; Zeilinger 1999; Brukner Zeilinger 2002; Fuchs 2002; Spekkens 2007; Goyal et al 2010; Hardy 2001, 2011, Chiribella et al 2011, D'Ariano 2017].

[11] Here we are concerned with non-relativistic quantum mechanics. When the ED framework is applied to relativistic quantum field theory it is the fields that are the only ontic variables [Ipek Caticha 2014, Ipek et al. 2018, 2020].

[12] See also [Guerra 1981, Guerra Morato 1983] and references therein.

him stochastic mechanics 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].

Incidentally, the fact that ED is committed to the reality of positions (or fields) does not stand in conflict with Bell's theorems on the impossibility of local/causal hidden variables [Bell 2004]. As we shall discuss in some detail in Chapter 13 the consequences of Bell's theorem and other no-go theorems (*e.g.*, [Pussey et al 2012]) are evaded by virtue of ED being a purely epistemic dynamics.

Yet another distinguishing feature is a deep concern with the nature of time — even at the non-relativistic level. The issue here is that any discussion of dynamics must inevitably include a notion of time but, being atemporal, the rules of inference are silent on this matter. One can make inferences about the past just as well as about the present or the future. This means that any model of dynamics based on inference must also include assumptions about time, and those assumptions must be explicitly stated. In ED an epistemic notion of time — an "entropic" time — is introduced as a book-keeping device *designed* to keep track of changes. The construction of entropic time involves several ingredients. One must introduce the notion of an instant; one must show that these instants are suitably ordered; and finally one must define a convenient measure of the duration or interval between the successive instants. It turns out that an arrow of time is generated automatically and entropic time is intrinsically directional.

Finally, as mentioned above, ED consists in the entropic updating of probabilities through information supplied by constraints. The challenge is to identify criteria that specify how these constraints are chosen and, in particular, how the constraints are themselves updated. As we saw in Chapter 10 the e-phase space of probabilities and their conjugate momenta is endowed with natural metric and symplectic structures. We propose that the preservation of these structures provides the natural criterion for updating the constraints. The result is a formalism in which the linearity of the Schrödinger equation and the emergence of complex numbers is derived rather than postulated. Interestingly, the introduction of Hilbert spaces turns out to be less a matter of necessity than of computational convenience. The chapter concludes with remarks on the connection between entropic time and the presumably more "physical" notion of time as measured by clocks — we shall argue that what clocks register is, in fact, entropic time.

11.2 The ontic microstates

The first step in any exercise in inference is to specify the quantities to be inferred. We consider N particles living in a flat Euclidean space $\mathbf{X}$. In Cartesian coordinates the metric is δ_{ab}. The particles are assumed to have *definite* positions x_n^a, which we denote collectively by x. The index $n = 1 \ldots N$ labels the particles, and $a = 1, 2, 3$ the three spatial coordinates. The configuration space for N particles is $\mathbf{X}_N = \mathbf{X} \times \ldots \times \mathbf{X}$. The positions of the particles are *unknown* and it is these values that we wish to infer. Since the positions are unknown the main target of our attention will be the probability distribution $\rho(x)$.

The previous paragraph may seem straightforward common sense but the assumptions it involves are not at all innocent. First, note that ED already suggests a new perspective on the old question of determinism vs. indeterminism. If we understand quantum mechanics as a generalization of a deterministic classical mechanics then it is natural to seek the cause of indeterminism. But within an inference framework that is designed to deal with insufficient information one must accept uncertainty, probabilities, and indeterminism as the expected and inevitable norm that requires no explanation. *It is the determinism of classical mechanics that demands an explanation.* Indeed, as we shall see in section **??** while most quantities are afflicted by uncertainties there are situations where for some very specially chosen variables one can, despite the lack of information, achieve complete predictability. This accounts for the emergence of classical determinism from an entropic dynamics that is intrinsically indeterministic.

Second, we note that the assumption that the particles have definite positions represents a major departure from the standard interpretation of quantum mechanics according to which definite values can be attained but only as the result of a measurement. In contrast, positions in ED play the very special role of defining the ontic state of the system. Let us be very explicit: in the ED description of the double slit experiment, we might not know which slit the particle goes through, but the particle definitely goes through one slit or the other.[13] Indeed, as far as positions are concerned, ED agrees with Einstein's view that spatially separated objects have definite separate ontic states [Einstein 1949a, p.85].

And third, we emphasize once again that $\rho(x)$ represents probabilities that are to be manipulated according to exactly the same rules described in previous chapters — the entropic and Bayesian methods. Just as there is no such thing as a quantum arithmetic, there is no quantum probability theory either.

11.3 The entropic dynamics of short steps

Having identified the microstates $x \in \mathbf{X}_N$ we can proceed to the dynamics. *The first assumption is that change happens.* We do not explain why motion happens — ED is silent about any underlying dynamics acting at the sub-quantum level. Instead our task is to produce an estimate of what kind of motion one might reasonably expect.

The goal is to find the probability $P(x'|x)d^{3N}x'$ that the system takes a step from the initial position $x \in \mathbf{X}_N$ into a volume element $d^{3N}x'$ centered at a new $x' \in \mathbf{X}_N$. This is done by maximizing the entropy,

$$S[P,Q] = -\int dx'\, P(x'|x) \log \frac{P(x'|x)}{Q(x'|x)} \ , \tag{11.1}$$

[13] See section 2.5.

relative to a prior $Q(x'|x)$, and subject to the appropriate constraints specified below. (For notational simplicity in multidimensional integrals such as (11.1) we will write dx' instead of $d^{3N}x'$.) It is through the choice of prior and constraints that the relevant pieces of information that define the dynamics are introduced.

11.3.1 The prior

Having decided that changes do happen, we need to be a bit more explicit about which changes are likely to expected. *The main dynamical assumption is that the particles follow trajectories that are continuous.* The assumption of continuity introduces an enormous simplification because it implies that a generic motion can be analyzed as the accumulation of many infinitesimally short steps. Therefore, our first goal will be to use eq.(11.1) to find the transition probability $P(x'|x)$ for an infinitesimally short step. The corresponding prior $Q(x'|x)$ is meant to describe the state of knowledge that is common to all short steps *before* we take into account the additional information that is specific to the particular short step being considered. We shall adopt a prior that incorporates the information that the particles take infinitesimally short steps but is otherwise maximally uninformative. In particular, it will reflect the translational and rotational invariance of the Euclidean space $\mathbf{X}$ and express total ignorance about any correlations. Such a prior can itself be derived from the principle of maximum entropy. Indeed, maximize

$$S[Q,\mu]=-\int dx'\,Q(x'|x)\log\frac{Q(x'|x)}{\mu(x'|x)}\ , \tag{11.2}$$

relative to a uniform measure $\mu(x'|x)$, subject to normalization, and subject to N independent constraints — one for each particle — that impose short steps and rotational invariance,

$$\langle\delta_{ab}\Delta x_n^a\Delta x_n^b\rangle=\kappa_n\ ,\quad (n=1\ldots N)\ , \tag{11.3}$$

where $\Delta x=x'-x$ and κ_n are small constants. The result is

$$Q(x'|x)\propto\mu\exp-\frac{1}{2}\sum\nolimits_n\alpha_n\delta_{ab}\Delta x_n^a\Delta x_n^b\ , \tag{11.4}$$

where the Lagrange multipliers α_n are constants that are independent of x but may depend on the index n in order to describe non-identical particles. The α_ns will be eventually be taken to infinity in order to enforce the fact that the steps are meant to be infinitesimally short. In Cartesian coordinates the uniform measure μ is a numerical constant that can be absorbed into the normalization and therefore has no effect on Q.[14] The result is a *product* of Gaussians,

$$Q(x'|x)\propto\exp-\frac{1}{2}\sum\nolimits_n\alpha_n\delta_{ab}\Delta x_n^a\Delta x_n^b\ , \tag{11.5}$$

[14] Indeed, as $\alpha_n\to\infty$ the prior Q becomes independent of any choice of $\mu(x')$ provided the latter is sufficiently smooth.

which describes the a priori lack of correlations among the particles. Next we specify the constraints that specify the information that is specific to each individual short step.

11.3.2 The phase constraint

In Newtonian dynamics one does not need to explain why a particle perseveres in its motion in a straight line; what demands an explanation — that is, a force — is why the particle deviates from inertial motion. In ED one does not require an explanation for why the particles move. It is taken for granted that things will not stay put; what requires an explanation is how the motion can be both directional and highly correlated. The information about such correlations is introduced through one constraint that acts simultaneously on all particles. The constraint involves a function $\varphi(x) = \varphi(x_1 \ldots x_N)$ on the $3N$-dimensional configuration space, $x \in \mathbf{X}_N$, that we shall refer to as the *drift potential.*

We shall assume that the displacements Δx_n^a are such that the expected change of $\varphi(x)$ is constrained to be

$$\langle \Delta\varphi(x) \rangle = \sum_{n=1}^{N} \frac{\partial \varphi}{\partial x_n^a} \langle \Delta x_n^a \rangle = \kappa'(x) \ , \tag{11.6}$$

where $\kappa'(x)$ is some small but for now unspecified function. This information is already sufficient to construct an interesting entropic dynamics which turns out to be a kind of diffusion where the expected "drift", $\langle \Delta x_n \rangle$, is determined by the "potential" φ.[15]

The physical origin of the potential $\varphi(x)$ is at this point unknown so how can one justify its introduction? First, we note that identifying the relevant constraints, such as (11.6), represents significant progress even when their physical origin remains unexplained. This situation has historical precedents. For example, in Newton's theory of gravity or in the theory of elasticity, the specification of the forces turned out to be very useful even though their microscopic origin had not yet been fully understood. Indeed, as we shall show the assumption of a constraint involving a configuration space function $\varphi(x)$ is instrumental to explain quantum phenomena such as entanglement, interference, and tunneling. A second more formal justification, is motivated by the geometrical discussion in chapter 10. We seek a dynamics in which evolution takes the form of curves or trajectories on the statistical manifold of probabilities $\{\rho\}$. Then, if the probabilities $\rho(x)$ are treated as generalized coordinates, it is only natural to expect that quantities $\varphi(x)$ must at some point be introduced to play the role of their conjugate momenta. What might be surprising is that the single function $\varphi(x)$ will play three roles that might appear to be totally unrelated to each other: first, as a constraint in an entropic inference; second, as the momenta conjugate

[15] However, to construct that particular dynamics that describes quantum systems we must further require that $\varphi/\hbar$ be a multi-valued function with the topological properties of an angle — $\varphi(x)$ and $\varphi(x) + 2\pi\hbar$ represent the same "angle" (see section 11.8.4 below).

to generalized coordinates; and third, as (essentially) the phase of the quantum wave function.

11.3.3 The gauge constraints

The minimal constraints described in the previous paragraphs lead to a rich entropic dynamics but by imposing additional constraints we can construct even more realistic models. To incorporate the effect of an external electromagnetic field we shall impose the additional constraints that the expected displacements $\langle \Delta x_n^a \rangle$ of each particle n satisfy

$$\langle \Delta x_n^a \rangle A_a(\vec{x}_n) = \kappa_n'' \quad \text{for} \quad n = 1 \dots N \ . \tag{11.7}$$

These N constraints involve a single vector field $A_a(\vec{x})$ that lives in the 3-dimensional physical space ($\vec{x} \in \mathbf{X}$). This ensures that all particles couple to one single electromagnetic field. The strength of the coupling is given by the values of the κ_n''. These are small quantities that could be specified directly but, as is often the case in entropic inference, it is much more convenient to specify them indirectly in terms of the corresponding Lagrange multipliers.

11.3.4 The transition probability

Following the by now standard procedure (see section 4.10), the distribution $P(x'|x)$ that maximizes the entropy $S[P,Q]$ in (11.1) relative to (11.5) and subject to (11.6), (11.7), and normalization is

$$P(x'|x) \propto \exp \sum\nolimits_n \{ -\frac{\alpha_n}{2} \delta_{ab} \Delta x_n^a \Delta x_n^b + \alpha' \left(\partial_{na}\varphi - \beta_n A_a(\vec{x}_n) \right) \Delta x_n^a \} \tag{11.8}$$

where α_n, α', and $\alpha'\beta_n$ are Lagrange multipliers, and $\partial_{na} = \partial / \partial x_n^a$.[16] It is convenient to rewrite $P(x'|x)$ as

$$P(x'|x) = \frac{1}{Z} \exp \left[-\frac{1}{2} \sum\nolimits_n \alpha_n \, \delta_{ab} \left(\Delta x_n^a - \overline{\Delta x}_n^a \right) \left(\Delta x_n^b - \overline{\Delta x}_n^b \right) \right] \tag{11.9}$$

where Z is a normalization constant. Thus, a generic displacement $\Delta x_n^a = x_n'^a - x_n^a$ can be expressed as the sum of an expected drift plus a fluctuation,

$$\Delta x_n^a = \overline{\Delta x}_n^a + \Delta w_n^a \ , \tag{11.10}$$

given by

$$\overline{\Delta x}_n^a = \langle \Delta x_n^a \rangle = \frac{\alpha'}{\alpha_n} \delta^{ab} \left[\partial_{nb}\varphi - \beta_n A_b(\vec{x}_n) \right] \quad , \tag{11.11}$$

$$\langle \Delta w_n^a \rangle = 0 \quad \text{and} \quad \langle \Delta w_n^a \Delta w_{n'}^b \rangle = \frac{1}{\alpha_n} \delta^{ab} \delta_{nn'} \ . \tag{11.12}$$

[16] The distribution (11.8) 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. The proof follows the standard argument originally due to Gibbs (see section 4.10).

The directionality of the motion and the correlations among the particles are introduced by a systematic drift in a direction determined by $\partial_{na}\varphi$ and A_a, while the position fluctuations remain isotropic and uncorrelated. As $\alpha_n \to \infty$, the trajectory is expected to be continuous. As we shall see below, whether the trajectory is differentiable or not depends on the particular choices of α_n and α'. Eqs. (11.11) and (11.12) also show that the effect of α' is to enhance or suppress the magnitude of the drift relative to the fluctuations.

11.3.5 Invariance under gauge transformations

The fact that the constraints (11.6) and (11.7) are not independent — they are both linear in the same displacements $\langle \Delta x_n^a \rangle$ — turns out to be significant: it leads to a gauge symmetry and provides the physical interpretation of the vector potential $A_a(\vec{x})$ as the corresponding gauge connection field. This is evident in eq.(11.8) where φ and A_a appear in the combination $\partial_{na}\varphi - \beta_n A_a$ which is invariant under a local gauge transformation,

$$A_a(\vec{x}_n) \to A_a(\vec{x}_n) + \partial_a \chi(\vec{x}_n) \ , \tag{11.13}$$

$$\varphi(x) \to \varphi(x) + \sum\nolimits_n \beta_n \chi(\vec{x}_n) \ , \tag{11.14}$$

where $\chi(\vec{x})$ is a function in 3d-space and the multipliers β_n will later be related to the electric charges q_n by $\beta_n = q_n/c$.

11.4 Entropic time

The transition probability $P(x'|x)$ in (11.9) describes a single short step. To predict motion over finite distances these short steps must be iterated and this is where *time* comes in. Time is introduced as a book-keeping device designed to keep track of the accumulation of short steps.

Since the foundation for any theory of time is dynamics, that is, the theory of change, it is important to be explicit about what changes one is talking about. To be clear, ED involves two different kinds of change. One consists of the ontic changes Δx of the positions that ED is designed to infer; the other reflects the epistemic changes of the evolving probabilities. An ontic dynamics such as, for example, Newtonian mechanics, leads to an ontic notion of time, while a dynamics of probabilities necessarily leads to an epistemic notion of time.

Our task here is to develop *an epistemic model of time* that includes (a) something one might identify as an "instant", (b) a sense in which these instants can be ordered, and (c) a convenient concept of "duration" that measures the separation or interval 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. Such a construction we shall call *entropic time* [Caticha 2010ab]. By design, *entropic time is epistemic and, therefore, it is not ontic.* Later, in section 11.11, after ED has been more fully developed, we shall return to the question of whether and how this epistemic notion of time is related to the presumably more "physical" time that is measured by clocks.

11.4.1 Time as an ordered sequence of instants

A trajectory in ED consists of a succession of short steps that take the system through a sequence of positions $(x_0, x_1, x_2, \ldots)$. Consider, for example, a generic kth step that takes the system from some unknown $x = x_{k-1}$ to an also unknown, neighboring next point $x' = x_k$. Integrating the joint probability $P(x_{k-1}, x_k)$ over x_{k-1} gives

$$P(x_k) = \int dx_{k-1} P(x_{k-1}, x_k) = \int dx_{k-1} P(x_k|x_{k-1}) P(x_{k-1}) \ . \tag{11.15}$$

This equation follows directly from the laws of probability and, therefore, it is true independently of any physical assumptions which means that it is not very useful as it stands. To make it useful, something else must be added.

There is something peculiar about configuration spaces. For example, when we represent the position of a particle as a *point* with coordinates (x^1, x^2, x^3) it is implicitly understood that the values of the three coordinates x^1, x^2, and x^3 hold simultaneously — no surprises here. Things get a bit more interesting when we describe a system of N particles by a single *point* $x = (\vec{x}_1, \vec{x}_2, \ldots \vec{x}_N)$ in $3N$-dimensional configuration space. The point x is meant to represent the state at one instant, that is, it is also implicitly assumed that all the $3N$ coordinate values are simultaneous. What is peculiar about configuration spaces is that they implicitly introduce a notion of simultaneity. Furthermore, when we express uncertainty about the values of an x by means of a probability distribution $P(x)$ it is also implicitly understood that the different possible values of x all refer to the same instant. The system could be at this point here or it could be at that point there, we might not know which, but whichever it is, the two possibilities refer to positions at one and the same instant.[17] And similarly, when we consider the transition probability from x to x', given by $P(x'|x)$, it is implicitly assumed that x refers to one instant, and the possible x's all refer to another instant. Thus, in ED, a probability distribution over configuration space provides a criterion of simultaneity.

We can now return to eq.(11.15): if $P(x_{k-1})$ happens to be the probability of different values of x at *an "initial" instant of entropic time* t, and $P(x_k|x_{k-1})$ is the transition probability from x_{k-1} at one instant to x_k at another instant, then we can interpret $P(x_k)$ as the probability of values of x_k at *a "later" instant of entropic time* $t' = t + \Delta t$. Accordingly, we write $P(x_{k-1}) = \rho_t(x)$ and $P(x_k) = \rho_{t'}(x')$ so that

$$\rho_{t'}(x') = \textstyle\int dx\, P(x'|x) \rho_t(x) \ . \tag{11.16}$$

[17] We could of course consider the joint probability $P(\vec{x}_1(t_1), \vec{x}_2(t_2))$ of particle 1 being at $\vec{x}_1$ at time t_1 and particle 2 being at $\vec{x}_2$ at time t_2, but the set of points $\{\vec{x}_1(t_1), \vec{x}_2(t_2)\}$ is not at all what one would call a configuration space.

Nothing in the laws of probability leading to eq.(11.15) forces the interpretation (11.16) on us — *this is the additional ingredient that allows us to construct time and dynamics in our model.* If the distribution $\rho_t(x)$ refers to one instant t, then the distribution $\rho_{t'}(x')$ generated by $P(x'|x)$ by means of eq.(11.16) defines what we mean by the "next" instant t'. The dynamics is defined by iterating this process. Entropic time is constructed instant by instant: $\rho_{t'}$ is constructed from ρ_t, $\rho_{t''}$ is constructed from $\rho_{t'}$, and so on.

The construction is intimately related to information and inference. *An instant is a complete epistemic state.* Its "completeness" consists in the instant being specified by information — codified into the distributions $\rho_t(x)$ and $P(x'|x)$ — that is sufficient for generating the next instant. Thus, *the present instant is defined so that, given the present, the future is independent of the past.*

Remark: It is common to use equations such as (11.16) to define a special kind of dynamics, called Markovian, that unfolds in a time defined by some external clocks. In such a Markovian dynamics the specification of the state at one instant is sufficient to determine its evolution into the future. Since the transition probability $P(x'|x)$, codifies information supplied through the prior and the constraints neither of which refer to anything earlier than the earlier point x it is clear that formally ED is a Markovian process. There is, however, an important difference. Equation (11.16) is not being used to define a (Markovian) dynamics in a pre-existing background time because ED makes no reference to external clocks. The system is its own clock and (11.16) is used both to define the dynamics and to construct time itself.[18]

Remark: In a relativistic theory there is a greater freedom in the choice of instants which translates into a greater flexibility with the notion of simultaneity. But even in the relativistic setting the notion of instant delineated above remains valid. The new element introduced by relativity is that these different notions of simultaneity must be consistent with each other. This requirement of consistency places severe constraints on the allowed forms of relativistic ED [Ipek et al 2018, 2020].

