Information theory and maximum entropy principles in non-equilibrium statistical physics

Chliamovitch, Gregor

Université de Genève — Faculté des sciences

Département d'informatique — Professeur Bastien Chopard
Département de physique théorique — Professeur Peter Wittwer

Information Theory and Maximum Entropy Principles in Non-Equilibrium Statistical Physics

THÈSE

présentée à la Faculté des sciences de l'Université de Genève pour obtenir le grade de Docteur ès sciences, mention interdisciplinaire informatique et physique

par

Gregor Chliamovitch

de Genève (GE)

Thèse N° 5078

Genève
Atelier d'impression Repromail
2017

UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES

Doctorat ès Sciences
Mention interdisciplinaire

Thèse de *Monsieur Gregor CHLIAMOVITCH*

intitulée :

"Information Theory and Maximum Entropy Principles in Non-Equilibrium Statistical Physics"

La Faculté des sciences, sur le préavis de Monsieur B. CHOPARD, professeur ordinaire et directeur de thèse (Département d'informatique), Monsieur P. WITTWER, professeur titulaire et codirecteur de thèse (Département de physique théorique), Monsieur A. DUPUIS, docteur (Département d'informatique) et Monsieur B. CESSAC, docteur (Institut National de la Recherche en Informatique et en Automatique, Université de Sophia Antipolis, France), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 30 mai 2017

Thèse - 5078 -

Le Doyen

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

Contents

Remerciements

Le travail que vous tenez entre les mains représente le produit final, sinon fini, de quatre années de vie et de labeur. Quant au labeur, ce fut l'occasion de voir de l'intérieur une collaboration scientifique transnationale. C'est ainsi un plaisir de saluer ici tous mes collègues du projet Sophocles dont les idées, en relief ou en creux, ont contribué à donner forme à ma recherche.

Parmi ceux-ci, les Genevois ont naturellement joué un rôle prépondérant. Je pense en premier lieu à mon directeur de thèse, Bastien Chopard, qu'il me faut remercier tout particulièrement à la fois pour sa supervision active et pour avoir toléré que je m'écarte assez largement des plans initiaux. Je pense aussi à Alexandre Dupuis, qui a bien voulu me témoigner son soutien scientifique et moral durant ces années, ainsi qu'à Anton Golub avec lequel nous avons échangé idées, grognements et réconfort mutuel.

Je remercie aussi ici Peter Wittwer, qui a bien voulu endosser la co-tutelle, et Bruno Cessac qui a accepté de faire partie du jury de thèse. Tous deux m'ont adressé de précieux commentaires quant au contenu du manuscrit. Enfin, je n'oublie pas que cette thèse n'aurait pas vu le jour sans Michel Droz, qui s'est substitué au bienveillant Hasard pour me mettre le pied à l'étrier.

Quant à la vie, il m'a fallu apprivoiser un à un les membres du SPC. Bien que nos rapports scientifiques n'aient pas toujours été étroits, c'est une chose heureuse et étonnante que tous soient devenus des amis parfois proches. Ce sont Aziza Merzouki, Christophe Charpilloz, Federico Brogi, Jean-Luc Falcone, Jonas Lätt, Mohamed Ben Belgacem, Orestis Malaspinas, Pierre Künzli, Sha Li, Xavier Meyer et Yann Thorimbert.

Mais plus encore, je prends congé avec ce manuscrit de dix ans de physique théorique, parcours semé d'embûches où certaines figures méritent un salut particulier; c'est le cas de Jan Lacki, qui a fait plus que quiconque pour que cette trajectoire aille à son terme. Et c'est bien sûr le cas de mes comparses Adrien, Deeraj, Maud (à qui le chapitre 6 est dédié), Stéphanie et Raphaël.

Enfin, je dois dire quelque chose de mes parents, qui ont toléré le désordre moral et physique caractéristique de cette longue période. Sans doute était-ce là le prix à payer pour voir leur rêve (pour mon père) ou leur cauchemar (pour ma mère) se réaliser !