11.4.2 The arrow of entropic time

The notion of time constructed according to eq.(11.16) is intrinsically directional. There is an absolute sense in which $\rho_{t'}(x')$ occurs after $\rho_t(x)$. To see how this comes about note that the same rules of probability that led us to (11.16) can also lead us to the time-reversed evolution,

$$\rho_t(x) = \int dx'\, P(x|x')\rho_{t'}(x')\,. \tag{11.17}$$

The temporal asymmetry is due to the fact that the distribution $P(x'|x)$, eq.(11.8), is a Gaussian derived using the maximum entropy method, while the time-

[18] In this respect, entropic time bears some resemblance with the relational notion of time advocated by J. Barbour in the context of classical physics (see *e.g.* [Barbour 1994]).

reversed version $P(x|x')$ is related to $P(x'|x)$ by Bayes' theorem,

$$P(x|x') = \frac{\rho_t(x)}{\rho_{t'}(x')} P(x'|x) \ , \tag{11.18}$$

and, therefore, it is not general Gaussian.

The puzzle of the arrow of time (see *e.g.* [Price 1996, Zeh 2007]) arises from the difficulty in deriving a temporal asymmetry from underlying laws of nature that are symmetric. The ED approach offers a fresh perspective on this issue because it does not assume any underlying laws of nature — whether they be symmetric or not. The asymmetry is the inevitable consequence of constructing time in a dynamics driven by entropic inference.

From the ED point of view the challenge does not consist in explaining the arrow of time — *entropic time only flows forward* — but rather in explaining how it comes about that despite the arrow of time some laws of physics, such as the Schrödinger equation, turn out to be time reversible. We will revisit this topic in section 11.11.

11.4.3 Duration

We have argued that the concept of time is intimately connected to the associated dynamics but at this point neither the transition probability $P(x'|x)$ that specifies the dynamics nor the corresponding entropic time have been fully defined yet. It remains to specify how the interval Δt between successive instants is encoded into the multipliers α_n and α'.

The basic criterion for this choice is convenience: *duration is defined so that motion looks simple.* The description of motion is simplest when it reflects the symmetry of translations in space and time. We therefore choose α' and α_n to be constants independent of x and t. The resulting entropic time resembles Newtonian time in that it flows "equably everywhere and everywhen."

The particular choice of duration Δt in terms of the multipliers α_n and α' can be motivated as follows. In Newtonian mechanics time is defined to simplify the dynamics. The prototype of a classical clock is a free particle that moves equal distances in equal times so that there is a well defined (constant) velocity. In ED time is also defined to simplify the dynamics, but now it is the dynamics of probabilities as prescribed by the transition probability. We define duration so that for short steps *the system's expected displacement $\langle\Delta x\rangle$ increases by equal amounts in equal intervals Δt so there is a well defined drift velocity.* Referring to eq.(11.11) this is achieved by setting the ratio α'/α_n proportional to Δt and thus, *the transition probability provides us with a clock.* For future convenience the proportionality constants will be expressed in terms of some particle-specific constants m_n,

$$\frac{\alpha'}{\alpha_n} = \frac{1}{m_n} \Delta t \ . \tag{11.19}$$

At this point the constants m_n receive no interpretation beyond the fact that their dependence on the particle label n recognizes that the particles need not be identical. Eventually, however, the m_ns will be identified with the particle masses.[19]

Having specified the ratio α'/α_n it remains to specify α' (or α_n). It turns out that different choices of α' lead to qualitatively different motions at the sub-quantum or "microscopic" level. Remarkably, however, all of these sub-quantum motions lead to the same dynamics at the quantum level [Bartolomeo Caticha 2016].

The freedom to choose α' and still reproduce quantum mechanics is yet another manifestation of the fact that ED is an epistemic dynamics of probabilities and not an ontic dynamics of positions. ED gives us the probability $P(x',t'|x,t)$ but it is silent on what caused the motion from x to x' and, beyond the fact that the ontic path is continuous, it is also silent on whether the path is smooth or not. Nevertheless, one must define the duration Δt and, to proceed further, one must commit to a definite choice of α'. The situation bears some resemblance to gauge theories, where for actual calculations it is necessary to choose a gauge, even if the actual choice should have no effect on physical predictions.

In the next sections we shall explore the consequences of setting $\alpha' = \text{const}$ which leads to the highly-irregular sub-quantum Brownian trajectories characteristic of Nelson's stochastic mechanics. Thus, we set

$$\alpha' = \frac{1}{\eta} \quad \text{so that} \quad \alpha_n = \frac{m_n}{\eta \Delta t} \ . \tag{11.20}$$

where a new constant η is introduced. Below we shall comment further on the significance of η and in section 11.6 we shall explore a different choice of α'. There we shall set $\alpha' \propto 1/\Delta t^2$ and show that it leads to an ED in which the particles follow the smooth trajectories characteristic of Bohmian mechanics.

11.5 Brownian sub-quantum motion and the evolution equation

It is convenient to introduce a notation tailored to configuration space. Let $x^A = x^a_n$, $\partial_A = \partial/\partial x^a_n$, and $\delta_{AB} = \delta_{nn'}\delta_{ab}$, where $A, B, \ldots$ are composite indices labeling both the particles $(n, n', \ldots)$ and their spatial coordinates $(a, b, \ldots)$. With the choice (11.20), the transition probability (11.9) for a drift potential φ becomes

$$P(x'|x) = \frac{1}{Z} \exp\left[-\frac{1}{2\eta\Delta t} m_{AB} \left(\Delta x^A - \overline{\Delta x}^A \right) \left(\Delta x^B - \overline{\Delta x}^B \right) \right] \tag{11.21}$$

Where we introduced the "mass" tensor, its inverse,

$$m_{AB} = m_n \delta_{AB} = m_n \delta_{nn'} \delta_{ab} \quad \text{and} \quad m^{AB} = \frac{1}{m_n} \delta^{AB} \ . \tag{11.22}$$

[19] If Δt and m_n are given units of time and mass then eqs.(11.11) and (11.19) fix the units of φ.

A generic displacement is then written as a drift plus a fluctuation,

$$\Delta x^A = \overline{\Delta x}^A + \Delta w^A = b^A \Delta t + \Delta w^A \ , \tag{11.23}$$

where, from eqs.(11.11) and (11.25), the drift velocity is

$$b^A(x) = \frac{\langle \Delta x^A \rangle}{\Delta t} = m^{AB} \left[\partial_B \varphi(x) - \bar{A}_B(x) \right] \ , \tag{11.24}$$

and $\bar{A}(x)$ is the electromagnetic vector expressed as a field in configuration space. Its components are

$$\bar{A}_A(x) = \bar{A}_{an}(x) = \beta_n A_a(x_n) \ . \tag{11.25}$$

From (11.21) we see that the fluctuation Δw^A obeys

$$\langle \Delta w^A \rangle = 0 \quad \text{and} \quad \langle \Delta w^A \Delta w^B \rangle = \eta m^{AB} \Delta t \ , \tag{11.26}$$

which shows that the constant η controls the strength of the fluctuations. Note that for very short steps, as $\Delta t \to 0$, the fluctuations become dominant: the drift is $\langle \Delta x^A \rangle \sim O(\Delta t)$ while $\Delta w^A \sim O(\Delta t^{1/2})$ which is characteristic of Brownian paths. Thus, with the choice (11.20), the trajectory is continuous but not differentiable: a particle has a definite position but its velocity, the tangent to the trajectory, is completely undefined. To state this more explicitly, since $\Delta w^A \sim O(\Delta t^{1/2})$ but $\langle \Delta w^A \rangle = 0$, we see that the limit $\Delta t \to 0$ and the expectation $\langle \cdot \rangle$ do not commute,

$$\lim_{\Delta t \to 0} \left\langle \frac{\Delta x^A}{\Delta t} \right\rangle \neq \left\langle \lim_{\Delta t \to 0} \frac{\Delta x^A}{\Delta t} \right\rangle \tag{11.27}$$

because

$$\lim_{\Delta t \to 0} \left\langle \frac{\Delta w^A}{\Delta t} \right\rangle = 0 \quad \text{while} \quad \left\langle \lim_{\Delta t \to 0} \frac{\Delta w^A}{\Delta t} \right\rangle = \infty \ . \tag{11.28}$$

11.5.1 The information metric of configuration space

Before studying the dynamics defined eq.(11.21) we take a brief detour to consider the geometry of the N-particle ontic configuration space, $\mathbf{X}_N$. Since the single particle space $\mathbf{X}$ is described by the Euclidean metric δ_{ab} we can expect that the N-particle configuration space, $\mathbf{X}_N = \mathbf{X} \times \ldots \times \mathbf{X}$, will also be flat, but for non-identical particles a question might be raised about the relative scales or weights associated to each $\mathbf{X}$ factor. Information geometry provides the answer.

The fact that to each point $x \in \mathbf{X}_N$ there corresponds a probability distribution $P(x'|x)$ means that to the ontic configuration space $\mathbf{X}_N$ we can associate a statistical manifold and, as we saw in chapter 7, its geometry is uniquely determined (up to an overall scale factor) by the information metric,

$$\gamma_{AB} = C \int dx' \, P(x'|x) \frac{\partial \log P(x'|x)}{\partial x^A} \frac{\partial \log P(x'|x)}{\partial x^B} \ , \tag{11.29}$$

where C is a positive constant. Substituting eqs.(11.21) into (11.29) in the limit of short steps ($\Delta t \to 0$) yields (see eq. 7.180)

$$\gamma_{AB} = \frac{C}{\eta \Delta t} m_{AB} \ . \tag{11.30}$$

The divergence as $\Delta t \to 0$ arises because the information metric measures statistical distinguishability. As $\Delta t \to 0$ the distributions $P(x'|x)$ and $P(x'|x+\Delta x)$ become more sharply peaked and increasingly easier to distinguish. Therefore, $\gamma_{AB} \to \infty$. To define a geometry that remains useful even for arbitrarily small Δt we can choose $C \propto \Delta t$,

$$\gamma_{AB} \propto m_{AB} \quad \text{or} \quad \gamma_{an,bn'} \propto m_n \delta_{ab} \delta_{nn'} \ . \tag{11.31}$$

Thus, up to overall constants the mass tensor is the metric of configuration space. In words: the metric for the N-particle configuration space, $\mathbf{X}_N = \mathbf{X} \times \ldots \times \mathbf{X}$ is a block diagonal matrix in which the block corresponding to each single particle is the flat Euclidean metric δ_{ab} scaled by the mass of the particle, $m_n \delta_{ab}$.

Ever since the work of H. Hertz in 1894 [Lanczos 1970] it has been standard practice to describe the motion of systems with many particles as the motion of a single point in an abstract space — the configuration space. The choice of the geometry of this *ontic* configuration space has been based on an examination of the kinetic energy of the system. Historically this choice has been regarded as a matter of convenience, a merely useful convention. We can now see that entropic dynamics points to a uniquely natural choice: up to a global scale factor the metric follows uniquely from information geometry.

11.5.2 The evolution equation in differential form

Entropic dynamics is generated by iterating eq.(11.16),

$$\rho_{t+\Delta t}(x') = \int dx \, P(x', t+\Delta t | x, t) \rho_t(x) \ , \tag{11.32}$$

where the times t and $t + \Delta t$ have been written down explicitly. As so often in physics it is more convenient to rewrite the evolution equation above in differential form. One might be tempted to Taylor expand in Δt and $\Delta x = x' - x$, but this is not possible because for small Δt the distribution $P(x', t + \Delta t|x, t)$, eq. (11.21), is very sharply peaked at $x' = x$. To handle such singular behavior one follows an indirect procedure that is well known from diffusion theory [Chandrasekhar 1943]: multiply by a smooth test function $f(x')$ and integrate over x',

$$\int dx' \, \rho_{t+\Delta t}(x') f(x') = \int dx \left[\int dx' \, P(x', t+\Delta t|x, t) f(x') \right] \rho_t(x) \ . \tag{11.33}$$

The test function $f(x')$ is assumed sufficiently smooth precisely so that it can be expanded about x. The important point here is that for Brownian paths eq.(11.26) implies that the terms $(\Delta x)^2$ contribute to $O(\Delta t)$. Then, dropping all terms of order higher than Δt, the integral in the brackets is

$$\begin{aligned}[\cdots] &= \int dx'\, P(x',t+\Delta t|x,t)\left(f(x)+\frac{\partial f}{\partial x^A}\Delta x^A+\frac{1}{2}\frac{\partial^2 f}{\partial x^A\partial x^B}\Delta x^A\Delta x^B+\ldots\right)\\ &= f(x)+b^A(x)\Delta t\frac{\partial f}{\partial x^A}+\frac{1}{2}\Delta t\,\eta m^{AB}\frac{\partial^2 f}{\partial x^A\partial x^B}+\ldots \qquad (11.34)\end{aligned}$$

where we used eq.(11.23) and (11.26),

$$\begin{aligned}\lim_{\Delta t\to 0^+}\frac{1}{\Delta t}\int dx'\, P(x',t+\Delta t|x,t)\Delta x^A &= b^A(x)\ ,\\ \lim_{\Delta t\to 0^+}\frac{1}{\Delta t}\int dx'\, P(x',t+\Delta t|x,t)\Delta x^A\Delta x^B &= \eta m^{AB}\ . \qquad (11.35)\end{aligned}$$

Dropping the primes on the left hand side of (11.33), substituting (11.34) into the right, and dividing by Δt, gives

$$\int dx\,\frac{1}{\Delta t}\left[\rho_{t+\Delta t}(x)-\rho_t(x)\right]f(x)=\int dx\left[b^A(x)\frac{\partial f}{\partial x^A}+\frac{1}{2}\,\eta m^{AB}\frac{\partial^2 f}{\partial x^A\partial x^B}\right]\rho_t(x)\ . \qquad (11.36)$$

Next integrate by parts on the right and let $\Delta t\to 0$. The result is

$$\int dx\,\partial_t\rho_t(x)f(x)=\int dx\left[-\frac{\partial}{\partial x^A}(b^A\rho_t)+\frac{1}{2}\,\eta m^{AB}\frac{\partial^2\rho_t}{\partial x^A\partial x^B}\right]f(x)\ . \qquad (11.37)$$

Since the test function $f(x)$ is arbitrary, we conclude that

$$\partial_t\rho_t=-\partial_A(b^A\rho_t)+\frac{1}{2}\,\eta m^{AB}\partial_A\partial_B\rho_t\ . \qquad (11.38)$$

Thus, the differential equation for the evolution of $\rho_t(x)$ takes the form of a Fokker-Planck equation. To proceed with our analysis we shall rewrite (11.38) in several different forms.

11.5.3 The current and osmotic velocities

The evolution equation (11.38) can also be rewritten as

$$\partial_t\rho_t=-\partial_A\left[\left(b^A+\frac{1}{2}\,\eta m^{AB}\partial_B\log\rho_t\right)\rho_t\right]\ . \qquad (11.39)$$

The interpretation is clear: the fact that the particles follow continuous paths implies that the probabilities are conserved locally in the N-particle configuration space, $\mathbf{X}_N$. This *continuity equation* can be written as

$$\partial_t\rho=-\partial_A\left(v^A\rho\right) \qquad (11.40)$$

where v^A, the velocity of the probability flow or *current velocity*, is

$$v^A=b^A+u^A \qquad (11.41)$$

where b^A is the drift velocity, eq.(11.24), and u^A, is the *osmotic velocity*,

$$u^A \overset{\text{def}}{=} -\eta m^{AB}\partial_B \log \rho^{1/2} \ . \tag{11.42}$$

The interpretation is straightforward: eq.(11.42) shows that the osmotic velocity u^A reflects the tendency for probability to flow down the density gradient in a diffusion process which is analogous to Brownian motion. Indeed, in Brownian motion the drift velocity b^A is the response to the gradient of an external potential while u^A is the response to the gradient of a concentration or chemical potential—the so-called *osmotic force*. The osmotic contribution to the probability flow is the actual diffusion current,

$$\rho u^a = -\frac{1}{2}\eta m^{AB}\partial_A \rho \ , \tag{11.43}$$

which can be recognized as the $3N$-dimensional configuration space version of Fick's law with a diffusion tensor given by $\eta m^{AB}/2$.[20]

Since both b^A and u^A involve gradients the current velocity for Brownian paths can be written ,

$$v^A = m^{AB}(\partial_B \phi - \bar{A}_B) \ , \tag{11.44}$$

where we introduced a new "potential,"

$$\phi = \varphi - \eta \log \rho^{1/2} \ , \tag{11.45}$$

that will be called the *phase*. (Eventually ϕ will be identified as the phase of the wave function.) Eqs.(11.43) and (11.45) show, once again, that the action of the constant η is to control the relative strength of diffusion and drift.

Next we shall rewrite the continuity equation (11.40) in yet another equivalent but very suggestive form involving functional derivatives.

11.5.4 A quick review of functional derivatives

Functional derivatives can be defined by analogy to the partial derivatives. A functional $F[\rho]$ is a function of a function, that is, a map that associates a number to a function ρ. We can think of $F[\rho]$ as a function of infinitely many variables $\rho(x) = \rho_x$ that are labeled by a continuous index x.

If $f(q) = f(q_1, q_2, \ldots q_n)$ is a function of several variables q_i labeled by a discrete index i, then small changes $q_i \to q_i + dq_i$ induce a small change df that to first order in dq_i is given by

$$df \overset{\text{def}}{=} \sum_i \frac{\partial f}{\partial q_i} dq_i \ . \tag{11.46}$$

[20] 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 the external force that is needed to balance the osmotic force (due to concentration gradients) in order to attain equilibrium. Here the osmotic velocity is the velocity associated to the actual diffusion current, eq. (11.43).

The partial derivative $\partial f/\partial q_i$ is *defined* as the coefficient of the term linear in dq_i.

Similarly, if $F[\rho] = F(\ldots\rho_x\ldots\rho_{x'}\ldots)$ is a function of infinitely many variables ρ_x labeled by a continuous index x, then a small change in the function $\rho(x) \to \rho(x)+\ \delta\rho(x)$ will induce a small change δF of the functional $F[\rho]$. To first order in $\delta\rho$ we have

$$\delta F \overset{\text{def}}{=} \textstyle\int dx \frac{\delta F}{\delta\rho(x)}\delta\rho(x) \tag{11.47}$$

where the functional derivative $\delta F/\delta\rho(x)$ is *defined* as the coefficient of the term linear in $\delta\rho(x)$.

The virtue of this approach is that it allows us to manipulate and calculate functional derivatives by just following the familiar rules of calculus such as Taylor expansions, integration by parts, etc. For example, if the functional $F[\rho]$ just returns the value of $\rho(x)$ at the point y, that is $F[\rho] = \rho_y$, then

$$\delta F = \delta\rho(y) = \textstyle\int dx \frac{\delta F}{\delta\rho(x)}\delta\rho(x) \quad \text{implies} \quad \frac{\delta\rho(y)}{\delta\rho(x)} = \delta(x-y)\ . \tag{11.48}$$

11.5.5 The evolution equation in Hamiltonian form

We can now return to rewriting the continuity equation (11.40) in an alternative form. The important observation is that a functional $\tilde{H}[\rho,\phi]$ can be found such that (11.40) can be written as

$$\partial_t\rho_t(x) = \frac{\delta\tilde{H}}{\delta\phi(x)}\ . \tag{11.49}$$

The desired $\tilde{H}$ satisfies

$$-\partial_A\left[\rho_t m^{AB}(\partial_B\phi - \bar{A}_B)\right] = \frac{\delta\tilde{H}}{\delta\phi(x)}\ , \tag{11.50}$$

which is a linear functional equation that can be easily integrated. The result is

$$\tilde{H}[\rho,\phi] = \int dx\, \frac{1}{2}\rho m^{AB}\left(\partial_A\phi - \bar{A}_A\right)\left(\partial_B\phi - \bar{A}_B\right) + F[\rho]\ , \tag{11.51}$$

where the unspecified functional $F[\rho]$ is an integration constant. (F could also depend on x and on t.) We can check that a variation $\phi \to \phi + \delta\phi$ followed by an integration by parts reproduces the correct functional derivative,

$$\begin{aligned}\delta\tilde{H} &= \int dx\, \frac{1}{2}\rho m^{AB}\left[\partial_A\delta\phi\left(\partial_B\phi - \bar{A}_B\right) + \left(\partial_A\phi - \bar{A}_A\right)\partial_B\delta\phi\right] \\ &= -\int dx\, \partial_A\left[\rho m^{AB}\left(\partial_B\phi - \bar{A}_B\right)\right]\delta\phi\ .\end{aligned}$$

The continuity equation (11.49) describes a dynamics in which the evolution of the probability density $\rho_t(x)$ is driven by two non-dynamical fields $\phi(x)$, and

$\bar{A}(x)$. This is an interesting ED in its own right but it is not QM. Indeed, a *quantum* dynamics consists in the coupled evolution of two dynamical fields: the density $\rho_t(x)$ and the phase of the wave function. This second field can be naturally introduced into ED by allowing the phase field $\phi_t(x)$ in (11.49) to become dynamical which amounts to an ED in which the constraint (11.6) is itself continuously updated at each instant in time. To complete the construction of ED we must identify the appropriate updating criterion (*e.g.*, along the lines of Chapter 10) to formulate an ED in which the phase field ϕ_t guides the evolution of ρ_t, and in return, the evolving ρ_t reacts back and induces the evolution of ϕ_t.