Abstract

This work stems from a research project ("Sophocles") whose ambition was to consider complex systems as structures through which information propagates, and, propagating, gets modified. Though this perspective in a sense bridges the gap between Lavoisier's old intuition (*"Rien ne se perd, rien ne se crée, tout se transforme."*) and modern information theory as pioneered by C. Shannon, it must be underlined that the seemingly intuitive notion of "complex system" does not admit any universally accepted definition, but instead denotes a set of phenomena and properties loosely related to each other. It is therefore necessary to look for unifying principles relating those disparate elements. Among these principles, the concept of *entropy* has played a central role not only in the development of statistical physics, but also in that of information theory.

However, the concept of *information* may also be understood in a more vernacular sense as the knowledge held by an entity about another. This is the perspective of chapter 1, where we propose an agent-based model inspired by the *Minority Game*, in which are introduced some agents having an extra information regarding the behaviour of others. Introducing this population impacts the global state of the system; our purpose in this chapter is to investigate this impact, depending on the size of this smart population and the amount of information it holds. It appears that being too well informed, or being too many to be well informed, does not necessarily lead to an actual competitive advantage.

After this preliminary chapter, we focus on information in the sense of modern information theory, to which chapter 2 provides an express introduction where are introduced all the technical tools used thereafter. In particular it is shown how *Shannon entropy* overlaps – almost uniquely – the intuitive concept of *uncertainty* contained in a probability distribution.

Interpreting entropy as a measure of uncertainty in turn allows setting up a heuristic criterion for choosing, among a set of distributions satisfying some structural or observational constraints, the one displaying the largest uncertainty. However different cases have to be distinguished, depending on the kind of constraints imposed on the distribution. In chapter 3 we expose the two most important cases, namely the one where moments of the distributions are constrained and the one where some given marginals are fixed. Some more or less standard results are summarized there, that are necessary for the understanding of the subsequent chapters containing most of the original material.

The original formulation of the principle of maximum entropy assumes that the target distribution does not depend on time. Nonetheless it is possible to generalize

the principle to the case of dynamical processes. In chapter 4 we discuss the case of a discrete Markov process whose transition matrix has to maximize the entropy rate. This is not achieved without pain, since the problem can be solved at the expense of a detailed balance assumption. Our analysis focuses on the quality of the reconstruction obtained using the criterion of maximum entropy rate and the way this reconstruction changes with the length of the sample used for inference. It is shown (analytically for the 2-state case and numerically for larger state spaces) that when the sample available is very short, some processes are better estimated using this criterion than by standard histogram sampling. This possibly opens the way to "real-time" estimation of non-stationary processes. As a conclusion, we expose briefly an alternative method which does not rely on detailed balance.

The last two chapters are devoted to inference from constrained marginal distributions. In chapter 5 we try to establish a link between information theory and complexity in the context of cellular automata. In the first part the principle of maximum entropy is used for decomposing the information content in an elementary cellular automaton. Unfortunately this decomposition happens to have little in common with other measures of complexity usually employed. In the second part we adopt a more local perspective and focus on the propagation and processing of information in an elementary neighborhood. Although here again the results obtained are not always in accordance with alternative measures of complexity, some interesting similarities appear.

Chapter 6 conveys a somewhat more speculative flavour. Starting from the trivial observation that the molecular chaos hypothesis (*Stosszahlansatz*) underpinning most of the kinetic theory of gases is a particular case of the criterion of maximum entropy, we generalize this hypothesis so as to truncate the BBGKY hierarchy at the second order, which leads to a kinetic equation for *pairs* of particles. Two consequences of this equation are, on the one side, the possibility of equilibrium solutions where particle momenta are correlated, and, on the other, the existence of a *bilocal* colisional invariant – hence of a conservation equation beside mass, momentum and kinetic energy conservation.

The main concern of this thesis thus revolves around assessing the criterion of maximum entropy in several contexts. We conclude it by sketching an alternative approach which, though very different from the one discussed here, aims at the same goal.

Résumé

Ce travail émane d'un projet de recherche européen ("Sophocles") dont l'ambition était d'envisager les systèmes complexes en tant qu'organisations dans lesquels l'information se propage, et, tout en se propageant, se trouve modifiée. Si cette perspective opère une forme de synthèse entre la vieille intuition de Lavoisier (*"Rien ne se perd, rien ne se crée, tout se transforme."*) et les développements modernes de la théorie de l'information initiée par C. Shannon, il faut toutefois souligner que la notion apparemment intuitive de "système complexe" n'admet pas de définition universellement acceptée, mais recouvre un ensemble de phénomènes et propriétés sans relations évidentes les uns avec les autres. Il convient donc de chercher des principes unificateurs qui permettent de jeter des ponts entre ces éléments disparates. Parmi ces principes, le concept d'*entropie* a joué un rôle central non seulement dans le développement de la physique statistique, mais aussi dans celui de la théorie de l'information.

Pour autant, la notion d'*information* peut aussi être comprise dans un sens plus vernaculaire, comme la connaissance que possède une entité sur une autre. C'est là la perspective du chapitre 1, dans lequel est proposé un modèle d'agents inspiré du *Minority Game*, mais dans lequel sont introduits certains agents disposant d'une connaissance supplémentaire quant au comportement des autres. L'introduction de cette population bien renseignée n'est pas anodine quant à l'état global du système; l'objectif du chapitre est d'étudier ces effets selon la taille de la population informée et la quantité d'information dont elle dispose. Il s'avère qu'être trop bien informé, ou trop nombreux à l'être, ne conduit pas toujours au bénéfice escompté.

Après ce chapitre liminaire, nous nous concentrons sur l'information au sens de le théorie de l'information moderne, à laquelle le chapitre 2 propose une introduction où sont définies les notions utilisées par la suite. En particulier, il y est montré comment l'entropie de Shannon donne corps – de manière presque unique – à la notion intuitive d'*incertitude* contenue dans une distribution de probabilité.

Cette interprétation de l'entropie comme mesure d'incertitude permet à son tour de spécifier un critère heuristique permettant de sélectionner, parmi un ensemble de distributions de probabilité satisfaisant certaines contraintes structurelles ou observationnelles, celle contenant l'incertitude maximale. Il importe toutefois de distinguer différents cas selon le type de contraintes imposées à la distribution en question; dans le chapitre 3 nous exposons les deux cas principaux, l'un où des moments de la distributions sont contraints et l'autre où certaines marginales sont fixées. Certains résultats plus ou moins standards y sont rappelés, qui sont nécessaires à la compréhension des chapitres suivants, lesquels contiennent l'essentiel du matériel original.

Dans la formulation originale du principe d'entropie maximum, la distribution cible est réputée indépendante du temps. Il est cependant possible de le généraliser au cas de processus dynamiques. Dans le chapitre 4, nous examinons le cas d'un processus Markovien discret dont la matrice de transition doit maximiser le taux d'entropie. Ceci ne va pas toutefois sans peine, puisque le problème devient soluble au prix d'un recours à une condition de bilan détaillé. Ceci fait, notre analyse se concentre sur la qualité de la reconstruction obtenue en utilisant le principe d'entropie maximum, en fonction de la longueur de l'échantillon sur lequel se base l'inférence. On montre (de manière analytique pour le cas à deux états et numérique pour des espaces plus grands) que lorsque l'échantillon à disposition est très court, certains processus sont mieux approximés par ce critère que par un échantillonage standard par histogramme. Ceci ouvre éventuellement une porte à l'estimation "en temps réel" de processus physiques – non nécessairement stationnaires. En conclusion, nous présentons une méthode alternative, permettant de relaxer la condition de bilan détaillé.

Les deux derniers chapitres sont consacrés à l'inférence à partir de distributions marginales contraintes. Dans le chapitre 5, nous tâchons d'établir un lien entre théorie de l'information et complexité dans le contexte particulier des automates cellulaires. Dans la première partie, nous utilisons le principe d'entropie maximum pour décomposer la quantité d'information contenue dans un automate cellulaire élémentaire. Il s'avère malheureusement que cette décomposition a peu à voir avec d'autres notions de complexité utilisées habituellement dans le domaine. Dans la seconde partie, nous adoptons une perspective plus locale en étudiant la propagation et le traitement de l'information au niveau d'un voisinage élémentaire. Quoiqu'ici encore les résultats obtenus ne se trouvent pas forcément en accord avec la théorie de la complexité usuelle, certaines similarités intéressantes se font jour.