Remark: One might suspect that the Hamiltonian $\tilde{H}$ in (11.51) will eventually lead us to the concept of energy and this will indeed turn out to be the case. But there is something peculiar about $\tilde{H}$: the variables that define $\tilde{H}$ are probabilities and the phase fields which are both epistemic quantities. It therefore follows that in the ED approach the energy is also an epistemic concept. In ED only positions are ontic; energy is not. Surprising as this may sound, it is not an impediment to formulating laws of physics that are empirically successful.

Remark: We note that once the evolution equation is written in Hamiltonian form in terms of the phase ϕ the constant η disappears from the formalism. This means that changes in ρ arise from the combined effect of drift and diffusion and it is no longer possible to attribute any particular effect to one or the other.

The fact that it is possible to enhance or suppress the fluctuations relative to the drift to achieve the same overall evolution shows that there is a whole family of ED models that differ at the "microscopic" or sub-quantum level.

Nevertheless, as we shall see, all members of this family lead to the same "emergent" Schrödinger equation at the "macroscopic" or quantum level. The model in which fluctuations are (almost) totally suppressed is of particular interest: the system evolves along the smooth lines of probability flow. This suggests that ED includes the Bohmian or causal form of quantum mechanics as a special limiting case. (For more on this see section 11.6).

11.5.6 The future and past drift velocities

An interesting consequence of the time asymmetry, eq.(11.18), is that the drift velocities towards the future and from the past do not coincide. Let us be more specific. Equation (11.24) gives the *mean drift velocity to the future*,

$$\begin{aligned} b^A(x) &= \lim_{\Delta t\to 0^+} \frac{\left\langle x^A(t+\Delta t)\right\rangle_{x(t)} - x^A(t)}{\Delta t} \\ &= \lim_{\Delta t\to 0^+} \frac{1}{\Delta t}\int dx'\, P(x'|x)\Delta x^A \ , \end{aligned} \tag{11.52}$$

where $x = x(t)$, $x' = x(t+\Delta t)$, and $\Delta x^A = x'^A - x^A$. Note that the expectation in (11.52) is conditional on the earlier position $x = x(t)$. One can also define a *mean drift velocity from the past*,

$$b_*^A(x) = \lim_{\Delta t\to 0^+} \frac{x^A(t) - \left\langle x^A(t-\Delta t)\right\rangle_{x(t)}}{\Delta t} \tag{11.53}$$

where the expectation is conditional on the later position $x = x(t)$. Shifting the time by Δt, b_*^A can be equivalently written as

$$\begin{aligned} b_*^A(x') &= \lim_{\Delta t\to 0^+} \frac{x^A(t+\Delta t) - \left\langle x^A(t)\right\rangle_{x(t+\Delta t)}}{\Delta t} \\ &= \lim_{\Delta t\to 0^+} \frac{1}{\Delta t}\!\int dx\, P(x|x')\Delta x^A \ , \end{aligned} \tag{11.54}$$

with the same definition of Δx^A as in eq.(11.52).

The two drift velocities, towards the future b^A and from the past b_*^A, do not coincide. The connection between them was derived by Nelson in [Nelson 1966, 1985] and independently by Jaynes [Jaynes 1989]. It turns out to be a straightforward consequence of Bayes' theorem, eq.(11.18). To derive it expand $\rho_{t'}(x')$ about x in (11.18) to get

$$P(x|x') = \frac{\rho_t(x)}{\rho_{t'}(x)}\left[1 - \partial_B \log\rho_{t'}(x)\,\Delta x^B + \ldots\right] P(x'|x)\,. \tag{11.55}$$

Next multiply $b_*^A(x')$ by a smooth test function $f(x')$ and integrate,

$$\int dx'\, b_*^A(x')f(x') = \lim_{\Delta t\to 0^+} \frac{1}{\Delta t}\int dx' \int dx\, P(x|x')\Delta x^A f(x') \ . \tag{11.56}$$

On the right hand side expand $f(x')$ about x and use (11.55),

$$\begin{aligned} \lim_{\Delta t\to 0^+} \frac{1}{\Delta t}\int dx' \int dx\, P(x'|x)\frac{\rho_t(x)}{\rho_{t'}(x)}[\Delta x^A f(x) \\ - \Delta x^A\Delta x^B f(x)\partial_B \log\rho_{t'}(x) + \Delta x^A\Delta x^C\partial_C f(x) + \ldots]\,. \end{aligned} \tag{11.57}$$

Interchange the orders of integration and integrate over x' using eq.(11.26),

$$\begin{aligned} \langle\Delta x^A\Delta x^B\rangle_x &= \langle\Delta w^A\Delta w^B\rangle_x + O(\Delta t^{3/2}) \\ &= \eta m^{AB}\Delta t + O(\Delta t^{3/2}) \ , \end{aligned} \tag{11.58}$$

to get

$$\lim_{\Delta t\to 0^+} \int dx\, \frac{\rho_t(x)}{\rho_{t'}(x)}[\frac{\Delta x^A}{\Delta t} f(x) - \eta m^{AB} f(x)\partial_B \log\rho_{t'}(x) \,+ \eta m^{AB}\partial_B f(x) + \ldots] \ . \tag{11.59}$$

Next take the limit $\Delta t \to 0^+$ and note that the third term vanishes (just integrate by parts). The result is

$$\int dx\, b_*^A(x)f(x) = \int dx\, \left[b^A(x) - \eta m^{AB}\partial_B \log\rho_t(x)\right] f(x)\,. \tag{11.60}$$

Since $f(x)$ is arbitrary we get the desired relation,

$$b_*^A(x) = b^A(x) - \eta m^{AB}\partial_B \log\rho_t(x) \ . \tag{11.61}$$

Incidentally, eq.(11.61) shows that the current v^A and osmotic u^A velocities, eqs.(11.41) and (11.42), can be expressed in terms of the sum and difference of the past and future drift velocities,

$$v^A = \frac{1}{2}\left(b_*^A + b^A\right) \quad \text{and} \quad u^A = \frac{1}{2}\left(b_*^A - b^A\right) \ . \tag{11.62}$$

11.6 An alternative: Bohmian sub-quantum motion

In section 11.5.5 we remarked that in the limit $\eta \to 0$ the Brownian trajectories become as smooth as one wishes. In this section we explore an alternative way to achieve the same result. Rather than $\alpha' = \text{const}$, eq.(11.20), and Brownian trajectories, we shall set $\alpha' \propto 1/\Delta t^2$ which leads directly to the smooth sub-quantum trajectories characteristic of a Bohmian mechanics. Thus, recalling eq.(11.19), we set

$$\alpha' = \frac{1}{\eta' \Delta t^2} \quad \text{so that} \quad \alpha_n = \frac{m_n}{\eta' \Delta t^3} \ , \tag{11.63}$$

where η' is a new constant.

Our goal is to show that this choice of α' has no effect on the dynamics provided a new "drift potential" $\varphi_{\text{Bohmian}} = \varphi'$ is suitably chosen. Then, with the choice (11.63), the new transition probability (11.9) becomes

$$P(x'|x) = \frac{1}{Z} \exp\left[-\frac{1}{2\eta' \Delta t} m_{AB} \left(\frac{\Delta x^A}{\Delta t} - b'^A(x)\right)\left(\frac{\Delta x^B}{\Delta t} - b'^B(x)\right)\right] \ , \tag{11.64}$$

where we used (11.11) to define the drift velocity,

$$b'^A(x) = \frac{\langle \Delta x^A \rangle}{\Delta t} = m^{AB}\left[\partial_B \varphi'(x) - \bar{A}_B(x)\right] \ , \tag{11.65}$$

and the fluctuations Δw^A are given by

$$\langle \Delta w^A \rangle = 0 \quad \text{and} \quad \langle \Delta w^A \Delta w^B \rangle = \eta' m^{AB} \Delta t^3 \ , \tag{11.66}$$

or

$$\left\langle \left(\frac{\Delta x^A}{\Delta t} - b'^A\right)\left(\frac{\Delta x^B}{\Delta t} - b'^B\right)\right\rangle = \eta' m^{AB} \Delta t. \tag{11.67}$$

It is noteworthy that $\langle \Delta x^A \rangle \sim O(\Delta t)$ and $\Delta w^A \sim O(\Delta t^{3/2})$. This means that as $\Delta t \to 0$ the dynamics is dominated by the drift and the fluctuations become negligible. Indeed, since $\Delta w'^A \sim O(\Delta t^{3/2})$ eq.(11.23) shows that the limit

$$\lim_{\Delta t \to 0} \frac{\Delta x^A}{\Delta t} = b'^A \tag{11.68}$$

is well defined. In words: the actual velocities of the particles coincide with the expected or drift velocities. From eq.(11.65) we see that these velocities are continuous functions. Since, as we shall later see, these smooth trajectories coincide with the trajectories postulated in Bohmian mechanics, we shall call them *Bohmian trajectories* to distinguish them from the Brownian trajectories discussed in section 11.5.2.

11.6.1 The evolution equation in differential form

We wish to rewrite the evolution equation (11.16),

$$\rho_{t+\Delta t}(x') = \int dx\, P(x', t+\Delta t|x, t)\rho_t(x) \ , \tag{11.69}$$

in differential form. Since for small Δt the transition probability $P(x', t{+}\Delta t|x, t)$ is very sharply peaked at $x' = x$ we proceed as in section 11.5.2. We multiply by a smooth test function $f(x')$ and integrate over x',

$$\int dx'\, \rho_{t+\Delta t}(x')f(x') = \int dx \left[\int dx'\, P(x', t+\Delta t|x, t)f(x')\right] \rho_t(x) \ . \tag{11.70}$$

The test function $f(x')$ is assumed sufficiently smooth precisely so that it can be expanded about x. Then, dropping all terms of order higher than Δt, as $\Delta t \to 0$ the integral in the brackets is

$$\begin{aligned}[\cdots] &= \int dx'\, P(x', t+\Delta t|x, t)\left(f(x) + \frac{\partial f}{\partial x^A}(x'^A - x^A) + ...\right) \\ &= f(x) + b'^A(x)\Delta t\frac{\partial f}{\partial x^A} + \ldots \end{aligned} \tag{11.71}$$

where we used eq.(11.23). Dropping the primes on the left hand side of (11.70), substituting (11.71) into the right, and dividing by Δt, gives

$$\int dx\, \frac{1}{\Delta t}\left[\rho_{t+\Delta t}(x) - \rho_t(x)\right] f(x) = \int dx\, b'^A(x)\frac{\partial f}{\partial x^A}\rho_t(x) \ . \tag{11.72}$$

Next integrate by parts on the right and let $\Delta t \to 0$. Since the test function $f(x)$ is arbitrary, we conclude

$$\partial_t\rho_t(x) = -\partial_A[\rho_t(x)b'^A(x)] \ , \tag{11.73}$$

which is the desired evolution equation for $\rho_t(x)$ written in differential form. This is a continuity equation where the current velocity is equal to the drift velocity, $v^A = b'^A$.

Thus, whether we deal with Brownian ($\alpha' = \text{const}$) or Bohmian ($\alpha' \propto 1/\Delta t^2$) trajectories we find *the same continuity equation*

$$\partial_t\rho_t(x) = -\partial_A[\rho_t(x)v^A(x)] \quad \text{with} \quad v^A = m^{AB}\left[\partial_B\phi(x) - \bar{A}_B(x)\right] \ , \tag{11.74}$$

provided the corresponding drift potentials φ_{Bohmian} and $\varphi_{\text{Brownian}}$ are chosen such that they lead to the same *phase* field,

$$\phi = \varphi_{\text{Bohmian}} = \varphi_{\text{Brownian}} - \eta \log \rho^{1/2} \ . \tag{11.75}$$

It also follows that whether we deal with Bohmian or Brownian paths, the evolution of probabilities can be expressed in the same Hamiltonian form given in eqs.(11.49) and (11.51).

A fractional Brownian motion? — Our choices of α' led to Brownian and Bohmian paths but more general *fractional* Brownian motions [Mandelbrot Van Ness 1968] are in principle possible. Consider

$$\alpha' = \frac{1}{\eta'' \Delta t^{\gamma-1}} \quad \text{and} \quad \alpha_n = \frac{m_n}{\eta'' \Delta t^{\gamma}} \ , \tag{11.76}$$

where γ and η'' are positive constants. We will not pursue this topic further except to note that for $\gamma < 2$ the sub-quantum motion is dominated by fluctuations and the trajectories are non-differentiable, while for $\gamma > 2$ the drift dominates and velocities are well defined.

11.7 The epistemic phase space

In ED we deal with two configuration spaces. One is the *ontic configuration space* $\mathbf{X}_N = \mathbf{X} \times \mathbf{X} \times \ldots$ of all particle positions, $x = (x_1 \ldots x_N) \in \mathbf{X}_N$. The other is the *epistemic configuration space* or *e-configuration space* $\mathcal{S}$ of all normalized probabilities,

$$\mathcal{S} = \left\{ \rho \,\middle|\, \rho(x) \geq 0; \int dx \rho(x) = 1 \right\} \ . \tag{11.77}$$

To formulate the coupled dynamics of ρ and ϕ we need a framework to study trajectories in the larger space $\{\rho, \phi\}$ that we will call the *epistemic phase space* or *e-phase space.*

As we saw in chapter 10, given a manifold such as $\mathcal{S}$ its associated cotangent bundle $T^*\mathcal{S}$ is a geometric object of particular interest because it comes automatically endowed with rich symplectic and metric structures.[21] This observation leads us to identify the e-phase space $\{\rho, \phi\}$ with the cotangent bundle $T^*\mathcal{S}$ and we adopt the preservation of those structures as the criterion for updating constraints. The discussion in chapter 10 can be borrowed essentially unchanged once our notation is adapted to account for the fact that the ontic variables we now deal are continuous rather than discrete.

[21] For previous work on the geometric and symplectic structure of quantum mechanics [Kibble 1979; Heslot 1985; Anandan and Aharonov 1990; Cirelli et al. 1990; Abe 1992; Hughston 1995; Ashekar and Schilling 1998; de Gosson, Hiley 2011; Elze 2012; Reginatto and Hall 2011, 2012]; [Caticha 2019, 2021b].

Notation: vectors, covectors, etc. A point $X \in T^*\mathcal{S}$ is represented as

$$X = (\rho(x), \pi(x)) = (\rho^x, \pi_x) \ , \tag{11.78}$$

where ρ^x represents coordinates on the base manifold $\mathcal{S}$, and π_x represents some generic coordinates on the space $T^*\mathcal{S}_\rho$ that is cotangent to $\mathcal{S}$ at the point ρ. Curves in $T^*\mathcal{S}$ allow us to define vectors. Let $X = X(\lambda)$ be a curve parametrized by λ, then the vector $\bar{V}$ tangent to the curve at $X = (\rho, \pi)$ has components $d\rho^x/d\lambda$ and $d\pi_x/d\lambda$, and is written

$$\bar{V} = \frac{d}{d\lambda} = \int dx \left[\frac{d\rho^x}{d\lambda}\frac{\delta}{\delta\rho^x} + \frac{d\pi_x}{d\lambda}\frac{\delta}{\delta\pi_x}\right] \ , \tag{11.79}$$

where $\delta/\delta\rho^x$ and $\delta/\delta\pi_x$ are the basis vectors. The directional derivative of a functional $F[X]$ along the curve $X(\lambda)$ is

$$\frac{dF}{d\lambda} = \int dx \left[\frac{\delta F}{\delta\rho^x}\frac{d\rho^x}{d\lambda} + \frac{\delta F}{\delta\pi_x}\frac{d\pi_x}{d\lambda}\right] \overset{\text{def}}{=} \tilde{\nabla}F[\bar{V}] \ , \tag{11.80}$$

where $\tilde{\nabla}$ is the functional gradient in $T^*\mathcal{S}$. The tilde '$\tilde{\ }$' on $\tilde{\nabla}$ serves to distinguish the functional gradient on $T^*\mathcal{S}$ from the spatial gradient $\nabla f = \partial_a f \nabla x^a$ on $\mathbf{X}_N$. The gradient of a generic functional $F[X] = F[\rho, \pi]$ is

$$\tilde{\nabla}F = \int dx \left[\frac{\delta F}{\delta\rho^x}\tilde{\nabla}\rho^x + \frac{\delta F}{\delta\pi_x}\tilde{\nabla}\pi_x\right] \ , \tag{11.81}$$

and the action of the basis covectors $\tilde{\nabla}\rho^x$ and $\tilde{\nabla}\pi_x$ on the vector $\bar{V}$ is defined by

$$\tilde{\nabla}\rho^x[\bar{V}] = \frac{d\rho^x}{d\lambda} \quad \text{and} \quad \tilde{\nabla}\pi_x[\bar{V}] = \frac{d\pi_x}{d\lambda} \ , \tag{11.82}$$

that is,

$$\tilde{\nabla}\rho^x[\frac{\delta}{\delta\rho^{x'}}] = \delta^x_{x'} \ , \quad \tilde{\nabla}\pi_x[\frac{\delta}{\delta\pi_{x'}}] = \delta^{x'}_x \ , \quad \text{and} \quad \tilde{\nabla}\rho^x[\frac{\delta}{\delta\pi_{x'}}] = \tilde{\nabla}\pi_x[\frac{\delta}{\delta\rho^{x'}}] = 0 \ . \tag{11.83}$$

The fact that the space $\mathcal{S}$ is constrained to normalized probabilities means that the coordinates ρ^x are not independent. This technical difficulty is handled by embedding the ∞-dimensional manifold $\mathcal{S}$ in a $(\infty+1)$-dimensional manifold $\mathcal{S}^+$ where the coordinates ρ^x are unconstrained. Thus, strictly, $\tilde{\nabla}F$ is a covector on $T^*\mathcal{S}^+$, that is, $\tilde{\nabla}F \in T^*\left(T^*\mathcal{S}^+\right)_X$ and $\tilde{\nabla}\rho^x$ and $\tilde{\nabla}\pi_x$ are the corresponding basis covectors.

Instead of keeping separate track of the ρ^x and π_x coordinates it is more convenient to combine them into a single index. A point $X = (\rho, \pi)$ will then be labelled by its coordinates

$$X^{\alpha x} = (X^{1x}, X^{2x}) = (\rho^x, \pi_x) \tag{11.84}$$

where αx is a composite index: $\alpha = 1, 2$ keeps track of whether x is an upper index ($\alpha = 1$) or a lower index ($\alpha = 2$). Then eqs.(11.79-11.81) are written as

$$\bar{V} = V^{\alpha x} \frac{\delta}{\delta X^{\alpha x}} \ , \quad \text{where} \quad V^{\alpha x} = \frac{dX^{\alpha x}}{d\lambda} = \begin{bmatrix} d\rho^x/d\lambda \\ d\pi_x/d\lambda \end{bmatrix} \ , \tag{11.85}$$

$$\frac{dF}{d\lambda} = \tilde{\nabla} F[\bar{V}] = \frac{\delta F}{\delta X^{\alpha x}} V^{\alpha x} \quad \text{and} \quad \tilde{\nabla} F = \frac{\delta F}{\delta X^{\alpha x}} \tilde{\nabla} X^{\alpha x} \ , \tag{11.86}$$

where the repeated upper and lower indices indicate a summation over α and an integration over x. Once we have introduced the composite indices αx to label tensor components there is no further need to draw a distinction between ρ^x and ρ_x — these are coordinates and not the components of a vector. From now on we shall write $\rho(x) = \rho_x = \rho^x$ switching from one notation to another as convenience dictates. On the other hand, for quantities such as $\delta\rho^x$ or $d\rho^x/d\lambda$ that are the components of vectors it is appropriate to keep x as an upper index.

11.7.1 The symplectic form in ED

In classical mechanics with configuration space $\{q^i\}$ the Lagrangian $L(q, \dot{q})$ is a function on the tangent bundle while the Hamiltonian $H(q, p)$ is a function on the cotangent bundle [Arnold 1997][Souriau 1997][Schutz 1980]. A symplectic form provides a mapping from the tangent to the cotangent bundles. When a Lagrangian is given the map is defined by $p_i = \partial L/\partial \dot{q}^i$ and this automatically defines the corresponding symplectic form. In ED there is no Lagrangian so in order to define the symplectic map we must look elsewhere. The fact that the preservation of a symplectic structure must reproduce the continuity equation (11.49) leads us to identify the phase ϕ_x as the momentum canonically conjugate to ρ^x. This identification of the e-phase space $\{\rho, \phi\}$ with $T^*\mathcal{S}$ is highly non-trivial. It amounts to asserting that there is a privileged symplectic form[22]

$$\Omega = \int dx \left[\tilde{\nabla}\rho_x \otimes \tilde{\nabla}\phi_x - \tilde{\nabla}\phi_x \otimes \tilde{\nabla}\rho_x \right] \ . \tag{11.88}$$

The action of $\Omega[\cdot, \cdot]$ on two vectors $\bar{V} = d/d\lambda$ and $\bar{U} = d/d\mu$ is given by

$$\Omega[\bar{V}, \bar{U}] = \int dx \left[V^{1x} U^{2x} - V^{2x} U^{1x} \right] = \Omega_{\alpha x, \beta x'} V^{\alpha x} U^{\beta x'} \ , \tag{11.89}$$

so that the components of Ω are

$$\Omega_{\alpha x, \beta x'} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \delta(x, x') \ , \tag{11.90}$$

where $\delta(x, x') = \delta_{xx'}$ is the Dirac δ function.

[22] Alternatively, the assumption is that the phase ϕ_x transforms as the components of a locally-defined Poincare 1-form

$$\theta = \int dx \, \phi_x \tilde{\mathsf{d}} \rho^x \ , \tag{11.87}$$

(where $\tilde{\mathsf{d}}$ is the exterior derivative on $T^*\mathcal{S}^+$) and the corresponding symplectic 2-form is $\Omega = -\tilde{\mathsf{d}}\theta$. By construction Ω is locally exact ($\Omega = -\tilde{\mathsf{d}}\theta$) and closed ($\tilde{\mathsf{d}}\Omega = 0$).

11.7.2 Hamiltonian flows

Next we reproduce the ∞-dimensional $T^*\mathcal{S}^+$ analogues of the finite dimensional Hamiltonian flows studied in section 10.4. Given a vector field $\bar{V}[X]$ in e-phase space we can integrate $V^{\alpha x}[X] = dX^{\alpha x}/d\lambda$ to find its integral curves $X^{\alpha x} = X^{\alpha x}(\lambda)$. We are particularly interested in those vector fields that generate flows that preserve the symplectic structure,

$$\pounds_V \Omega = 0 \ , \tag{11.91}$$

where the Lie derivative is given by (see eq.(10.39))

$$(\pounds_V \Omega)_{\alpha x,\beta x'} = V^{\gamma x''}\tilde{\nabla}_{\gamma x''}\Omega_{\alpha x,\beta x'} + \Omega_{\gamma x'',\beta x'}\tilde{\nabla}_{\alpha x}V^{\gamma x''} + \Omega_{\alpha x,\gamma x''}\tilde{\nabla}_{\beta x'}V^{\gamma x''} \ . \tag{11.92}$$

Since by eq.(11.90) the components $\Omega_{\alpha x,\beta x'}$ are constant, $\tilde{\nabla}_{\gamma x''}\Omega_{\alpha x,\beta x'} = 0$, we rewrite $\pounds_V \Omega$ as

$$(\pounds_V \Omega)_{\alpha x,\beta x'} = \tilde{\nabla}_{\alpha x}(\Omega_{\gamma x'',\beta x'}V^{\gamma x''}) - \tilde{\nabla}_{\beta x'}(\Omega_{\gamma x'',\alpha x}V^{\gamma x''}) \ , \tag{11.93}$$

which is the exterior derivative (basically, the curl) of the covector $\Omega_{\gamma x'',\alpha x}V^{\gamma x''}$. By Poincare's lemma, requiring $\pounds_V \Omega = 0$ (a vanishing curl) implies that $\Omega_{\beta x',\alpha x}V^{\beta x'}$ is the gradient of a scalar function, which we will denote $\tilde{V}[X]$,

$$\Omega_{\beta x',\alpha x}V^{\beta x'} = \tilde{\nabla}_{\alpha x}\tilde{V} \ . \tag{11.94}$$

Using (11.90) this is more explicitly written as

$$\int dx \left[\frac{d\rho_x}{d\lambda}\tilde{\nabla}\phi_x - \frac{d\phi_x}{d\lambda}\tilde{\nabla}\rho_x\right] = \int dx \left[\frac{\delta\tilde{V}}{\delta\rho_x}\tilde{\nabla}\rho_x + \frac{\delta\tilde{V}}{\delta\phi_x}\tilde{\nabla}\phi_x\right] \ , \tag{11.95}$$

or

$$\frac{d\rho_x}{d\lambda} = \frac{\delta\tilde{V}}{\delta\phi_x} \quad \text{and} \quad \frac{d\phi_x}{d\lambda} = -\frac{\delta\tilde{V}}{\delta\rho_x} \ , \tag{11.96}$$

which are Hamilton's equations for a Hamiltonian function $\tilde{V}$. Thus $\bar{V}$ is the Hamiltonian vector vector field associated to the Hamiltonian function $\tilde{V}$.

Remark: The Hamiltonian flows that might potentially be of interest tend to be those associated to Lie groups and, in particular, those that generate symmetry transformations. Then, to each element of the Lie algebra one can associate a corresponding Hamiltonian function. This map from the Lie algebra to Hamiltonian functions is commonly called "the moment map".

From (11.89), the action of the symplectic form Ω on two Hamiltonian vector fields $\bar{V} = d/d\lambda$ and $\bar{U} = d/d\mu$ generated respectively by $\tilde{V}$ and $\tilde{U}$ is

$$\Omega[\bar{V},\bar{U}] = \int dx \left[\frac{d\rho_x}{d\lambda}\frac{d\phi_x}{d\mu} - \frac{d\phi_x}{d\lambda}\frac{d\rho_x}{d\mu}\right] \ , \tag{11.97}$$

which, using (11.96), gives

$$\Omega[\bar{V},\bar{U}] = \int dx \left[\frac{\delta\tilde{V}}{\delta\rho_x}\frac{\delta\tilde{U}}{\delta\phi_x} - \frac{\delta\tilde{V}}{\delta\phi_x}\frac{\delta\tilde{U}}{\delta\rho_x}\right] \overset{\text{def}}{=} \{\tilde{V},\tilde{U}\} \ , \tag{11.98}$$

where on the right we introduced the Poisson bracket notation.

These results are summarized by the continuum version of eqs.(10.67-10.70): **(1)** The condition for the flow generated by the vector field $V^{\alpha x}$ to preserve the symplectic structure, $\pounds_V \Omega = 0$, is that the flow be generated by a Hamiltonian function $\tilde{V}$ according to eq.(11.96) or equivalently,

$$V^{\alpha x} = \frac{dX^{\alpha x}}{d\lambda} = \{X^{\alpha x}, \tilde{V}\} \ . \qquad (11.99)$$

(2) The action of Ω on two Hamiltonian vector fields (11.98) is the Poisson bracket of the associated Hamiltonian functions,

$$\Omega[\bar{V}, \bar{U}] = \Omega_{\alpha x, \beta x'} V^{\alpha x} U^{\beta x'} = \{\tilde{V}, \tilde{U}\} \ . \qquad (11.100)$$

The ED that preserves the symplectic structure Ω and reproduces the continuity equation (11.49) is generated by the Hamiltonian functional $\tilde{H}[\rho, \phi]$ in (11.51),

$$\partial_t \rho_x = \frac{\delta \tilde{H}}{\delta \phi_x} \ , \quad \partial_t \phi_x = -\frac{\delta \tilde{H}}{\delta \rho_x} \ , \qquad (11.101)$$

and the evolution of a generic functional $f[\rho, \phi]$ is given by the Poisson bracket,

$$\partial_t f = \{f, \tilde{H}\} \ . \qquad (11.102)$$

The dynamics, however, is not yet fully determined because the integration constant $F[\rho]$ in (11.51) remains to be specified.

11.7.3 The normalization constraint

Since the particular flow that we will associate with time evolution is required to reproduce the continuity equation it will also preserve the normalization constraint,

$$\tilde{N} = 0 \quad \text{where} \quad \tilde{N} = 1 - |\rho| \quad \text{and} \quad |\rho| \overset{\text{def}}{=} \int dx \, \rho_x \ . \qquad (11.103)$$

Indeed, one can check that

$$\partial_t \tilde{N} = \{\tilde{N}, \tilde{H}\} = 0 \ . \qquad (11.104)$$

The Hamiltonian flow (11.99) generated by $\tilde{N}$ and parametrized by ν is given by the vector field

$$\bar{N} = N^{\alpha x} \frac{\delta}{\delta X^{\alpha x}} \quad \text{with} \quad N^{\alpha x} = \frac{dX^{\alpha x}}{d\nu} = \{X^{\alpha x}, \tilde{N}\} \ . \qquad (11.105)$$

More explicitly, the components are

$$N^{1x} = \frac{d\rho_x}{d\nu} = 0 \quad \text{and} \quad N^{2x} = \frac{d\phi_x}{d\nu} = 1 \ . \qquad (11.106)$$

From eq.(11.104) we see that if $\tilde{N}$ is conserved along $\bar{H}$, then $\tilde{H}$ is conserved along $\tilde{N}$,

$$\frac{d\tilde{H}}{d\nu} = \{\tilde{H}, \tilde{N}\} = 0 \ , \tag{11.107}$$

so that $\tilde{N}$ is the generator of a global gauge symmetry. Integrating (11.106) one finds the integral curves generated by $\tilde{N}$,

$$\rho_x(\nu) = \rho_x(0) \quad \text{and} \quad \phi_x(\nu) = \phi_x(0) + \nu \ . \tag{11.108}$$

This shows that the symmetry generated by $\tilde{N}$ is to shift the phase ϕ by a constant ν independent of x without otherwise changing the dynamics.

As discussed in the last chapter this gauge symmetry is the consequence of embedding the e-phase space $T^*\mathcal{S}$ of normalized probabilities into the larger embedding space $T^*\mathcal{S}^+$ of unnormalized probabilities. One consequence of introducing a superfluous coordinate is to also introduce a superfluous momentum. The extra coordinate is eliminated by imposing the constraint $\tilde{N} = 0$. The extra momentum is eliminated by declaring it unphysical, that is, shifting the phase ϕ_x by a constant, eq.(11.108), does not lead to a new state. The two states (ρ, ϕ) and $(\rho, \phi + \nu)$ are declared to be equivalent; they lie on the same "ray".

11.8 The information geometry of e-phase space

The construction of the Hamiltonian $\tilde{H}$ on e-phase space involves **three steps**. The goal of dynamics is to determine the evolution of the state (ρ_t, ϕ_t). From a given initial state (ρ_0, ϕ_0) two slightly different Hamiltonians will lead to slightly different final states, say (ρ_t, ϕ_t) or $(\rho_t + \delta\rho_t, \phi_t + \delta\phi_t)$. Will these small changes make any difference? Can we quantify the extent to which we can *distinguish* between two neighboring states? This is precisely the kind of question that metrics are designed to address. One should then expect that a suitable choice of metric will help us constrain the form of $\tilde{H}$. In this section we take the **first step** and transform the e-phase space $T^*\mathcal{S}^+$ from a manifold that is merely symplectic to a manifold that is both symplectic and Riemannian.

As discussed in chapter 10 the e-phase space is endowed with a natural metric inherited from information geometry. It is also endowed with a natural symplectic structure inherited from the entropic dynamics as expressed in the continuity equation, (11.49). We shall impose the preservation of these two structures — a Hamilton-Killing or HK flow — as the natural criterion for updating the constraints. In section 11.9 we shall implement this **second step** and identify the general form of the Hamiltonian functions $\tilde{H}$ that generate HK flows.

In ED entropic time is constructed so that time (duration) is defined by a clock provided by the system itself. In section 11.10 we take the final **third step** in the construction of $\tilde{H}$ and require that the generator of time evolution be defined in terms of the very same clock that provides the measure of time. There we impose that the Hamiltonian $\tilde{H}$ agree with (11.51) so as to reproduce the ED evolution of ρ_t.

11.8.1 The embedding space $T^*\mathcal{S}^+$

In section (10.6.2) we assigned a metric to the statistical manifold of discrete unnormalized probabilities. The generalization of the metric from finite dimensions, eqs.(10.101) and (10.102), to the infinite-dimensional case is straightforward.[23] The length element for $\mathcal{S}^+$ is[24]

$$\delta\ell^2 = g_{xx'}\delta\rho^x\delta\rho^{x'} \quad \text{with} \quad g_{xx'} = A(|\rho|)\, n_x n_{x'} + \frac{\hbar}{2\rho_x}\delta_{xx'} \ , \tag{11.109}$$

where n is a special covector which, in ρ coordinates, has components $n_x = 1$, so that

$$\delta\ell^2 = A(|\rho|)\left(\int dx\,\delta\rho^x\right)^2 + \int dx \frac{\hbar}{2\rho_x}(\delta\rho^x)^2 \ , \tag{11.110}$$

and the freedom in the function $A(|\rho|)$ reflects the flexibility in the choice of spherically symmetric embedding. The corresponding inverse tensor, see eq.(10.103), is

$$g^{xx'} = \frac{2\rho_x}{\hbar}\delta_{xx'} + C\rho_x\rho_{x'} \quad \text{where} \quad C(|\rho|) = \frac{-2A}{\hbar A|\rho| + \hbar^2/2} \ . \tag{11.111}$$

The metric structure for $T^*\mathcal{S}^+$ is obtained following the same argument that led to eqs.(10.78) and (10.104). The simplest geometry that is invariant under flow reversal and is determined by the information geometry of $\mathcal{S}^+$, which is fully described by the tensor $g_{xx'}$ and its inverse $g^{xx'}$, is given by the length element

$$\delta\tilde{\ell}^2 = G_{\alpha x,\beta x'}\delta X^{\alpha x}\delta X^{\beta x'} = g_{xx'}\delta\rho^x\delta\rho^{x'} + g^{xx'}\delta\phi_x\delta\phi_{x'} \ . \tag{11.112}$$

More explicitly, $\delta\tilde{\ell}^2$ is

$$\delta\tilde{\ell}^2 = A\left(\int dx\,\delta\rho^x\right)^2 + C\left(\int dx\,\rho_x\delta\phi_x\right)^2 + \int dx\left(\frac{\hbar}{2\rho_x}(\delta\rho^x)^2 + \frac{2\rho_x}{\hbar}\delta\phi_x^2\right) \ , \tag{11.113}$$

or

$$\delta\tilde{\ell}^2 = A|\delta\rho|^2 + C|\rho|^2\langle\delta\phi\rangle^2 + \int dx\left(\frac{\hbar}{2\rho_x}(\delta\rho^x)^2 + \frac{2\rho_x}{\hbar}\delta\phi_x^2\right) \ . \tag{11.114}$$

In 2×2 matrix form the tensor G and its inverse G^{-1} can be written as

$$[G_{xx'}] = \begin{bmatrix} g_{xx'} & 0 \\ 0 & g^{xx'} \end{bmatrix} \ , \ [G^{xx'}] = \begin{bmatrix} g^{xx'} & 0 \\ 0 & g_{xx'} \end{bmatrix} . \tag{11.115}$$

[23] The mathematics of infinite dimensional spaces is tricky. The term 'straightforward' should be qualified with some fine print to the effect that "we adopt the standard of mathematical rigor typical of theoretical physics." Ultimately the argument is justified by the fact that it leads to useful models that are empirically successful. For relevant references see [Cirelli et al 1990][Pistone Sempi 1995] and also [Jaynes 2003, appendix B].

[24] The quantities $\delta\rho^x$ are the components of a vector so in (11.109) it makes sense to keep x as an upper index and adopt Einstein's summation convention.

Using G^{-1} to raise the first index of the symplectic form $\Omega_{\alpha x,\beta x'}$,

$$[\Omega_{xx'}] = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \delta_{xx'} \quad , \tag{11.116}$$

as in eq.(10.79),

$$G^{\alpha x,\gamma x''}\Omega_{\gamma x'',\beta x'} = -J^{\alpha x}{}_{\beta x'} \ , \tag{11.117}$$

we find

$$[J^{x}{}_{x'}] = \begin{bmatrix} 0 & -g^{xx'} \\ g_{xx'} & 0 \end{bmatrix} \ . \tag{11.118}$$

And just as in the discrete case the square of the J tensor is minus the identity,

$$J^{\alpha x}{}_{\gamma x''}J^{\gamma x''}{}_{\beta x'} = -\delta^{\alpha x}_{\beta x'} = -\delta^{\alpha}_{\beta}\delta_{xx'} \quad \text{or} \quad JJ = -\mathbf{1} \ , \tag{11.119}$$

which means that J endows $T^*\mathcal{S}^+$ with a complex structure.

11.8.2 The metric induced on the e-phase space $T^*\mathcal{S}$

The e-phase space $T^*\mathcal{S}$ is obtained from the embedding space $T^*\mathcal{S}^+$ by restricting to normalized probabilities, $|\rho| = 1$, and by identifying the gauge equivalent points (ρ^x, ϕ_x) and $(\rho^x, \phi_x + \nu)$. Consider two rays defined by the neighboring points (ρ^x, ϕ_x) and (ρ'^x, ϕ'_x) with $|\rho| = |\rho'| = 1$. The distance induced on $T^*\mathcal{S}$, that is, the distance between the two neighboring rays, is defined as the shortest $T^*\mathcal{S}^+$ distance between (ρ^x, ϕ_x) and points on the ray defined by (ρ'^x, ϕ'_x). Since the $T^*\mathcal{S}^+$ distance between (ρ^x, ϕ_x) and $(\rho^x + \delta\rho^x, \phi_x + \delta\phi_x + \nu)$ is

$$\delta\tilde{\ell}^2(\nu) = g_{xx'}\delta\rho^x\delta\rho^{x'} + g^{xx'}(\delta\phi_x + \nu)(\delta\phi_{x'} + \nu) \ , \tag{11.120}$$

the metric on $T^*\mathcal{S}$ is defined by

$$\delta\tilde{s}^2 \overset{\text{def}}{=} \min_{\nu} \delta\tilde{\ell}^2 \ . \tag{11.121}$$

The value of ν that minimizes (11.120) is

$$\nu_{\min} = -\langle\delta\phi\rangle = -\int dx \rho_x \delta\phi_x \ . \tag{11.122}$$

Then the metric that measures the distance between neighboring rays on $T^*\mathcal{S}$ is obtained by substituting $\nu_{\min}$ back into (11.120), and setting $|\rho| = 1$ and $|\delta\rho| = 0$. The result is

$$\delta\tilde{s}^2 = \int dx \left[\frac{\hbar}{2\rho_x}(\delta\rho^x)^2 + \frac{2\rho_x}{\hbar}(\delta\phi_x - \langle\delta\phi\rangle)^2\right] \ . \tag{11.123}$$

As we saw in section 10.6.3 this is the *Fubini-Study metric.*

11.8.3 A simpler embedding

To avoid the inconvenience of dealing with normalized probabilities we return to the embedding space $T^*\mathcal{S}^+$. We take advantage of the fact that the $T^*\mathcal{S}$ metric (11.123) is independent of the particular choice of the *function* $A(|\rho|)$ and choose $A(|\rho|)=0$ so that the embedding spaces $\mathcal{S}^+$ and $T^*\mathcal{S}^+$ are assigned the simplest possible geometries, namely, they are flat. With this choice the $T^*\mathcal{S}^+$ metric, eq.(11.112), becomes

$$\delta\tilde{\ell}^2=\int dx\left[\frac{\hbar}{2\rho_x}(\delta\rho^x)^2+\frac{2\rho_x}{\hbar}\delta\phi_x^2\right]=G_{\alpha x,\beta x'}\delta X^{\alpha x}\delta X^{\beta x'} \tag{11.124}$$

where G and its inverse G^{-1} are

$$[G_{xx'}]=\begin{bmatrix}\frac{\hbar}{2\rho_x}\delta_{xx'} & 0\\ 0 & \frac{2\rho_x}{\hbar}\delta_{xx'}\end{bmatrix}, \tag{11.125}$$

and

$$[G^{xx'}]=\begin{bmatrix}\frac{2\rho_x}{\hbar}\delta_{xx'} & 0\\ 0 & \frac{\hbar}{2\rho_x}\delta_{xx'}\end{bmatrix}. \tag{11.126}$$

The tensor J, eq.(11.118) that defines the complex structure becomes

$$J^{\alpha x}{}_{\beta x'}=-G^{\alpha x,\gamma x''}\Omega_{\gamma x'',\beta x'}\quad\text{or}\quad [J^x{}_{x'}]=\begin{bmatrix}0 & -\frac{2\rho_x}{\hbar}\delta_{xx'}\\ \frac{\hbar}{2\rho_x}\delta_{xx'} & 0\end{bmatrix}. \tag{11.127}$$

11.8.4 Refining the choice of cotangent space

As we saw in section 10.6.4 in the finite-dimensional case the cotangent spaces that are relevant to quantum mechanics are flat n-dimensional "hypercubes" that are only locally isomorphic to the old $\mathbb{R}^n$. Here we make the analogous move for the ∞-dimensional case and we choose cotangent spaces that obey periodic boundary conditions: the coordinates (ρ^x,ϕ_x) and $(\rho^x,\phi_x+2\pi\hbar)$ label the same point in $T^*\mathcal{S}^+$.