Le chapitre 6 est d'un caractère plus spéculatif. Partant de l'observation triviale que l'hypothèse du chaos moléculaire (*Stosszahlansatz*) sur laquelle se fonde la théorie cinétique des gaz est un cas particulier du critère d'entropie maximum, nous généralisons cette hypothèse de manière à tronquer la hiérarchie BBGKY au second ordre. Ceci conduit à une équation cinétique pour les *paires* de particules. Deux conséquences de cette équation sont, d'une part, la possibilité de solutions d'équilibre où les impulsions des particules sont corrélées, et, d'autre part, l'existence d'un invariant de collision *bilocal* – et donc d'une équation de conservation bilocale en sus de celles pour la masse, l'impulsion et l'énergie cinétique.

L'essentiel de cette thèse est ainsi consacré à l'évaluation critique du critère d'entropie maximum dans différents contextes. Nous la concluons avec un bref aperçu d'une approche qui, quoique très différente, concourt au même résultat.

Introduction

The material presented here stems from the work carried through during a three-years collaborative project ("Sophocles") funded by the European Union Framework Program. Although this thesis diverges to some extent from the goals announced initially, it would be fair to start by describing briefly the scope and aim of the project in order to situate the present work within its original context.

'Sophocles' is a quasi-acronym for *self-organized information processing, criticality and emergence in multilevel systems*. It therefore lets intervene several sophisticated notions such as *self-organization*, *criticality*, *processing*, *emergence* and *multi-levelness*, whose contour and definition are often equivocal. These keywords are actually facets of the broader concept of *complexity* which – even though it strangely does not show up in the title – forms the core of 'Sophocles'. This is this notion of complexity that we shall be mostly interested in this chapter. It could be enough to stick to an implicit definition of a complex system as a multi-level entity that self-organizes, processes information and displays criticality and emergence, but we shall instead try to follow some avatars of complexity through history in order to shape our own picture of the subject. This picture is certainly only one among many, but it reflects our own understanding and as such announces some orientations of the work presented later.

From Sophocles to Aristotle

Complex originates from the latin *cum* ("with") and *plectere*, which itself comes from the Greek *plekein* ("to weave", "to braid", but also "to build" or "to compose"). *Complexus* can therefore mean "entwined", "encircled" or "infolded". It therefore conveys a notion of unity resulting from the union of many. Since the word can be traced back to greek origins, it is not surprising that the notion itself appears first in a scientific context under the calamus of none other than Aristotle. In book VII of the *Metaphysiscs* [5], the philosopher discusses the relation that parts entertain to the whole. Though the style may sound somewhat old-fashioned nowadays, the following extract from section VII, 10 sets the tone:

> *For they cannot even exist if severed from the whole; for it is not a finger in any and every state that is the finger of a living thing, but a dead finger is a finger only in name.*