Having thus refined our choice of cotangent spaces, the fact that $T^*\mathcal{S}^+$ is endowed with a complex structure suggests introducing complex coordinates,

$$\psi_x=\rho_x^{1/2}e^{i\phi_x/\hbar}\quad\text{and}\quad i\hbar\psi_x^*=i\hbar\rho_x^{1/2}e^{-i\phi_x/\hbar}, \tag{11.128}$$

so that a point $\Psi\in T^*\mathcal{S}^+$ has coordinates

$$\Psi^{\mu x}=\begin{pmatrix}\Psi^{1x}\\ \Psi^{2x}\end{pmatrix}=\begin{pmatrix}\psi_x\\ i\hbar\psi_x^*\end{pmatrix}, \tag{11.129}$$

where the index μ takes two values, $\mu=1,2$.

We can check that the transformation from real coordinates (ρ,ϕ) to complex coordinates $(\psi,i\hbar\psi^*)$ is indeed canonical. From (11.128) we have

$$\begin{aligned}\delta\rho_x&=\psi_x^*\delta\psi_x+\psi_x\delta\psi_x^*,\\ \delta\phi_x&=\frac{\hbar}{2i\rho_x}\left(\psi_x^*\delta\psi_x-\psi_x\delta\psi_x^*\right).\end{aligned} \tag{11.130}$$

The action of Ω on any two vectors $\bar{V}=d/d\lambda$ and $\bar{U}=d/d\mu$,

$$\Omega[\bar{U},\bar{V}]=\Omega_{\alpha x,\beta x'}U^{\alpha x}V^{\beta x'}=\int dx\,\left(\frac{d\rho_x}{d\lambda}\frac{d\phi_x}{d\mu}-\frac{d\phi_x}{d\lambda}\frac{d\rho_x}{d\mu}\right)\;,$$

transforms into

$$\Omega[\bar{U},\bar{V}]=\int dx\,\left(\frac{d\psi}{d\lambda}\frac{di\hbar\psi^*}{d\mu}-\frac{di\hbar\psi^*}{d\lambda}\frac{d\psi}{d\mu}\right)=\Omega_{\mu x,\nu x'}\delta\Psi^{\mu x}\delta\Psi^{\nu x'} \tag{11.131}$$

so that in ψ coordinates the symplectic form,

$$[\Omega_{xx'}]=\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}\delta_{xx'}\;, \tag{11.132}$$

retains the same form as (11.90).

Expressed in ψ coordinates the Hamiltonian flow generated by the normalization constraint (11.103),

$$\tilde{N}=0\quad\text{with}\quad\tilde{N}=1-\int dx\,\psi_x^*\psi_x\;, \tag{11.133}$$

and parametrized by ν is given by the vector field

$$\bar{N}=N^{\mu x}\frac{\delta}{\delta\Psi^{\mu x}}\quad\text{with}\quad N^{\mu x}=\frac{d\Psi^{\mu x}}{d\nu}=\{\Psi^{\mu x},\tilde{N}\}\;, \tag{11.134}$$

or

$$\bar{N}=\begin{pmatrix}\frac{i}{\hbar}\psi_x\\ \psi_x^*\end{pmatrix}\;. \tag{11.135}$$

Its integral curves are given by

$$\frac{d\psi_x}{d\nu}=\frac{i}{\hbar}\psi_x\quad\text{so that}\quad\psi_x(\nu)=\psi_x(0)e^{i\nu/\hbar}\;. \tag{11.136}$$

We had seen that if $\tilde{N}$ is conserved along $\bar{H}$, then $\tilde{H}$ is conserved along $\bar{N}$,

$$\frac{d\tilde{H}}{d\nu}=\{\tilde{H},\tilde{N}\}=0\;. \tag{11.137}$$

Therefore $\tilde{N}$ is the generator of a global “gauge” symmetry and the Hamiltonian $\tilde{H}$ is invariant under the transformation $\psi_x(0)\to\psi_x(\nu)$. The interpretation is that as we embed the e-phase space $T^*\mathcal{S}$ into the larger space $T^*\mathcal{S}^+$ we introduce two additional degrees of freedom. We eliminate one by imposing the constraint $\tilde{N}=0$; we eliminate the other by declaring that two states $\psi_x(0)$ and $\psi_x(\nu)$ that lie on the same ray (or gauge orbit) are equivalent in the sense that they represent the same epistemic state.

In ψ coordinates the metric on $T^*\mathcal{S}^+$, eqs.(11.124) and (11.125), becomes

$$\delta\ell^2=-2i\int dx\,\delta\psi_x\delta i\hbar\psi_x^*=\int dxdx'G_{\mu x,\nu x'}\,\delta\Psi^{\mu x}\delta\Psi^{\nu x'}\;, \tag{11.138}$$

where in matrix form the metric tensor $G_{\mu x,\nu x'}$ and its inverse $G^{\mu x,\nu x'}$ are

$$[G_{xx'}] = -i\delta_{xx'}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad [G^{xx'}] = i\delta_{xx'}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} . \tag{11.139}$$

Finally, using $G^{\mu x,\nu x'}$ to raise the first index of $\Omega_{\nu x',\gamma x''}$ gives the Ψ components of the tensor J

$$J^{\mu x}{}_{\gamma x''} \overset{\text{def}}{=} -G^{\mu x,\nu x'}\Omega_{\nu x',\gamma x''} \quad \text{or} \quad [J^{x}{}_{x''}] = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}\delta_{xx''} . \tag{11.140}$$

11.9 Hamilton-Killing flows

Our next goal will be to find those Hamiltonian flows $\bar{H}$ that also happen to preserve the metric tensor, that is, we want $\bar{H}$ to be a Killing vector. The condition for $\bar{H}$ is $\pounds_H G = 0$, or

$$(\pounds_H G)_{\mu x,\nu x'} = H^{\lambda x''}\partial_{\lambda x''}G_{\mu x,\nu x'} + G_{\lambda x'',\nu x'}\partial_{\mu x}H^{\lambda x''} + G_{\mu x,\lambda x''}\partial_{\nu x'}H^{\lambda x''} = 0 . \tag{11.141}$$

Remark: We note that while we adopt the H notation that is usually associated with evolution in the time parameter t, the vector field $\bar{H}$ might refer to any flow that preserves the normalization $\tilde{N}$, the symplectic form Ω, and the metric G.

In complex coordinates eq.(11.139) gives $\partial_{\lambda x''}G_{\mu x,\nu x'} = 0$, and the Killing equation simplifies to

$$(\pounds_H G)_{\mu x,\nu x'} = G_{\lambda x'',\nu x'}\partial_{\mu x}H^{\lambda x''} + G_{\mu x,\lambda x''}\partial_{\nu x'}H^{\lambda x''} = 0 , \tag{11.142}$$

or

$$[(\pounds_Q G)_{xx'}] = -i\begin{bmatrix} \frac{\delta H^{2x'}}{\delta\psi_x} + \frac{\delta H^{2x}}{\delta\psi_{x'}} \; ; & \frac{\delta H^{1x'}}{\delta\psi_x} + \frac{\delta H^{2x}}{\delta i\hbar\psi^*_{x'}} \\ \frac{\delta H^{2x'}}{\delta i\hbar\psi^*_x} + \frac{\delta H^{1x}}{\delta\psi_{x'}} \; ; & \frac{\delta H^{1x'}}{\delta i\hbar\psi^*_x} + \frac{\delta H^{1x}}{\delta i\hbar\psi^*_{x'}} \end{bmatrix} = 0 . \tag{11.143}$$

If we further require that $H^{\mu x}$ be a Hamiltonian flow, $\pounds_H\Omega = 0$, then we substitute

$$H^{1x} = \frac{\delta\tilde{H}}{\delta i\hbar\psi^*_x} \quad \text{and} \quad H^{2x} = -\frac{\delta\tilde{H}}{\delta\psi_x} \tag{11.144}$$

into (10.130) to get

$$\frac{\delta^2\tilde{H}}{\delta\psi_x\delta\psi_{x'}} = 0 \quad \text{and} \quad \frac{\delta^2\tilde{H}}{\delta\psi^*_x\delta\psi^*_{x'}} = 0 . \tag{11.145}$$

Therefore in order to generate a flow that preserves $\tilde{N}$, G, and Ω, the functional $\tilde{H}[\Psi,\Psi^*]$ must be gauge invariant and *linear* in both ψ and ψ^*. Therefore,

$$\tilde{H}[\psi,\psi^*] = \int dxdx'\,\psi^*_x\hat{H}_{xx'}\psi_{x'} , \tag{11.146}$$

where $\hat{H}_{xx'}$ is a possibly non-local kernel. The actual Hamilton-Killing flow is

$$\frac{d\psi_x}{dt} = H^{1x} = \frac{\delta \tilde{H}}{\delta i\hbar\psi_x^*} = \frac{1}{i\hbar}\int dx'\, \hat{H}_{xx'}\psi_{x'} \ , \tag{11.147}$$

$$\frac{di\hbar\psi_x^*}{dt} = H^{2x} = -\frac{\delta \tilde{H}}{\delta \psi_x} = -\int dx'\, \psi_{x'}^* \hat{H}_{xx'} \ . \tag{11.148}$$

Taking the complex conjugate of (11.147) and comparing with (11.148), shows that the kernel $\hat{H}_{xx'}$ is Hermitian,

$$\hat{H}_{xx'}^* = H_{x'x} \ , \tag{11.149}$$

and we can check that the corresponding Hamiltonian functionals $\tilde{H}$ are real,

$$\tilde{H}[\psi,\psi^*]^* = \tilde{H}[\psi,\psi^*] \ .$$

An example: translations — The Hamiltonian flows of interest are those associated to Lie groups and, in particular, those that generate symmetry transformations. For example, the generator of translations is total momentum. Under a spatial displacement by ε^a, a function $f(x)$ transforms as

$$f(x) \to f_\varepsilon(x) = f(x-\varepsilon) \quad \text{or} \quad \delta_\varepsilon f(x) = f_\varepsilon(x) - f(x) = -\varepsilon^a \frac{\partial f}{\partial x^a} \ . \tag{11.150}$$

The change of a functional $F[\rho,\phi]$ is

$$\delta_\varepsilon F[\rho,\phi] = \int dx \left(\frac{\delta F}{\delta \rho_x}\delta_\varepsilon \rho_x + \frac{\delta F}{\delta \phi_x}\delta_\varepsilon \phi_x \right) = \{F, \tilde{P}_a \varepsilon^a\} \tag{11.151}$$

where

$$\tilde{P}_a = \int dx\, \rho_x {\textstyle\sum_n} \frac{\partial \phi}{\partial x_n^a} = \int dx\, \rho_x \frac{\partial \phi_x}{\partial X_{\text{cm}}^a} \tag{11.152}$$

is interpreted as the expectation of the total linear momentum, and X_{cm}^a are the coordinates of the center of mass,

$$X_{\text{cm}}^a = \frac{1}{M}\sum_n m_n x_n^a \quad \text{where} \quad M = \sum_n m_n \ . \tag{11.153}$$

Transforming to complex coordinates we can check that

$$\tilde{P}_a = \int dx\, \psi^* \left(\sum_n \frac{\hbar}{i}\frac{\partial}{\partial x_n^a} \right) \psi = \int dx\, \psi^* \frac{\hbar}{i}\frac{\partial}{\partial X_{\text{cm}}^a}\psi \ , \tag{11.154}$$

and the corresponding kernel $\hat{P}_{axx'}$ is

$$\hat{P}_{axx'} = \delta_{xx'} \sum_n \frac{\hbar}{i}\frac{\partial}{\partial x_n^a} = \delta_{xx'}\frac{\hbar}{i}\frac{\partial}{\partial X_{\text{cm}}^a} \ . \tag{11.155}$$

11.10 The e-Hamiltonian

In previous sections we supplied the e-phase space $T^*\mathcal{S}^+$ with a symplectic structure, a Riemannian metric and, as a welcome by-product, also with a complex structure. Then we showed that the condition for the simplest form of dynamics — one that preserves normalization and the metric, symplectic, and complex structures — is a Hamilton-Killing flow generated by a Hamiltonian $\tilde{H}$ that is linear in both ψ and ψ^*,

$$\tilde{H}[\psi,\psi^*] = \int dx dx'\, \psi^*_x \hat{H}_{xx'} \psi_{x'} \ . \tag{11.156}$$

The last ingredient in the construction of $\tilde{H}$ is that the e-Hamiltonian that generates evolution in entropic time must be defined in terms of the same clock that provides the measure of entropic time. In other words, $\tilde{H}$ has to agree with (11.51) in order to reproduce the entropic dynamics of ρ given by the continuity eq.(11.49).

To proceed we use the identity

$$\frac{1}{2}\rho m^{AB}(\partial_A\phi - \bar{A}_A)(\partial_B\phi - \bar{A}_B) = \frac{\hbar^2}{2} m^{AB}(D_A\psi)^* D_B\psi - \frac{\hbar^2}{8\rho^2} m^{AB}\partial_A\rho\partial_B\rho \tag{11.157}$$

where

$$D_A = \partial_A - \frac{i}{\hbar}\bar{A}_A \quad \text{and} \quad \bar{A}_A(x) = \beta_n A_a(x_n) \ . \tag{11.158}$$

The proof is straightforward: just substitute $\psi = \rho^{1/2}e^{-\phi/\hbar}$ into the right hand side of (11.157). Rewriting $\tilde{H}[\rho,\Phi]$ in (11.51) in terms of ψ and ψ^* we get

$$\tilde{H}[\psi,\psi^*] = \int dx \left(-\frac{\hbar^2}{2} m^{AB}\psi^* D_A D_B \psi \right) + F'[\rho] \ . \tag{11.159}$$

where

$$F'[\rho] = F[\rho] - \int dx \frac{\hbar^2}{8\rho^2} m^{AB}\partial_A\rho\partial_B\rho \ . \tag{11.160}$$

According to (11.156), in order for $\tilde{H}[\psi,\psi^*]$ to generate an HK flow we must impose that $F'[\rho]$ be linear in both ψ and ψ^*,

$$F'[\rho] = \int dx dx'\, \psi^*_x \hat{V}_{xx'} \psi_{x'} \tag{11.161}$$

for some Hermitian kernel $\hat{V}_{xx'}$, but $F'[\rho]$ must remain independent of ϕ,

$$\frac{\delta F'[\rho]}{\delta\phi_x} = 0 \ . \tag{11.162}$$

Substituting $\psi = \rho^{1/2}e^{i\phi/\hbar}$ into (11.161) and using $\hat{V}^*_{x'x} = \hat{V}_{xx'}$ leads to

$$\frac{\delta F'}{\delta\phi_x} = \frac{2}{\hbar}\rho_x^{1/2} \int dx' \rho_{x'}^{1/2}\, \mathrm{Im}\left(\hat{V}_{xx'} e^{-i(\phi_x - \phi_{x'})/\hbar} \right) = 0 \ . \tag{11.163}$$

This equation must be satisfied for all choices of $\rho_{x'}$. Therefore, it follows that

$$\operatorname{Im}\left(\hat{V}_{xx'}e^{-i(\phi_x-\phi_{x'})/\hbar}\right)=0\ . \tag{11.164}$$

Furthermore, this last equation must in turn hold for all choices of ϕ_x and $\phi_{x'}$. Therefore, the kernel $\hat{V}_{xx'}$ must be local in x,

$$\hat{V}_{xx'}=\delta_{xx'}V_x\ , \tag{11.165}$$

where $V_x=V(x)$ is some real function.

We conclude that the Hamiltonian that generates a Hamilton-Killing flow and agrees with the ED continuity equation must be of the form

$$\tilde{H}[\psi,\psi^*]=\int dx\psi^*\left(-\frac{\hbar^2}{2}m^{AB}D_AD_B+V(x)\right)\psi\ . \tag{11.166}$$

The evolution of Ψ is given by the Hamilton equation,

$$\partial_t\psi_x=\{\psi_x,\tilde{H}\}=\frac{\delta\tilde{H}}{\delta i\hbar\psi_x^*}\ , \tag{11.167}$$

which is the Schrödinger equation,

$$i\hbar\partial_t\psi=-\frac{\hbar^2}{2}m^{AB}D_AD_B\psi+V\psi\ . \tag{11.168}$$

In more standard notation it reads

$$i\hbar\partial_t\psi=\sum\nolimits_n\frac{-\hbar^2}{2m_n}\delta^{ab}\left(\frac{\partial}{\partial x_n^a}-\frac{i}{\hbar}\beta_nA_a(x_n)\right)\left(\frac{\partial}{\partial x_n^b}-\frac{i}{\hbar}\beta_nA_b(x_n)\right)\psi+V\psi\ . \tag{11.169}$$

At this point we can finally provide the physical interpretation of the various constants introduced along the way. Since the Schrödinger equation (11.169) is the tool we use to analyze experimental data we can identify $\hbar$ with Planck's constant, m_n will be interpreted as the particles' masses, and the β_n are related to the particles' electric charges q_n (in Gaussian units) by

$$\beta_n=\frac{q_n}{c}\ . \tag{11.170}$$

For completeness we write the Hamiltonian in the (ρ,Φ) variables,

$$\begin{aligned}\tilde{H}[\rho,\Phi]=\int d^{3N}x\,\rho\Bigg[&\sum\nolimits_n\frac{\delta^{ab}}{2m_n}\left(\frac{\partial\phi}{\partial x_n^a}-\frac{q_n}{c}A_a(x_n)\right)\left(\frac{\partial\phi}{\partial x_n^b}-\frac{q_n}{c}A_b(x_n)\right)\\&+\sum\nolimits_n\frac{\hbar^2}{8m_n}\frac{\delta^{ab}}{\rho^2}\frac{\partial\rho}{\partial x_n^a}\frac{\partial\rho}{\partial x_n^b}+V(x_1\ldots x_n)\Bigg]\ .\end{aligned} \tag{11.171}$$

The Hamilton equations for ρ and ϕ are the continuity equation (11.49),

$$\partial_t\rho=\frac{\delta\tilde{H}}{\delta\phi}=-\sum\nolimits_n\frac{\partial}{\partial x_n^a}\left[\rho\frac{\delta^{ab}}{m_n}\left(\frac{\partial\phi}{\partial x_n^b}-\frac{q_n}{c}A_b(x_n)\right)\right]\ , \tag{11.172}$$

and the quantum analogue of the Hamilton-Jacobi equation,

$$\begin{aligned}\partial_t\phi = -\frac{\delta\tilde{H}}{\delta\rho} = {\textstyle\sum_n}\frac{-\delta^{ab}}{2m_n}\left(\frac{\partial\phi}{\partial x_n^a}-\frac{q_n}{c}A_a(x_n)\right)\left(\frac{\partial\phi}{\partial x_n^b}-\frac{q_n}{c}A_b(x_n)\right) \\ + {\textstyle\sum_n}\frac{\hbar^2}{2m_n}\frac{\delta^{ab}}{\rho^{1/2}}\frac{\partial^2\rho^{1/2}}{\partial x_n^a\partial x_n^b} - V(x_1\ldots x_n)\Bigg] \,. \qquad (11.173)\end{aligned}$$

To summarize: we have just shown that an ED that preserves normalization, the symplectic and the metric structures of the e-phase space $T^*\mathcal{S}^+$ leads to a linear Schrödinger equation. In particular, such an ED reproduces the Bohmian quantum potential in (11.173) with the correct coefficients $\hbar^2/2m_n$.

The action — Now that we have Hamilton's equations (11.101) one can invert the usual procedure and *construct* an action principle from which they can be derived. Define the differential

$$\delta A = \int dt \int dx \left[\left(\partial_t\rho_x - \frac{\delta\tilde{H}}{\delta\phi_x}\right)\delta\phi_x - \left(\partial_t\phi_x + \frac{\delta\tilde{H}}{\delta\rho_x}\right)\delta\rho_x\right] \qquad (11.174)$$

and then integrate to get the action,

$$A[\rho,\phi] = \int dt \left(\int dx\phi_x\partial_t\rho_x - \tilde{H}[\rho,\phi]\right) \,. \qquad (11.175)$$

By construction, imposing $\delta A = 0$ leads to eq.(11.101). Introducing the action $A[\rho,\phi]$ is a maneuver that is useful as a convenient summary of the dynamical equations and in the formal derivation of important consequences such as Noether's theorem. From the ED perspective, however, it does not appear to be particularly fundamental.

11.11 Entropic time, physical time, and time reversal

Now that the dynamics has been fully developed we revisit the question of time. The derivation of laws of physics as examples of inference led us to introduce the notion of an entropic time which includes assumptions about the concept of instant, of simultaneity, of ordering, and of duration. It is clear that entropic time is useful but is this the actual, "physical" time? The answer is an unqualified *yes*. By deriving the Schrödinger equation (from which we can obtain the classical limit) we have shown that the t that appears in the laws of physics is entropic time. Since these are the equations that we routinely use to design and calibrate our clocks we conclude that *what clocks "measure" is entropic time.* No notion of time that is in any way deeper or more "physical" is needed.

This may at first be surprising. What ED has led us to is an epistemic notion of time and one might ask how do epistemic quantities, such as probabilities, temperatures, or energies, ever get to be "measured". As we shall discuss in some detail in chapter 13, the answer is that epistemic quantities are not measured; they are inferred from those other ontic quantities, such as positions, that are actually directly accessible to observations.

Most interestingly, the entropic model automatically includes an arrow of time. The statement that the laws of physics are invariant under time reversal has nothing to do with particles travelling backwards in time. It is instead the assertion that the laws of physics exhibit a certain symmetry. For a classical system described by coordinates q and momenta p the symmetry is the statement that if $\{q_t, p_t\}$ happens to be one solution of Hamilton's equations then we can construct another solution $\{q_t^T, p_t^T\}$ where

$$q_t^T = q_{-t} \quad \text{and} \quad p_t^T = -p_{-t} \ , \tag{11.176}$$

but both solutions $\{q_t, p_t\}$ and $\{q_t^T, p_t^T\}$ describe evolution forward in time.

An alternative statement of time reversibility is the following**:** if there is one trajectory of the system that takes it from state $\{q_0, p_0\}$ at time t_0 to state $\{q_1, p_1\}$ at the later time t_1, then there is another possible trajectory that takes the system from state $\{q_1, -p_1\}$ at time t_0 to state $\{q_0, -p_0\}$ at the later time t_1. The merit of this re-statement is that it makes clear that nothing needs to travel back in time. Indeed, rather than time reversal the symmetry might be more appropriately described as momentum or motion or flow reversal.

Since ED is a Hamiltonian dynamics one can expect that similar considerations will apply to QM and indeed they do. It is straightforward to check that given one solution $\{\rho_t(x), \phi_t(x)\}$ that evolves forward in time, we can construct another solution $\{\rho_t^T(x), \phi_t^T(x)\}$ that is also evolving forward in time. The reversed solution is

$$\rho_t^T(x) = \rho_{-t}(x) \quad \text{and} \quad \phi_t^T(x) = -\phi_{-t}(x) \ . \tag{11.177}$$

These transformations constitute a symmetry — *i.e.*, the transformed $\psi_t^T(x)$ is a solution of the Schrödinger equation — provided the motion of the sources of the external potentials is also reversed, that is, the potentials $A_a(\vec{x}, t)$ and $V(x, t)$ are transformed according to

$$A_a^T(\vec{x}, t) = -A_a(\vec{x}, -t) \quad \text{and} \quad V^T(x, t) = V(x, -t) \ . \tag{11.178}$$

Expressed in terms of wave functions the time reversal transformation is

$$\psi_t^T(x) = \psi_{-t}^*(x) \ . \tag{11.179}$$

The proof that this is a symmetry is straightforward; just take the complex conjugate of (11.169), and let $t \to -t$.

Remark: In section 10.5.1 we saw that the metric structure of the e-phase space was designed so that potential time-reversal violations would be induced at the dynamical level of the Hamiltonian and not at the kinematical level of the geometry of e-phase space. The Hamiltonian (11.166) offers us an explicit example: it leads to a dynamics that can either preserve or violate the time-reversal symmetry according to whether the potentials A and V obey the extra condition (11.178) or not.

11.12 Hilbert space

The formulation of the ED of spinless particles is now complete. We note, in particular, that the notion of Hilbert spaces turned out to be unnecessary to the formulation of quantum mechanics. However, as we argued in the last chapter, the introduction of Hilbert spaces is nevertheless a very useful tool designed for the specific purpose of exploiting the full calculational advantages of linearity. The rest of this section is a straightforward adaptation of the discussion in section 10.8 from the finite-dimensional to the infinite-dimensional case.

We shall assume that the e-phase space $T^*\mathcal{S}^+$ of unnormalized probabilities is flat. Then the points $\Psi \in T^*\mathcal{S}^+$ are are also vectors and we can deploy the already available metric and symplectic tensors to construct an inner product.

The inner product — We introduce the Dirac notation to represent the wave functions ψ_x as vectors $|\psi\rangle$ in a Hilbert space. The finite-dimensional inner product given by eq.(10.154) is written in infinite dimensions as

$$\langle\psi_1|\psi_2\rangle \overset{\text{def}}{=} \frac{1}{2\hbar}\int dx\, dx'\, (\psi_{1x}, i\hbar\psi^*_{1x})\, [G_{xx'} + i\Omega_{xx'}] \begin{pmatrix}\psi_{2x'} \\ i\hbar\psi^*_{2x'}\end{pmatrix} . \tag{11.180}$$

A straightforward calculation using (11.132) and (11.139) gives

$$\langle\psi_1|\psi_2\rangle = \int dx\, \psi_1^*\psi_2 \ . \tag{11.181}$$

The map $\psi_x \leftrightarrow |\psi\rangle$ is defined by

$$|\psi\rangle = \int dx\, |x\rangle\psi_x \quad \text{where} \quad \psi_x = \langle x|\psi\rangle \ , \tag{11.182}$$

where, in this "position" representation, the vectors $\{|x\rangle\}$ form a basis that is orthogonal and complete,

$$\langle x|x'\rangle = \delta_{xx'} \quad \text{and} \quad \int dx\, |x\rangle\langle x| \ = \hat{1} \ . \tag{11.183}$$

The complex structure — The tensors Ω and G were originally meant to act on tangent vectors but now they can also act on all points $\psi \in T^*\mathcal{S}^+$. For example,

The action of the mixed tensor $J = G^{-1}\Omega$, eq.(11.140), on a point Ψ is

$$[(J\Psi)^x] = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} \begin{pmatrix}\psi_x \\ i\hbar\psi^*_x\end{pmatrix} = \begin{pmatrix} i\psi_x \\ i(i\hbar\psi_x)^* \end{pmatrix} , \tag{11.184}$$

which shows that J plays the role of multiplication by i, that is, when acting on a point ψ the action of J is represented by an operator $\hat{J}$,

$$\psi \overset{J}{\to} i\psi \quad \text{is} \quad |\psi\rangle \overset{J}{\to} \hat{J}|\psi\rangle = i|\psi\rangle \ . \tag{11.185}$$

Hermitian and unitary operators — The bilinear Hamilton functionals $\tilde{Q}[\psi,\psi^*]$ with kernel $\hat{Q}_{xx'}$ in eq.(11.146) can now be written in terms of a Hermitian operator $\hat{Q}$ and its matrix elements,

$$\tilde{Q}[\psi,\psi^*] = \langle\psi|\hat{Q}|\psi\rangle \quad \text{and} \quad \hat{Q}_{xx'} = \langle x|\hat{Q}|x'\rangle \ . \tag{11.186}$$

The corresponding Hamilton-Killing flows parametrized by λ are given by

$$i\hbar\frac{d}{d\lambda}\langle x|\psi\rangle = \langle x|\hat{Q}|\psi\rangle \quad \text{or} \quad i\hbar\frac{d}{d\lambda}|\psi\rangle = \hat{Q}|\psi\rangle \ . \tag{11.187}$$

These flows are described by unitary transformations

$$|\psi(\lambda)\rangle = \hat{U}_Q(\lambda)|\psi(0)\rangle \quad \text{where} \quad \hat{U}_Q(\lambda) = \exp\left(-\frac{i}{\hbar}\hat{Q}\lambda\right) \ . \tag{11.188}$$

Commutators — The Poisson bracket of two Hamiltonian functionals $\tilde{U}[\psi,\psi^*]$ and $\tilde{V}[\psi,\psi^*]$,

$$\{\tilde{U},\tilde{V}\} = \int dx \left(\frac{\delta\tilde{U}}{\delta\psi_x}\frac{\delta\tilde{V}}{\delta i\hbar\psi_x^*} - \frac{\delta\tilde{U}}{\delta i\hbar\psi_x^*}\frac{\delta\tilde{V}}{\delta\psi_x}\right) \ ,$$

can be written in terms of the commutator of the associated operators,then

$$\{\tilde{U},\tilde{V}\} = \frac{1}{i\hbar}\langle\psi|[\hat{U},\hat{V}]|\psi\rangle \ . \tag{11.189}$$

Thus the Poisson bracket is the expectation of the commutator.

11.13 Summary

We conclude with a summary of the main ideas.

- Ontological clarity: Particles have definite but unknown positions and follow continuous trajectories.
- ED is a dynamics of probabilities. The probability of a short step is found by maximizing entropy subject to a constraint expressed in terms of the phase field ϕ that introduces directionality and correlations, and when appropriate gauge constraints that account for external electromagnetic fields.

- Entropic time: An epistemic dynamics of probabilities inevitably leads to an epistemic notion of time. The construction of time involves the introduction of the concept of an instant, that the instants are suitably ordered, and a convenient definition of duration. By its very construction there is a natural arrow of entropic time.
- The evolution of probabilities is given by a continuity equation that is local in configuration space but leads to non-local correlations in physical space. Rewriting the continuity equation in Hamiltonian form allows us to identify the pair (ρ, ϕ) as canonically conjugate variables.
- The epistemic phase space $\{\rho, \phi\}$ has a natural symplectic geometry.
- The e-phase space is assigned a metric structure based on the information metric of the statistical manifold of probabilities plus the requirement of symmetry under flow reversal. The joint presence of symplectic and metric structures implies the existence of a complex structure.
- The introduction of wave functions ψ as complex coordinates suggests that the cotangent spaces $T^*\mathcal{S}^+_\rho$ that are relevant to quantum mechanics are "hypercubes" with opposite faces identified.
- The dynamics that preserves the symplectic and metric structures as well as the normalization of probabilities is shown to be described by a *linear* Schrödinger equation. The particular form of the Hamiltonian is determined by requiring that it reproduce the ED evolution of probabilities in entropic time.
- Since clocks are calibrated according to the Schrödinger equation, it follows that what clocks measure is entropic time. Therefore, entropic time is "physical" time.
- The ontic motion of the particles at the sub-quantum level remains unspecified. ED allows us to infer different sub-quantum motions (*i.e.*, Brownian vs. Bohmian trajectories) and they all lead to the same emergent quantum mechanics but there is no underlying ontic dynamics.

ED achieves ontological clarity by a sharp distinction between the ontic and the epistemic — positions of particles on one side and probabilities ρ and phases ϕ on the other. ED is an epistemic dynamics of probabilities and not an ontic dynamics of particles. Of course, the probabilities will allow us to infer how the particles move — but nothing in ED describes what it is that has pushed the particles around. ED is a mechanics without a mechanism. Christiaan Huygens would have been disappointed: the theory *works* but it does not *explain.* But then, the question of 'what *counts* as an explanation?' has always been a tricky one.

We can elaborate on this point from a different direction. The empirical success of ED suggests that its epistemic probabilities are in agreement with ontic features of the physical world. It is highly desirable to clarify the precise nature of this agreement. Consider, for example, a fair die. Its property of being a perfect cube is an ontic property of the die that is reflected at the epistemic level in the assignment of equal probabilities to each face of the die. In this example we see that the epistemic probabilities achieve objectivity, and therefore usefulness, by corresponding to something ontic. The situation in ED is similar except for one crucial aspect. The ED probabilities are objective and they are empirically successful. They must therefore reflect something real. However, ED is silent about what those underlying ontic symmetries might possibly be.

To the extent that ED is a dynamics of probabilities it follows that quantum mechanics is incomplete in the sense of Einstein — QM does not provide us with a completely detailed and therefore deterministic description of the system and its dynamics. On the other hand, the fact that epistemic tools such as information geometry have turned out to be so central to the derivation of the mathematical formalism of QM lends strong support to the conjecture that QM is not only incomplete, but that it is also incompletable.

The trick of embedding the e-phase space $T^*\mathcal{S}$ in the larger space $T^*\mathcal{S}^+$ that is *chosen* to be a *flat* vector space is clever but optional. It allows one to make use of the calculational advantages of linearity; it allows, for example, to rewrite QM in the form of Feynman's path integrals. This recognition that Hilbert spaces are not fundamental is one of the significant contributions of the entropic approach to our understanding of QM — another being the derivation of linearity. The distinction between Hilbert spaces being necessary in principle as opposed to merely convenient in practice is not of purely academic interest. It could be important in the search for a quantum theory that includes gravity: Shall we follow the usual approaches to quantization that proceed by replacing classical dynamical variables by an algebra of linear operators acting on some abstract space? Or, in the spirit of an entropic dynamics, shall we search for an appropriately constrained dynamics of probabilities and information geometries? [Ipek Caticha 2020]

Chapter 12

Topics in Quantum Theory*

In the Entropic Dynamics (ED) framework quantum theory is derived as an application of entropic inference. In this chapter the immediate goal is to demonstrate that the entropic approach can prove its worth through the clarification and removal of conceptual difficulties.

We will tackle three topics that are central to quantum theory: in section 12.1 we discuss the connections between linearity, the superposition principle, the single-valuedness of wavefunctions, and the quantization of charge. Second, we consider the introduction and interpretation of quantities other than position, including momentum, and the corresponding uncertainty relations. Third, we discuss the classical limit.[1]

12.1 Linearity and the superposition principle

The Schrödinger equation is linear, that is, a linear combination of solutions is a solution too. However, this *mathematical* linearity does not guarantee the *physical* linearity that is usually referred to as the superposition principle. The latter is the physical assumption that if there is one experimental setup that prepares a system in the (epistemic) state ψ_1 and there is another setup that prepares the system in the state ψ_2 then, at least in principle, it is possible to construct yet a third setup that can prepare the system in the superposition

$$\psi_3 = \alpha_1\psi_1 + \alpha_2\psi_2 \ , \tag{12.1}$$

where α_1 and α_2 are arbitrary complex numbers. Mathematical linearity refers to the fact that solutions can be expressed as sums of solutions and there is no implication that any of these solutions will necessarily describe physical situations.[2] Physical linearity on the other hand — the superposition principle —

[1] The presentation follows closely the work presented in [Caticha 2019] [Johnson Caticha 2011; Nawaz Caticha 2011]. More details can be found in [Johnson 2011; Nawaz 2012].

[2] The diffusion equation provides an illustration. Fourier series were originally invented to describe the diffusion of heat: a physical distribution of temperature, which can only

refers to the fact that the superposition of physical solutions is also a physical solution. The point to be emphasized is that the superposition principle is a physical hypothesis of wide applicability that need not, however, be universally true.

12.1.1 The single-valuedness of ψ

The question "Why should wave functions be single-valued?" has been around for a long time. In this section we shall argue that the single- or multi-valuedness of the wave functions is closely related to the question of linearity and the superposition principle.[3]

First we show that the mathematical linearity of (11.169) is not sufficient to imply the superposition principle. The idea is that even when $|\psi_1|^2 = \rho_1$ and $|\psi_2|^2 = \rho_2$ are probabilities it is not generally true that $|\psi_3|^2$ from eq.(12.1) will also be a probability. Consider moving around a closed loop Γ in configuration space. Since the phase $\phi(x)$ can be multi-valued the corresponding wave function could in principle be multi-valued too. Suppose a generic ψ changes by a phase factor,

$$\psi \to \psi' = e^{i\delta}\psi \ , \tag{12.2}$$

then the superposition ψ_3 of two wave functions ψ_1 and ψ_2 changes into

$$\psi_3 \to \psi_3' = \alpha_1 e^{i\delta_1}\psi_1 + \alpha_2 e^{i\delta_2}\psi_2 \ . \tag{12.3}$$

The problem is that even if $|\psi_1|^2 = \rho_1$ and $|\psi_2|^2 = \rho_2$ are single-valued (because they are probability densities), the quantity $|\psi_3|^2$ need not in general be single-valued. Indeed,

$$|\psi_3|^2 = |\alpha_1|^2\rho_1 + |\alpha_2|^2\rho_2 + 2\,\mathrm{Re}[\alpha_1\alpha_2^*\psi_1\psi_2^*] \ , \tag{12.4}$$

changes into

$$|\psi_3'|^2 = |\alpha_1|^2\rho_1 + |\alpha_2|^2\rho_2 + 2\,\mathrm{Re}[\alpha_1\alpha_2^* e^{i(\delta_1-\delta_2)}\psi_1\psi_2^*] \ , \tag{12.5}$$

take positive values, is expressed as a sum of sines and cosines which cannot individually represent physical distributions. Despite the unphysical nature of the individual sine and cosine components the Fourier expansion is nevertheless very useful.

[3] Our discussion parallels [Schrödinger 1938]. Schrödinger invoked time reversal invariance which was a very legitimate move back in 1938 but today it is preferable to develop an argument which does not invoke symmetries that are already known to be violated. The answer proposed in [Pauli 1939] is worthy of note. (See also [Merzbacher 1962].) Pauli proposed that admissible wave functions must form a basis for representations of the transformation group that happens to be pertinent to the problem at hand. In particular, Pauli's argument serves to discard double-valued wave functions for describing the orbital angular momentum of scalar particles. The question of single-valuedness was later revived in [Takabayashi 1952, 1983] in the context of the hydrodynamical interpretation of QM, and later rephrased by in [Wallstrom 1989, 1994] as an objection to Nelson's stochastic mechanics: are these theories equivalent to QM or do they merely reproduce a subset of its solutions? Wallstrom's objection to Nelson's stochastic mechanics is that it leads to phases and wave functions that are either both multi-valued or both single-valued. Both alternatives are unsatisfactory because on one hand QM requires single-valued wave functions, while on the other hand single-valued phases exclude states that are physically relevant (*e.g.*, states with non-zero angular momentum).

so that in general

$$|\psi_3'|^2 \neq |\psi_3|^2 \ , \tag{12.6}$$

which precludes the interpretation of $|\psi_3|^2$ as a probability. That is, even when the epistemic states ψ_1 and ψ_2 describe actual physical situations, their superpositions need not.

The problem does not arise when

$$e^{i(\delta_1-\delta_2)} = 1 \ . \tag{12.7}$$

If we were to group the wave functions into classes each characterized by its own δ then we could have a limited version of the superposition principle that applies within each class. We conclude that beyond the linearity of the Schrödinger equation we have a superselection rule that restricts the validity of the superposition principle to wave functions that belong to the same δ-class.

To find the allowed values of δ we argue as follows. It is natural to assume that if $\{\rho, \phi\}$ (at some given time t_0) is a physical state then the state with reversed momenta $\{\rho, -\phi\}$ (at the same time t_0) is an equally reasonable physical state. Basically, the idea is that if particles can be prepared to move in one direction, then they can also be prepared to move in the opposite direction. In terms of wave functions the statement is that if ψ_{t_0} is a physically allowed initial state, then so is $\psi^*_{t_0}$.[4] Next we consider a generic superposition

$$\psi_3 = \alpha_1 \psi + \alpha_2 \psi^* \ . \tag{12.8}$$

Is it physically possible to construct superpositions such as (12.8)? The answer is that while constructing ψ_3 for an arbitrary ψ might be difficult in practice there is strong empirical evidence that there exist no superselection rules to prevent us from doing so in principle. Indeed, it is easy to construct superpositions of wavepackets with momentum $\vec{p}$ and $-\vec{p}$, or superpositions of states with opposite angular momenta, $Y_{\ell m}$ and $Y_{\ell,-m}$. *We shall assume that in principle the superpositions (12.8) are physically possible.*

According to eq.(12.2) as one moves in a closed loop Γ the wave function ψ_3 will transform into

$$\psi_3' = \alpha_1 e^{i\delta} \psi + \alpha_2 e^{-i\delta} \psi^* \ , \tag{12.9}$$

and the condition (12.7) for $|\psi_3|^2$ to be single-valued is

$$e^{2i\delta} = 1 \quad \text{or} \quad e^{i\delta} = \pm 1 \ . \tag{12.10}$$

Thus, we are restricted to two discrete possibilities ±1. Since the wave functions are assumed sufficiently well behaved (continuous, differentiable, etc.) we conclude that they must be either single-valued, $e^{i\delta} = 1$, or double-valued, $e^{i\delta} = -1$.

Thus, the superposition principle appears to be valid in a sufficiently large number of cases to be a useful rule of thumb but it is restricted to either single-valued or double-valued wave functions. The argument above does not exclude

[4] We make no symmetry assumptions such as parity or time reversibility. It need not be the case that there is any symmetry that relates the time evolution of $\psi^*_{t_0}$ to that of ψ_{t_0}.

the possibility that a multi-valued wave function might describe an actual physical situation. What the argument implies is that the superposition principle would not extend to such states.