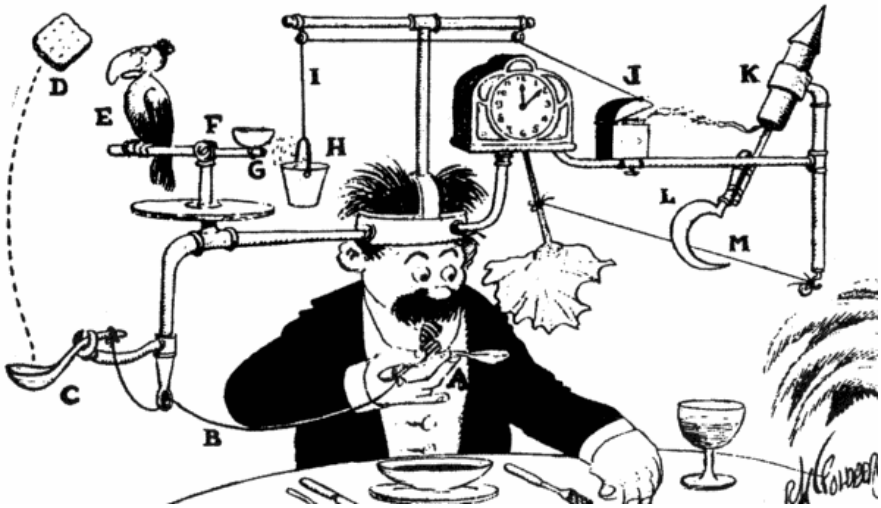


Figure 1: Rube Goldberg's self-operating napkin: complex or trivially clumsy ?

Here clearly the idea of unity dominates, but making a temporal leap forwards we can see that it is not always so. In his *Treatise on Human Nature* [50], Hume clearly emphasizes that what he calls "complex" can be resolved into parts:

> *There is another division of our perceptions, which it will be convenient to observe, and which extends itself both to our impressions and ideas. This division is into* Simple *and* Complex. *Simple perceptions or impressions and ideas are such as admit of no distinction nor separation. The complex are the contrary to these, and may be distinguished into parts. Tho' a particular colour, taste, and smell are qualities all united together in this apple, 'tis easy to perceive they are not the same, but are at least distinguishable from each other.*

Obviously merging two antagonist notions into one word has led to some confusion as to the use of "complexity" that still perdures nowadays. In this thesis "complex" will be preferentially reserved for systems in which the unitary aspect is essential, while composite systems whose parts do not form an organic whole can be said "complicated", "messy" or the like. Let us emphasize that there is no clear-cut separation between complexity and complication; for how should be classified the MacGuffin illustrated in figure 1 ?

It is premonitory but not surprising that Aristotle picked a biological example to illustrate the subtle relation between an entity and its parts. Indeed it soon turned obvious that while celestial bodies and physical phenomena could be adequately grasped through mathematical reasoning, living beings displayed some 'added value' that could not easily be cast into such kind of mathematical mould[1]. One could actually argue that

[1]No doubt that for the scholastic continuators of Aristotle's philosophy this added value could be identified with soul, though himself did not necessarily intend so.

this led to the regrettable habit of qualifying as 'complex' any object or phenomenon which escapes the grasp of physics: after biology made advances significative enough to consider that it could eventually enter our mathematico-physical frames, complex scientists were prompt to turn their attention to more 'complicated' entities like social [43, 98], economical [70, 15] or technical ecosystems. In particular financial systems deserve a mention due to the remarkable variety and originality of some transversal approaches they raised [104]. Indeed even advanced theoretical concepts from quantum [8] or particle [51] physics have started an unexpected new career in this completely different field of study.

An appealing feature of human ecosystems as an object of mathematical study is that they can easily be represented as networks of interacting agents (see [75] for an overview). Networks lend themselves to some extent to analytical as well as numerical investigation, and in particular in the theory of random networks several calculations are reminiscent of classical results in statistical mechanics [75]. More generally a wide range of physical tools and results have been transfered successfully to the theory of networks [42].

Though the theory of complex networks is interesting in itself, it becomes considerably richer when vertices are not simple mathematical points that take their sole interest from the links between them but have a dynamics of their own that upgrades them from *points* to *agents*. As an instance recent advances as to the way information percolates through networks of agents can be found in [33, 34]. It should be emphasized that 'Information' has there to be understood in a more colloquial sense than the kind of information we shall mostly be interested in this thesis, even though [84] addresses the problem in a way which is closer to the spirit of information theory. Networks also provide a convenient setting for investigating the tendency of some systems to aggregate into hierarchical organizations (hence *multi-levelness*) [33, 34]. A network structure is however by no means necessary to the development of hierarchies, and the same questions for continous systems of agents have long since raised a keen interest [10, 9, 12, 11, 109, 110, 4, 3]. The *talon d'Achille* of agent-based modelling lies in the difficulty to find interaction rules that do not simply jump out of the theoreticist's mind but represent faithfully the interactions of real social or economical agents. Possibly one of the most relevant approaches is still to rely on experiment (see for instance [64]), complemented by a grain of game theory [106].