12.1.2 Charge quantization

Next we analyze the conditions for the electromagnetic gauge symmetry to be compatible with the superposition principle. We shall confine our attention to systems that are described by single-valued wave functions ($e^{i\delta}=+1$).[5] The condition for the wave function to be single-valued is

$$\Delta\frac{\phi}{\hbar}=\oint_\Gamma d\ell^A\partial_A\frac{\phi}{\hbar}=2\pi k_\Gamma\ , \tag{12.11}$$

where k_Γ is an integer that depends on the loop Γ. Under a local gauge transformation

$$A_a(\vec{x})\to A_a(\vec{x})+\partial_a\chi(\vec{x}) \tag{12.12}$$

the phase ϕ transforms according to (**??**),

$$\phi(x)\to\phi'(x)=\phi(x)+\textstyle\sum_n\frac{q_n}{c}\chi(\vec{x}_n)\ . \tag{12.13}$$

The requirement that the gauge symmetry and the superposition principle be compatible amounts to requiring that the gauge transformed states also be single-valued,

$$\Delta\frac{\phi'}{\hbar}=\oint_\Gamma d\ell^A\partial_A\frac{\phi'}{\hbar}=2\pi k'_\Gamma\ . \tag{12.14}$$

Thus, the allowed gauge transformations are restricted to functions $\chi(\vec{x})$ such that

$$\textstyle\sum_n\frac{q_n}{\hbar c}\oint_\Gamma d\ell_n^a\partial_{na}\chi(\vec{x}_n)=2\pi\Delta k_\Gamma \tag{12.15}$$

where $\Delta k_\Gamma=k'_\Gamma-k_\Gamma$ is an integer. Next consider a loop γ_n in which we follow the coordinates of the nth particle around some closed path in 3-dimensional space while all the other particles are kept fixed. Then

$$\frac{q_n}{\hbar c}\oint_{\gamma_n}d\ell_n^a\partial_{an}\chi(\vec{x}_n)=2\pi\Delta k_{\gamma_n} \tag{12.16}$$

where Δk_{γ_n} is an integer. But $\chi(\vec{x})$ and $\gamma_n=\gamma$ are just a function and a loop in 3-dimensional space, which implies that the integral on the left,

$$\oint_{\gamma_n}d\ell_n^a\partial_{an}\chi(\vec{x}_n)=\oint_\gamma d\ell^a\partial_a\chi(\vec{x})\ , \tag{12.17}$$

is independent of n. Therefore the charge q_n divided by an integer Δk_{γ_n} must be independent of n which means that q_n must be an integer multiple of some basic charge q_0. We conclude that the charges q_n are quantized.

[5] Double-valued wave functions with $e^{i\delta}=-1$ will, of course, find use in the description of spin-1/2 particles [Caticha Carrara 2019].

The issue of charge quantization is ultimately the issue of deciding which is the gauge group that generates electromagnetic interactions. We could for example decide to restrict the gauge transformations to single-valued gauge functions $\chi(\vec{x})$ so that (12.16) is trivially satisfied irrespective of the charges being quantized or not. Under such a restricted symmetry group the single-valued (or double-valued) nature of the wave function is unaffected by gauge transformations. If, on the other hand, the gauge functions $\chi(\vec{x})$ are allowed to be multi-valued, then the compatibility of the gauge transformation (12.12-12.13) with the superposition principle demands that charges be quantized.

The argument above cannot fix the value of the basic charge q_0 because it depends on the units chosen for the vector potential A_a. Indeed since the dynamical equations show q_n and A_a appearing only in the combination $q_n A_a$ we can change units by rescaling charges and potentials according to $Cq_n = q'_n$ and $A_a/C = A'_a$ so that $q_n A_a = q'_n A'_a$.[6]

Remark: A similar conclusion — that charge quantization is a reflection of the compactness of the gauge group — can be reached following an argument due to C. N. Yang [Yang 1970]. Yang's argument assumes that a Hilbert space has been established and one has access to the unitary representations of symmetry groups. Yang considers a gauge transformation

$$\Psi(x) \to \Psi(x) \exp i {\textstyle\sum_n} \frac{q_n}{c} \chi(\vec{x}_n) \ , \tag{12.18}$$

with $\chi(\vec{x})$ independent of $\vec{x}$. If the q_ns are not commensurate there is no value of χ (except 0) that makes (12.18) be the identity transformation. The gauge group — translations on the real line — would not be compact. If, on the other hand, the charges are integer multiples of a basic charge q_0, then two values of χ that differ by an integer multiple of $2\pi c/q_0$ give identical transformations and the gauge group is compact. In the present ED derivation, however, we deal with the space $T^*\mathcal{S}$ which is a complex projective space. We cannot adopt Yang's argument because a gauge transformation χ independent of $\vec{x}$ is already an identity transformation — it leads to an equivalent state in the same ray — and cannot therefore lead to any constraints on the allowed charges.

12.2 Momentum in Entropic Dynamics*

12.2.1 Uncertainty relations

See [Nawaz Caticha 2011][Bartolomeo Caticha 2016]

12.3 The classical limit*

See [Demme Caticha 2016].

[6] For conventional units such that the rescaled basic charge is $Cq_0 = e/3$ with $\alpha = e^2/\hbar c = 1/137$ the scaling factor is $C = (\alpha\hbar c)^{1/2}/3q_0$. A more natural set of units might be to set $q_0 = \hbar c$ so that the gauge functions $\chi(\vec{x})$ are angles.

12.4 Elementary applications*

See [DiFranzo 2018].

12.4.1 The free particle*

12.4.2 The double-slit experiment*

12.4.3 Tunneling*

12.4.4 Entangled particles*

Chapter 13

The quantum measurement problem*

See [Johnson Caticha 2011][Vanslette Caticha 2017][Caticha 2022].

13.1 Measuring position: amplification*

13.2 “Measuring” other observables and the Born rule*

13.3 Evading no-go theorems*

13.4 Weak measurements*

13.5 The qubit*

13.6 Contextuality*

13.7 Delayed choice experiments*

Chapter 14

Entropic Dynamics of Fermions*

Chapter 15

Entropic Dynamics of Spin*

See [Caticha Carrara 2019][Carrara Caticha 202?]

15.1 Geometric algebra*

15.1.1 Multivectors and the geometric product*

15.1.2 Spinors*

15.2 Spin and the Pauli equation*

15.3 Entangled spins*

Chapter 16

Entropic Dynamics of Bosons*

See [Caticha 2012b][Ipek Caticha 2014][Ipek et al 2018, 2020]

16.1 Boson fields*

16.2 Boson particles*

Chapter 17

Entropy V: Quantum Entropy*

See [Vanslette 2017]

17.1 Density matrices*

17.2 Ranking density matrices*

17.3 The quantum maximum entropy method*

17.4 Decoherence*

17.5 Variational approximation methods – II*

17.5.1 The variational method*

17.5.2 Quantum density functional formalism*

See [Yousefi Caticha 2022]

Chapter 18

Epilogue: Towards a Pragmatic Realism*

See [Caticha 2014a].

18.1 Background: the tensions within realism*

18.2 Pragmatic realism*

References

[**Abe 1992**] S. Abe, "Quantum-state space metric and correlations," Phys. Rev. **A 46**, 1667 (1992).

[**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).

[**Anandan Aharonov 1990**] J. Anandan and Y. Aharonov, "Geometry of Quantum Evolution," Phys. Rev. Lett. **65**, 1697-1700 (1990).

[**Arnold 1997**] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer, Berlin/Heidelberg, 1997).

[**Ashtekar Schilling 1998**] A. Ashtekar and T. A. Schilling, "Geometrical Formulation of Quantum Mechanics," in *On Einstein's Path*, ed. by A. Harvey (Springer, New York, 1998).

[**Atkinson Mitchell 1981**] C. Atkinson and A. F. S. Mitchell, "Rao's distance measure", Sankhyā **43**A, 345 (1981).

[**Ay et al 2017**] N. Ay, J. Jost, H. Vân Lê, L. Schwanchhöfer, *Information Geometry* (Springer, 2017).

[**Balasubramanian 1996**] V. Balasubramanian, "A Geometric Formulation of Occam's Razor for Inference of Parametric Distributions" (arXiv:adap-org/9601001).

[**Balasubramanian 1997**] V. Balasubramanian, "Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions", Neural Computation **9**, 349 (1997).

[**Balian 1991, 1992**] R. Balian, *From microphysics to macrophysics: methods and applications of statistical mechanics* (Vol. I and II) (Springer, Heidelberg, 1991 and 1992).

[**Balian 1999**] R. Balian, "Incomplete descriptions and relevant entropies", Am. J. Phys. **67**, 1078 (1999).

[**Balian 2014**] R. Balian, "The Entropy-Based Quantum Metric", Entropy **16**, 3878 (2014).

[**Balian Balazs 1987**] R. Balian and N. L. Balazs, "Equiprobability, Inference, and Entropy in Quantum Theory", Ann. Phys. **179,** 97 (1987).

[**Balian Veneroni 1981**] R. Balian and M. Vénéroni, "Time-dependent Variational Principle for Predicting the Expectation Value of an Observable", Phys. Rev. Lett. **47**, 1353 (1981).

[**Balian Veneroni 1987**] R. Balian and M. Vénéroni, "Incomplete Descriptions, Relevant Information, and Entropy Production in Collision Processes", Ann. Phys. **174,** 229 (1987).

[**Balian Veneroni 1988**] R. Balian and M. Vénéroni, "Static and Dynamic Variational Principles for Expectation Values of Observables", Ann. Phys. **187,** 29 (1988).

[**Ballentine 1970**] L. Ballentine, "The statistical interpretation of quantum mechanics", Rev. Mod. Phys. **42**, 358 (1970).

[**Ballentine 1986**] L. Ballentine, "Probability theory in quantum mechanics," Am. J. Phys. **54**, 883 (1986).

[**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).

[**Bartolomeo Caticha 2015**] D. Bartolomeo and A. Caticha, "Entropic Dynamics: the Schroedinger equation and its Bohmian limit", in Bayesian Inference and Maximum Entropy Methods in Science and Engineering, ed. by A.Giffin and K. Knuth, AIP Conf. Proc. **1757**, 030002 (2016) (arXiv.org:1512.09084).

[**Bartolomeo Caticha 2016**] D. Bartolomeo and A. Caticha, "Trading drift and fluctuations in entropic dynamics: quantum dynamics as an emergent universality class," J. Phys: Conf. Series **701**, 012009 (2016) (arXiv.org:1603.08469).

[**Bell 2004**] J. S. Bell, *Speakable and Unspeakable in Quantum Mechanics* (2nd. ed., Cambridge U. Press, Cambridge 2004).

[**Bennett 1982**] C. Bennett, "The thermodynamics of computation—A review", Int. J. Th. Phys. **21**, 905 (1982).

[**Bennett 2003**] C. Bennett, "Notes on Landauer's principle, reversiblecomputation, and Maxwell's demon", Studies in History and Philosophy of Modern Physics **34**, 501 (2003).

[**Blanchard et al 1986**] P. Blanchard, S. Golin and M. Serva, "Repeated mesurements in stochastic mechanics", Phys. Rev. **D34**, 3732 (1986).

[**Bohm 1952**] D. Bohm, "A Suggested Interpretation of the Quantum Theory in Terms of 'Hidden' Variables, I and II", Phys. Rev. **85**, 166 and 180 (1952).

[**Bohm Hiley 1993**] D. Bohm and B. J. Hiley, *The Undivided Universe: an ontological interpretation on quantum theory* (Routledge, New York 1993).

[**Bohr 1934**] N. Bohr, *Atomic Theory and the Description of Nature* (1934, reprinted by Ox Bow Press, Woodbridge Connecticut, 1987).

[**Bohr 1958**] N. Bohr, *Essays 1933-1957 on Atomic Physics and Human Knowledge* (1958, reprinted by Ox Bow Press, Woodbridge Connecticut, 1987).

[**Bohr 1963**] N. Bohr, *Essays 1958-1962 on Atomic Physics and Human Knowledge* (1963, reprinted by Ox Bow Press, Woodbridge Connecticut, 1987).

[**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).

[**Brillouin 1953**] L. Brillouin, "The Negentropy Principle of Information", J. Appl. Phys. **24**, 1152 (1953).

[**Brody Hughston 1997**] D. J. Brodie and L. P. Hughston, "Statistical Geometry in Quantum Mechanics," Phil. Trans. R. Soc. London **A 454**, 2445 (1998); arXiv:gr-qc/9701051.

[**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).

[**Carrara Caticha 2017**] N. Carrara and A. Caticha, "Quantum phases in entropic dynamics", in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by A. Polpo et al., Springer Proc. Math. Stat. **239**, 1 (2018); arXiv.org:1708.08977.

[**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 2009a**] A. Caticha, "From Entropic Dynamics to Quantum Theory" in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by P. Goggans and C.-Y. Chan, AIP Conf. Proc. **1193**, 48 (2009); arXiv.org/abs/0907.4335.

[**Caticha 2009b**] 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**, 20 (2010); arXiv.org: 1011.0723.

[**Caticha 2012a**] A. Caticha, "Entropic inference: some pitfalls and paradoxes we can avoid", in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by U. von Toussaint *et al.*, AIP Conf. Proc. **1553**, 200 (2013); arXiv.org/abs/1212.6967.

[**Caticha 2012b**] A. Caticha, "The entropic dynamics of relativistic quantum fields", in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by U. von Toussaint *et al.*, AIP Conf. Proc. **1553**, 176 (2013); arXiv.org/abs/1212.6946.

[**Caticha 2012c**] A. Caticha, *Entropic Inference and the Foundations of Physics*, (EBEB 2012, São Paulo, Brazil); http://www.albany.edu/physics/ACaticha-EIFP-book.pdf.

[**Caticha 2014a**] A. Caticha, "Towards an Informational Pragmatic Realism", Mind and Machines **24**, 37 (2014); arXiv.org:1412.5644.

[**Caticha 2014b**] A. Caticha, "Entropic Dynamics: an Inference Approach to Time and Quantum Theory", J. Phys: Conf Series **504** (2014) 012009; arXiv.org:1403.3822.

[**Caticha 2015a**] A. Caticha, "Entropic Dynamics", Entropy **17**, 6110-6128 (2015); arXiv.org:1509.03222.

[**Caticha 2015b**] A. Caticha, "Geometry from Information Geometry", in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, A.Giffin and K. Knuth (eds.), AIP Conf. Proc. **1757**, 030001 (2016); arXiv.org:1512.09076.

[**Caticha 2017a**] A. Caticha, "Entropic Dynamics: Mechanics without Mechanism", in *Recent Advances in Info-Metrics*, ed. by M. Chen and A. Golan; arXiv.org:1704.02663

[**Caticha 2017b**] A. Caticha, "Entropic Dynamics: Quantum Mechanics from Entropy and Information Geometry", Annalen der Physik, 1700408 (2018); https://doi.org/10.1002/andp.201700408; arXiv.org:1711.02538.

[**Caticha 2019**] A. Caticha, "The Entropic Dynamics approach to Quantum Mechanics," Entropy **21**, 943 (2019); arXiv.org:1908.04693.

[**Caticha 2021a**] A. Caticha, "Entropy, Information, and the Updating of Probabilities," Entropy **23**, 895 (2021); arXiv.org:2107.04529.

Caticha 2021b A. Caticha, "Quantum mechanics as Hamilton-Killing flows on a statistical manifold," Phys. Sci. Forum **3**, 12 (2021); arXiv:2107.08502.

[**Caticha 2022**] A. Caticha, "Entropic Dynamics and Quantum "Measurement"", MaxEnt 2022, Paris, France.

[**Caticha et al 2014**] A. Caticha, D. Bartolomeo, and M. Reginatto, "Entropic Dynamics: from entropy and information geometry to Hamiltonians and quantum mechanics", in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by A. Mohammad-Djafari and F, Barbaresco, AIP Conf. Proc. **1641**, 155 (2015); arXiv.org:1412.5629.

[**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 Carrara 2019**] A. Caticha and N. Carrara, “The entropic dynamics of spin,” arXiv:2007.15719.

[**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.

[**Caticha Golan 2014**] A. Caticha and A. Golan, “An Entropic framework for Modeling Economies”, Physica **A 408**, 149 (2014).

[**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.

[**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).

[**Caves et al 2007**] C. Caves, C. Fuchs, and R. Schack, “Subjective probability and quantum certainty”, Studies in History and Philosophy of Modern 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).

[**Chaitin 1975**] G. J. Chaitin, “A theory of program size formally identical to information theory”, J. Assoc. Comp. Mach. **22**, 329-340 (1975).

[**Chandrasekhar 1943**] S. Chandrasekhar, “Stochastic Problems in Physics and Astronomy” Rev. Mod. Phys. **15**, 1 (1943).

[**Chiribella et al 2011**] G. Chiribella, G. M. D’Ariano, and P. Perinotti, “Informational derivation of quantum theory,” Phys. Rev. **A 84**, 012311 (2011).

[**Cirelli et al 1990**] R. Cirelli, A. Manià, and L. Pizzochero, “Quantum mechanics as an infinite-dimensional Hamiltonian system with uncertainty structure: Part I and II,” J. Math. Phys. **31**, 2891 and 2898 (1990).

[**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).

[**Csiszar 1996**] I. Csiszár, “MaxEnt, Mathematics, and Information Theory”, p. 35 in *Maximum Entropy and Bayesian Methods*, K. M. Hanson and R. N. Silver (eds.) (Kluwer Academic Publishers, Netherlands 1996).

[**Csiszar 2008**] I. Csiszár, “Axiomatic Characterizations of Information Measures”, Entropy **10**, 261 (2008).

[**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).

[**DAriano 2017**] G. M. D’Ariano, “Physics without physics: the power of information-theoretical principles,” Int. J. Th. Phys. **56**, 97 (2017).

[**de Gosson HiIey 2011**] M. A. de Gosson and B. J. Hiley, “Imprints of the Quantum World in Classical Mechanics,” Found. Phys. **41**, 1415 (2011).

[**De Martino et al 1984**] S. De Martino and S. De Siena, “On Uncertainty Relations in Stochastic Mechanics”, Il Nuovo Cimento **79B**, 175 (1984).

[**Demme Caticha 2016**] A. Demme and A. Caticha, "The classical limit of entropic quantum dynamics," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by G. Verdoolaege, AIP Conf. Proc. **1853**, 090001 (2017) (arXiv.org:1612.01905).

[**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).

[**DiFranzo 2018**] S. DiFranzo, "The Entropic Dynamics Approach to the Paradigmatic Quantum Mechanical Phenomena", Ph.D. thesis, University at Albany (2018).

[**Dirac 1948**] P. A. M. Dirac, *Quantum Mechanics* (3rd edition, Oxford University Press, 1948).

[**Dohrn Guerra 1978**] D. Dohrn and F. Guerra, "Nelson's stochastic mechanics on Riemannian manifolds", Lett. Nuovo Cimento **22**, 121 (1978).

[**Doran Lasenby 2003**] C. Doran and A. Lasenby, *Geometric Algebra for Physicists* (Cambridge U.P., Cambridge UK, 2003).

[**Earman 1992**] J. Earman, *Bayes or Bust?: A Critical Examination of Bayesian Confirmation Theory* (MIT Press, Cambridge, 1992).

[**Earman Redei 1996**] J. Earman and M. Rédei, "Why ergodic theory does not explain the success of equilibrium statistical mechanics", Brit. J. Phil. Sci. **47**, 63-78 (1996).

[**Ehrenfest 1912**] P. Ehrenfest and T. Ehrenfest, *The Conceptual Foundations of the Statistical Approach in Mechanics* (Cornell U.P., Ithaca, New York, 1959).

[**Einstein 1949a**] A. Einstein, "Autobiographical Notes" in *Albert Einstein: Philosopher Scientist*, ed. by P. A. Schilpp (Open Court, La Salle, Illinois, 1949).

[**Einstein 1949b**] A. Einstein, "Reply to Criticisms" in *Albert Einstein: Philosopher Scientist*, ed. by P. A. Schilpp (Open Court, La Salle, Illinois, 1949).

[**Ellis 1985**] B. Ellis, "What science aims to do" in *Images of Science* ed. by P. Churchland and C. Hooker (U. of Chicago Press, Chicago 1985); reprinted in [Papineau 1996].

[**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).

[**Elze 2011**] H.-T. Elze, "Linear dynamics of quantum-classical hybrids," Phys. Rev. **A 85**, 052109 (2012).

[**Everett 1957**] H. Everett III, "Relative state formulation of quantum mechanics", Rev. Mod. Phys. **29**, 454 (1957).

[**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).

[**Fine 1996**] A. Fine, *The Shaky Game – Einstein Realism and the Quantum Theory* (University of Chicago Press, Chicago 1996)

[**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).

[**Golan 2018**] A. Golan, *Foundations of Info-Metrics: Modeling, Inference, and Imperfect Information* (Oxford U. P., New York, 2018).

[**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).

[**Godfrey-Smith 2003**] P. Godfrey-Smith, *Theory and Reality* (U. of Chicago Press, Chicago 2003).

[**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 et al 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).

[**Grad 1967**] H. Grad, "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).

[Guillemin Sternberg 1984] V. Guillemin and S. Sternberg, *Symplectic techniques in physics* (Cambridge U. Press, Cambridge 1984).

[Hacking 2001] I. Hacking, *An Introduction to Probability and Inductive Logic* (Cambridge U. Press, Cambridge 2001).

[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).