An interesting link between living organisms and physical systems relies on the conjecture that (at least from a cosmic perspective) life is a rare – if not unique – phenomenon. A central concept of modern statistical mechanics is that of *phase transiton*, namely the fact that the transition from one equilibrium state to another is not smooth but occurs sharply at a *critical point* [93]. It is therefore tempting to conjecture that frail life is possible only around such a critical point, as developed for instance in [73]. Critical transitions also have an appealing feature for complex scientists because in physics transitions often come along with oscillatory phenomena. The detection of such oscillatory behaviours would provide a convenient way to predict catastrophic events when such an anticipation is needed (and it often is), and is obviously a popular subject of investiagtion [81]. Recent applications to financial markets can be found in [104, 82].

As we can see, the original project therefore makes intervene many different concepts that are all believed to catch a particular aspect of complexity, but whose logical relations to each other are not evident; though possible indeed, it is not likely that one can build a completely unified picture of complexity from such disparate building blocks. There is however in statistical mechanics one such unifying concept that plays an important role, and that we must briefly discuss now.

Complexity and entropy

Although there was certainly a gap between physics and chemistry on the one side and biology on the other from the early days of modern physics (*i.e.* after Galileo's work), the situation turned considerably worse around the mid-point of the 19th century due to the advances of thermodynamics. Until that date the reach of classical mechanics was limited to relatively simple, stable, well-ordered systems, and drawing a parallel between a living being and a planetary system or a clock mechanism, though globally wrong, could not be considered as utterly absurd (witness Descartes' theory of mechanistic animals [35]). Things changed drastically with the advent of thermodynamics, since not content with providing a framework for the study of physical or chemical processes, the second principle of thermodynamics seemed to provide a general explicative principle to all physical phenomena (what would be called nowadays a "theory of everything"): entropy increases forever – or more carefully stated *something* increases forever, and let *entropy* be a generic name for it[2]. Though of utmost importance for physics, the second principle unfortunately opposes the intuition we can shape from most natural phenomena: even if the long term stability of biological or social systems can be considered as a matter of opinion to a large extent[3], living organisms tend to be remarkably resilient to thermodynamical decay and human societies are prompt to oscillate around strongly hierarchical metastable structures - despite recurrent claims that this is but a step on the way towards an everlasting positive (in the moral sense – needless to say !) equilibrium.

While this schism among natural sciences led to an outburst of what would be considered today as parascience (spiritism, vitalism, *etc.*), it also prompted the will to reconsider the old problem of casting biology into the mathematico-physical scheme through the spyglass of the new physics. A typical instance of this attitude (not unexpected from a physicist actually) can be found in Schrödinger's pioneering monography on the foundations of genetics [90], where the author discusses the stability of genetic codes in entropic terms, with an emphasis on the role quantum mechanics possibly plays here.

Around the same time, the progresses made, among others[4], by Maxwell, Gibbs and

[2] An interesting and diversified overview of the versatile history of the concept of entropy can be found in [46]

[3] "In the long run we are all dead." (J. M. Keynes)

[4] Unfortunately it is not the place to sing the folk song of early statistical mechanics. When men-

Boltzmann from thermodynamics to modern statistical mechanics and kinetic theory of gases contributed to shed a new light on the nature of entropy and the meaning of the second principle. Entropy clearly appeared then to be related to the probabilistic description of large mechanical systems. It is however Shannon's merit [91] to have shown that entropy had a central place not only in statistical mechanics but also in the theory of probability in the large as a measure of uncertainty (though first introduced in the context of signal processing, its wider range of applicability was soon highlighted in Khinchin's mathematically rigorous account [59] of the theory).

This was not without consequence for statistical physics itself since within a few years Jaynes [52, 53] proposed to get rid of its mechanical part and turn the theory into an almost exclusively statistical one. This was achieved at the expense of a considerable shift of focus as to the role given to entropy, and actually laid the first stone of the maximum entropy ("Maxent") approach to statistical mechanics. This approach will play an essential role in this thesis and will be introduced in details in chapter 3; for the moment, let us simply state that the key element of the theory is to consider entropy, taken as a measure of uncertainty, as a heuristic criterion for discriminating between models drawn directly from experiment.

Jaynes' formulation could not however supersede the standard mechanistic approach, partly because letting the entropy stand on the forescene seems to introduce a subjective ingredient that is apparently at odds with the supposed objectivity of the standard approach (this alleged paradox will be discussed in more detail in chapter 6). However, it is precisely this subjective aspect that makes the principle of maximum entropy a versatile instrument whose wide utility progressively came to be recognized in the last decades. This movement was prompted partly by the work of Schneidman and coworkers [87] on retinal neurons where the principle of maximum entropy is used for reconstructing the distribution of neuronal patterns from pairwise correlations and firing rates. Such maximum entropy models based on observables that can be expressed as low-order momenta are especially popular due to their tight connection with the almighty Ising model (see chapter 3), and in the meanwhile closely similar analyses have been carried through, among others, on the structure of stock markets [20] or the structure of bird flocks [14]. Long restricted to probability distributions that are fixed in time, the maximum entropy formalism can actually be extended so as to encompass a temporal dimension. Recent efforts in this direction are to be found in [100, 71, 22, 38, 39, 37, 41]. Bringing in a temporal aspect however raises technical subleties that are discussed in chapter 4.

Outline

The historical sketch above is bound to be biased by the author's readings and personal taste, and it pretends in no way to exhaust the many beautiful developments met in attempting to grasp complexity. However it highlights that on the one hand the theory

tioning such early works we shall simply refer to Uffink's historical survey of foundations of statistical mechanics [99], where all appropriate references regarding this period are listed and commented.

of complex systems ramifies into many subdomains whose relation to each other is sometimes loose, and on the other that entropy provides a unifying pattern to which some (but not all) of the aforementioned aspects of complex systems can be integrated.[5] In particular the principle of maximum entropy is extremely appealing both from a down-to-earth perspective (in that it sets up a scheme for infering direcly from observation) as well as from a theoretical viewpoint (since the rationale for adopting it as a guiding principle is not trivial[6]).

In writing this thesis we have attempted to account for both unity and diversity. On the one side we have tried to focus as much as possible on the criterion of maximum entropy and the several ways it can turn relevant in the study of complex systems. This approach will be the topic of chapters 3, 4, part of 5 and 6. Although this constitutes our most original contribution to 'Sophocles', focussing exclusively on the maximum entropy principle would not convey a complete picture of the project. Therefore we shall also allow us the licence to divert from our essential focus in chapter 1 and the second part of chapter 5. Our aim was however in no way to give an exhaustive account of the many directions explored within the project, and the areas that could not be connected directly to our own line of thinking based on information theory had to be left in the shadows.

The thesis is therefore built around three main pieces of work and is structured as follows:

1. Chapter 1 stands apart from the following chapters, inasmuch as it makes no use of information theory. Our purpose in this liminary part was to show that information could be interpreted in a quite vernacular sense as a knowledge that part of a system has about another. To do so we build a toy model which is a generalization of the so-called *Minority Game*. This relatively simple dynamics allows reproducing some features that one might attribute intuitively to information, but also gives caution against the danger of being over-informed.

2. Chapter 2 is a contextual chapter that introduces the concepts of information theory that will be extensively used in what follows. Shannon entropy is shown to be a necessary consequence of Khinchin's axioms, and possible generalizations resulting from relaxing the axioms are briefly presented. Mutual information and total information are then derived as well as their relation to Kullback-Leibler divergence.

3. Once the probabilistic meaning of entropy is made clear the criterion of maximum entropy is stated and discussed in chapter 3. We present both the variant based on

[5] We said above that the tools of fundamental physics fertilized the study of complex systems, but to be fair the converse is also true; the second principle leads naturally to the concept of *entropic force*, which amounts to recasting a statistical property in newtonian terms. See [74, 103, 108].