[Hall Reginatto 2016] M. J. W. Hall and M. Reginatto, *Ensembles in Configuration Space* (Springer, Switzerland, 2016).

[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).

[**Heisenberg 1958**] W. Heisenberg, *Physics and Philosophy. The Revolution in Modern Science* (Harper, New York, 1958).

[**Hempel 1967**] C. G. Hempel, "The white shoe: No red herring", Brit. J. Phil. Sci. **18**, 239 (1967).

[**Hermann 1965**] R. Hermann, "Remarks on the Geometric Nature of Quantum Phase Space," J. Math. Phys. **6**, 1768 (1965).

[**Heslot 1985**] A. Heslot, "Quantum mechanics as a classical theory," Phys. Rev. **D31**, 1341-1348 (1985).

[**Hestenes 1966**] D. Hestenes, *Space-Time Algebra* (Gordon and Breach, New York, 1966; 2nd ed. Springer, Switzerland, 2015).

[**Hestenes Sobczyk 1984**] D. Hestenes and G. Sobczyk, *Clifford Algebra to Geometric Calculus* (Reidel, Dordrecht, 1984).

[**Holland 1993**] P. R. Holland, *The quantum Theory of Motion* (Cambridge U. Press, Cambridge 1993).

[**Howard 1985**] D. Howard, "Einstein on locality and separability," Stud. Hist. Phil. **16**, 171-201 (1985).

[**Howard 2004**] D. Howard, "Who invented the "Copenhagen Interpretation"? A study in mythology", Philosophy of Science **71**, 669 (2004).

[**Howson Urbach 1993**] C. Howson and P. Urbach, *Scientific Reasoning, the Bayesian Approach* (Open Court, Chicago 1993).

[**Hughston 1995**] L. P. Hughston, "Geometric aspects of quantum mechanics," in *Twistor Theory*, ed. by S. A. Huggett (Marcel Dekker, New York, 1995).

[**Ipek Caticha 2014**] S. Ipek and A. Caticha, "Entropic Quantization of Scalar Fields", in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by A. Mohammad-Djafari and F. Barbaresco, AIP Conf. Proc. **1641**, 345 (2015); arXiv.org:1412.5637.

[**Ipek Caticha 2016**] S. Ipek and A. Caticha, "Relational Entropic Dynamics of Many Particles", in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by A.Giffin and K. Knuth, AIP Conf. Proc. **1757**, 030003 (2016); arXiv.org:1601.01901.

[**Ipek et al 2018**] S. Ipek, M. Abedi, and A. Caticha, "Entropic Dynamics: Reconstructing Quantum Field Theory in Curved Spacetime", Class. Quantum Grav. **36**, 205013 (2019); arXiv:1803.07493 [gr-qc].

[**Ipek Caticha 2020**] S. Ipek and A.Caticha, "The Entropic Dynamics of Quantum Scalar Fields coupled to Gravity," Symmetry **12**, 1324 (2020); arXiv: 2006.05036.

[**Ipek 2021**] S. Ipek, *The Entropic Dynamics of Relativistic Quantum Fields in Curved Spacetime*, Ph.D. Thesis, University at Albany, State University of New York, 2021; arXiv:2105.07042 [gr-qc].

[**Jaeger 2009**] G. Jaeger, *Entanglement, Information, and the Interpretation of Quantum Mechanics* (Springer, Berlin 2009).

[**James 1897**] W. James, *The Will to Believe* (1897, reprinted by Dover, New York 1956).

[**James 1907**] W. James, *Pragmatism* (1907, reprinted by Dover, 1995).

[**James 1911**] W. James, *The Meaning of Truth* (1911, reprinted by 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).

[**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 (1957).

[**Jaynes 1957c**] E. T. Jaynes, "Information Theory and Statistical Mechanics. II", Phys. Rev. **108**, 171 (1957).

[**Jaynes 1963**] E. T. Jaynes, "Information Theory and Statistical Mechanics," in *Statistical Physics*, *Brandeis Lectures in Theoretical Physics*, K. Ford (ed.), Vol. 3, p.181 (Benjamin, New York, 1963).

[**Jaynes 1965**] E. T. Jaynes, "Gibbs vs. Boltzmann Entropies", Am. J. Phys. **33**, 391 (1965).

[**Jaynes 1968**] E. T. Jaynes, "Prior Probabilities", IEEE Trans. on Systems Science and Cybernetics **SSC-4**, 227 (1968) and at http://bayes.wustl.edu.

[**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 1980**] E. T. Jaynes, “The Minimum Entropy Production Principle”, Ann. Rev. Phys. Chem. 31, 579 (1980) 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 1986**] E. T. Jaynes, “Predictive Statistical Mechanics,” in *Frontiers of Nonequilibrium Statistical Physics*, G.T. Moore and M.O. Scully (eds.) (Plenum Press, New York, 1986) 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**] E. T. Jaynes, “Macroscopic Prediction”, pp. 254-269 in *Complex Systems|Operational Approaches in Neurobiology, Physics, and Computers*, ed. by H. Haken (Springer, Berlin, 1985) 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) and at http://bayes.wustl.edu.

[**Jaynes 1990**] E. T. Jaynes, “Probability in Quantum Theory”, in *Complexity, Entropy and the Physics of Information*, ed. by W. H Zurek (Addison-Welsey, Reading MA, 1990) and at http://bayes.wustl.edu.

[**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).

[**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).

[**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**, 130 (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. **1443**, 104 (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).

[**Khinchin 1949**] A. I. Khinchin, *Mathematical Foundations of Statistical Mechanics* (Dover, New York, 1949).

[**Kibble 1979**] T. W. B. Kibble,"Geometrization of Quantum Mechanics," Comm. Math. Phys. **65**, 189-201 (1979).

[**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).

[**Knuth 2006**] K. H. Knuth, "Valuations on lattices and their application to information theory," Proceedings of the 2006 IEEE World Congress on Computational Intelligence (IEEE WCCI 2006) (doi: 10.1109/FUZZY.2006.1681717).

[**Knuth Skilling 2012**] K. H. Knuth, J. Skilling, "Foundations of Inference," Axioms **1**, 38-73 (2012).

[**Kolmogorov 1965**] A. N. Kolmogorov, "Three approaches to the quantitative definition of information", Problems Inform. Transmission **1**, 4-7 (1965).

[**Koopman 1955**] B. O. Koopman, “Quantum Theory and the Foundations of Probability”, Proc. Symp. Appl. Math. Vol. **VII**, 97-102 (1955).

[**Kullback 1959**] S. Kullback, *Information Theory and Statistics* (Wiley, New York 1959).

[**Landauer 1961**] R. Landauer, “Information is Physical”, Physics Today, May 1991, 23.

[**Landau Lifshitz 1977a**] L. D. Landau and E. M. Lifshitz, *Statistical Physics* (Pergamon, New York 1977).

[**Landau Lifshitz 1977b**] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics* (3rd edition, Pergamon, New York 1977).

[**Landau Lifshitz 1993**] L. D. Landau and E. M. Lifshitz, *Mechanics* (Butterworth, Oxford 1993).

[**Lanczos 1970**] C. Lanczos, *The Variational Principles of Mechanics* (4th edition, Dover, New York 1986).

[**Lebowitz 1993**] J. Lebowitz, “Boltzmann’s entropy and time’s arrow”, Physics Today, September 1993, 32.

[**Lebowitz 1999**] J. Lebowitz, “Statistical mechanics: a selective review of two central issues”, Rev. Mod. Phys. **71**, S346 (1999).

[**Lee and Presse 2012**] J. Lee and S. Pressé, “Microcanonical origin of the maximum entropy principle for open systems”, Phys. Rev. E **86**, 041126 (2012).

[**Leifer 2014**] M. S. Leifer, “Is the quantum state real? An extended review of ψ-ontology theorems”, Quanta **3**, 67 (2014); arXiv.org: 1409.1570.

[**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).

[**Mackey 1989**] M. C. Mackey, “The dynamic origin of increasing entropy”, Rev. Mod. Phys. **61**, 981 (1989).

[**Mandelbrot Van Ness 1968**] B. B. Mandelbrot and J. W. Van Ness, “Fractional Brownian motions, fractional noises, and applications,” SIAM Review **10**, 422 (1968).

[**Marchildon 2004**] L. Marchildon, "Why Should We Interpret Quantum Mechanics?", Found. Phys. **34**, 1453 (1998).

[**McDuff Salamon 2017**] D. McDuff and D. Salamon, *Introduction to Symplectic Topology* (Oxford U.P., Oxford, 2017).

[**Mehra 1998**] J. Mehra, "Josiah Willard Gibbs and the Foundations of Statistical Mechanics", Found. Phys. **28**, 1785 (1998).

[**Mehrafarin 2004**] M. Mehrafarin, "Quantum mechanics from two physical postulates," Int. J. Theor. Phys., **44**, 429 (2005); arXiv:quant-ph/0402153.

[**Merzbacher 1962**] E. Merzbacher, "Single Valuedness of Wave Functions", Am. J. Phys. **30**, 237 (1962).

[**Molitor 2015**] M. Molitor, "On the relation between geometrical quantum mechanics and information geometry," J. Geom. Mech. **7**, 169 (2015).

[**Myung et al 2000**] I. J. Myung, V. Balasubramanian, M. A. Pitt, "Counting probability distributions: Differential geometry and model selection" Proc. Nat. Acad. Sci. **97**, 11170–11175 (2000).

[**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", in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by K. Knuth *et al.*, AIP Conf. Proc. **1443**, 112 (2012) (arXiv:1108.2629).

[**Nawaz et al 2016**] S. Nawaz, M. Abedi, and A. Caticha, "Entropic Dynamics on Curved Spaces", in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by A.Giffin and K. Knuth, AIP Conf. Proc. **1757**, 030004 (2016) (arXiv.org:1601.01708).

[**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. P., Princeton 1967; 2nd ed. 2001, http://www.math.princeton.edu/nelson/books.html).).

[**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).

[**von Neumann 1955**] J. von Neumann, *Mathematical Foundations of Quantum Mechanics* (Princeton University Press, 1955).

[**Newton 1693**] Isaac Newton's third letter to Bentley, February 25, 1693 in *Isaac Newton's papers and letters on Natural Philosophy and related documents*, ed. by I. B. Cohen (Cambridge, 1958), p. 302.

[**Norton 2011**] J. D. Norton, "Waiting for Landauer", Studies in History and Philosophy of Modern Physics **36**, 184 (2011).

[**Norton 2013**] J. D. Norton, "The End of the Thermodynamics of Computation: A No-Go Result", Philosophy of Science **80**, 1182 (2013).

[**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).

[**de la Peña and Cetto 2014**] L. de la Peña and A.M. Cetto, *The Emerging Quantum: The Physics Behind Quantum Mechanics* (Springer, 2014).

[**Peres 1993**] A. Peres, *Quantum Theory: Concepts and Methods* (Kluwer, Dordrecht 1993).

[**Pistone Sempi 1995**] An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one", Ann. Statist. **23**, 1543–1561 (1995).

[**Plastino and Plastino 1994**] A. R. Plastino and A. Plastino, "From Gibbs microcanonical ensemble to Tsallis generalized canonical distribution", Phys. Lett. **A193**, 140 (1994).

[**Presse et al 2013**] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill, "Nonadditive Entropies Yield Probability Distributions with Biases not Warranted by the Data", Phys. Rev. Lett. **111**, 180604 (2013).

[**Price 1996**] H. Price, *Time's Arrow and Archimedes' Point* (Oxford U. Press, Oxford 1996).

[**Pusey et al 2012**] M. F. Pusey, J. Barrett, and T. Rudolph, "On the reality of the quantum state," Nature Physics **8**, 475-478 (2012); arXiv:1111.3328.

[**Putnam 1975**] H. Putnam, *Mathematics, Matter, and Method*, Vol. 1 (Cambridge U. Press, Cambridge 1975).

[**Putnam 1979**] H. Putnam, "How to be an internal realist and a transcendental idealist (at the same time)" in *Language Logic and Philosophy*, Proc. of the 4th International Wittgenstein Symposium (Kirchberg/Wechsel, Austria 1979).

[**Putnam 1981**] H. Putnam, *Reason, Truth and History* (Cambridge U. Press, Cambridge 1981).

[**Putnam 1987**] H. Putnam, *The Many Faces of Realism* (Open Court, LaSalle, Illinois 1987).

[**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).

[**Ramsey**] F. P. Ramsey, *Philosophical Papers*, ed. by D. H. Mellor (Cambridge U.P., Cambridge, 1990).

[**Reginatto Hall 2011**] M. Reginatto and M.J.W. Hall, "Quantum theory from the geometry of evolving probabilities," AIP Conf. Proc. **1443**, 96 (2012); arXiv:1108.5601.

[**Reginatto Hall 2012**] M. Reginatto and M.J.W. Hall, "Information geometry, dynamics and discrete quantum mechanics," AIP Conf. Proc. **1553**, 246 (2013); arXiv:1207.6718.

[**Reginatto 2013**] M. Reginatto, "From information to quanta: A derivation of the geometric formulation of quantum theory from information geometry," arXiv:1312.0429.

[**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).

[**Riesz 1958**] M. Riesz, *Clifford Numbers and Spinors* (The Institute for Fluid Dynamics and Applied Mathematics, Lecture Series No.38, U. of Maryland, 1958).

[**Rissanen 1978**] J. Rissanen, "Modeling by shortest data description", Automatica **14**, 465 (1978).

[**Rissanen 1986**] J. Rissanen, "Stochastic complexity and modeling", Ann. Stat. **14**, 1080 (1986).

[**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*, W. T. Grandy Jr. and L. H. Schick (eds.) (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*, R. L. Fry (ed.) 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*, K. Knuth *et al.* (eds.) AIP Conf. Proc. Vol. **803**, 80 (2006) (omega.albany.edu:8008/CIC/me05.pdf).

[**Rovelli 1996**] C. Rovelli, "Relational Quantum Mechanics", Int. J. Theor. Phys. **35**, 1637 (1996); arXiv:9609002[quant-ph].

[**Rozanov 1977**] Y. A. Rozanov, *Probability Theory, A Concise Course* (Dover, New York 1977).

[**Ruppeiner 1979**] G. Ruppeiner, "Thermodynamics: a Riemannian geometric model", Phys. Rev. **A 20**, 1608 (1979).

[**Ruppeiner 1995**] G. Ruppeiner, "Riemannian geometry in thermodynamic fluctuation theory", Rev. Mod. Phys. **63**, 605 (1995).

[**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. Schrödinger, "About the Heisenberg Uncertainty Relation", Sitzungberichten der Preussischen Akademie der Wissenschaften (Phys. Math. Kasse) **19**, 296 (1930); English translation by A. Agelow and M. Batoni: arxiv.org/abs/quant-ph/9903100.

[**Schrodinger 1938**] E. Schrödinger, "The multivaluedness of the wave function," Ann. Phys. **32**, 49 (1938).

[**Schutz 1980**] B. Schutz, *Geometrical Methods of Mathematical Physics* (Cambridge U.P., UK, 1980).

[**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, 623 (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).

[**Shore Johnson 1981**] J. E. Shore and R. W. Johnson, "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*, pp. 173-187, ed. by G. J. Erickson and C. R. Smith (Kluwer, Dordrecht 1988).

[**Skilling 1989**] J. Skilling, "Classic Maximum Entropy", *Maximum Entropy and Bayesian Methods*, pp. 45-52, ed. by J. Skilling (Kluwer, Dordrecht 1989).

[**Skilling 1990**] J. Skilling, "Quantified Maximum Entropy", *Maximum Entropy and Bayesian Methods*, pp. 341-350, ed. by P. F. Fougère (Kluwer, Dordrecht 1990).

[**Smolin 1986a**] L. Smolin, "On the nature of quantum fluctuations and their relation to gravitation and the principle of inertia", Class. Quantum Grav. **3**, 347 (1986).

[**Smolin 1986b**] L. Smolin, "Quantum fluctuations and inertia", 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).

[**Solomonov 1964**] R. Solomonov, "A formal theory of inductive inference", Information and Control **7**, 1-22 and 224-254 (1964).

[**Souriau 1997**] J.-M. Souriau, *Structure of Dynamical Systems – A Symplectic View of Physics*, translation by C.H. Cushman-deVries (Birkhäuser, Boston 1997).

[**Spekkens 2005**] R. W. Spekkens, "Contextuality for preparations, transformations and unsharp measurements", *Phys. Rev. A* ***2005***, *71*, 052108; arXiv:quant-ph/0406166.

[**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).

[**Takabayasi 1952**] T. Takabayasi, "On the Formulation of Quantum Mechanics associated with Classical Pictures," Prog. Theor. Phys. **8**, 143 (1952).

[**Takabayasi 1983**] T. Takabayasi, "Vortex, Spin and Triad for Quantum Mechanics of Spinning Particle," Prog. Theor. Phys. **70**, 1 (1983).

[**'t Hooft 1999**] G. 't Hooft, "Quantum Gravity as a Dissipative Deterministic System", Class. Quant. Grav. **16**, 3263 (1999) (arXiv:gr-qc/9903084).

[**ter Haar 1955**] D. ter Haar, "Foundations of Statistical Mechanics", Rev. Mod. Phys. **27**, 289 (1955).

[**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 Horn 2003**] K. Van Horn, “Constructing a Logic of Plausible Inference: a Guide to Cox’s Theorem”, Int. J. Approx. Reasoning **34**, 3 (2003).

[**Vanslette 2017**] K. Vanslette, “Entropic Updating or Probabilities and Density Matrices”, Entropy **19**, 664 (2017).

[**Vanslette Caticha 2016**] K. Vanslette and A. Caticha, “Quantum measurement and weak values in entropic quantum dynamics,” in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by G. Verdoolaege, AIP Conf. Proc. **1853**, 090003 (2017) (arXiv.org:1701.00781).

[**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 2001**] J. Uffink, “Bluff Your Way in the Second Law of Thermodynamics”, Studies in History and Philosophy of Modern Physics **32**(3), 305 (2001).

[**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).

[**Uffink 2006**] J. Uffink, "Compendium of the foundations of classical statistical physics" in *Handbook for Philosophy of Physics*, J. Butterfield and J. Earman (eds) (2006).

[**Uffink 2009**] J. Uffink, "Boltzmann's H-theorem, its discontents, and the birth of statistical mechanics", Studies in History and Philosophy of Modern Physics **40**, 174 (2009).

[**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 Fraasen 1997**] B. C. van Fraasen, "Structure and Perspective: Philosophical Perplexity and Paradox", in *Logic and Scientific Methods*, p. 511, ed. by M. L. Dalla Chiara *et al.* (Kluwer, Netherlands 1997).

[**van Fraasen 2006a**] B. C. van Fraasen, "Structure: Its Shadow and Substance", Brit. J. Phil. Sci. **57**, 275 (2006).

[**van Fraasen 2006b**] B. C. van Fraasen, "Representation: The Problem for Structuralism", Philosophy of Science **73**, 536 (2006).

[**Vanslette 2017**] K. Vanslette, "Entropic Updating of Probabilities and Density Matrices", Entropy **19**, 664 (2017); arXiv:1710.09373.

[**Vanslette Caticha 2017**] K. Vanslette and A. Caticha, "Quantum measurement and weak values in entropic quantum dynamics," AIP Conf. Proc. **1853**, 090003 (2017); arXiv:1701.00781.

[**von Toussaint 2011**] U. von Toussaint, "Bayesian inference in physics", Rev. Mod. Phys. **83**, 943 (2011).

[**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).

[**Wehrl 1978**] A. Wehrl, “General properties of entropy”, Rev. Mod. Phys. **50**, 221 (1978).

[**Weinhold 1975**] F. Weinhold, “Metric geometry of equilibrium thermodynamics”, J. Chem. Phys. **63**, 2479 (1975).

[**Weinhold 1976**] F. Weinhold, “Geometry and thermodynamics”, Phys. Today **29**, No. 3, 23 (1976).

[**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 23**, 357 (1981).

[**Yang 1970**] C. N. Yang, “Charge Quantization, Compactness of the Gauge Group, and Flux Quantization,” Phys. Rev. **D1**, 2360 (1970).

[**Yousefi Caticha 2021**] A. Yousefi and A. Caticha, “An entropic approach to classical Density Functional Theory,” Phys. Sci.Forum **3**, 13 (2021); arXiv:2108.01594.

[**Zeh 2002**] H. D. Zeh, “The Wave Function: It or Bit?” (arXiv.org/abs/quant-ph/0204088).

[**Zeh 2007**] H. D. Zeh, *The Physical Basis of the Direction of Time* (5th edition, Springer, Berlin 2007).

[**Zeh 2016**] H. D. Zeh, “The strange (hi)story of particles and waves”, Z. Naturf. A **71**, 195 (2016); revised version: arXiv:1304.1003v23.

[**Zeilinger 1999**] A. Zeillinger, “A Foundational Principle for Quantum Mechanics”, Found. Phys. **29**, 631 (1999).

[**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).