[6] This is particularly true when the maximum entropy approach is used in the context of bayesian inference. Our emphasis on physical models will keep us from going into these questions, but the reader wishing to get introduced to the bayesian mysteries will refer with profit to [54, 56].

constrained moments and the variant based on constrained marginals. Some standard results are presented, as well as an alternative justification of the maximum entropy criterion which is more 'objective' than the usual heuristic justification.

4. Chapter 4 is concerned with the problem of stating a maximum entropy criterion for temporal processes. While some authors avoid the problem by 'temporalizing' statical distributions, we consider the case of a discrete Markov process whose transition matrix has to maximize the entropy rate. Following Van der Straeten [100], a practicable way to proceed is to impose detailed balance.

   Our contribution in this chapter is, first, to consider observables for which the maximum entropy approach makes more sense than the observables originally considered in [100] (which were counting operators quite similar to the transition rates themselves). Second, the focus of our analysis is on the quality of the reconstructed process depending on the length of available samples. Analytical results can be obtained for 2-state processes and numerical results are shown for higher-dimensional processes. The possibility of applying this approach to empirical times series is discussed, even though *a priori* the processes generating such series satisfy none of the hypotheses. The material of this chapter was published in [26, 25]. The weak point of our approach lying in the detailed balance assumption, we present briefly a promising way to remedy this issue, proposed recently in [102].

5. In chapter 5 we tie a closer link between information theory and complexity in the particular context of cellular automata. The chapter splits into two parts. In the first we use the principle of maximum entropy (in its marginal-based form) to decompose the informational content of a system in a theoretically satisfying way, proposed by Schneidman & al. [88]. Our contribution is to put this decomposition at work and assess its relevance for simple dynamical systems. It turns out that this decomposition is not as informative as expected, which brings into question either the relevance of the decomposition, or the relevance of our expectations, or both. In the second part we reconsider the question from a more local perspective by studying the processing of information at the level of an elementary neighbourhood. We show that these so-called *informational features* can be expressed in terms of the lookup table defining the dynamics when the cells of the automaton are decorrelated. Moreover, we show that clustering elementary cellular automata according to their features structurates the space of features in a way that is in partial accordance with more standard notions of complexity.

6. The material of chapter 6 conveys a more speculative flavour. Starting from the trivial observation that the principle of maximum entropy allows reformulating Boltzmann's assumption of molecular chaos, we make a proposal to generalize it so as to close the BBGKY hierarchy at the second order. This leads to a kinetic equation for pairs of particles (published in [27]). Two consequences of this equation seem to be that particles can be correlated at equilibrium (which we argue should result from the conservation of the fine-grained entropy under Liouville's

equation), and that the usual conservation equations of mass, momentum and energy have to be complemented by a fourth macroscopic balance equation.

7. At the time of concluding and summarizing we expose a method recently proposed by Obuchi & al. [77, 78] for assessing the relevance of maximum entropy distributions. This method, based on randomized observables, shares some goals with the present thesis and as such offers comparative interest.

Doubtless we have thus significantly derived compared to our initial aims; we hope that the reader will find the journey illuminating nonetheless.

Chapter 1

Minority Game and the role of information

In the next chapter and thereafter we shall focus our attention on information theory, in which "information" can be understood as a well-defined mathematical information (see eq. (2.33)). Before proceeding in this direction we shall however in this chapter digress slightly from our principal aim, and consider information in a more colloquial, vernacular way. In this context information can be defined - if needed - as a knowledge an agent possesses about the system. Our purpose in this chapter is to investigate the role of information in the setup of a modified Minority Game [23] in which sophisticated informed players coexist with naive agents playing according to the rules of the standard minority game. It will turn out that in this context information is both useful and dangerous.

It should be noted that introducing differentiated agents constitutes to a certain extent a deviation from the usual practices of statistical mechanics. Large parts of statistical mechanics rely on the use of the central limit theorem, and since the CLT applies to sets of independent and *identically distributed* variables, it is in general assumed (implicitely or explicitely) that the system is constituted of agents which are homogeneous in terms of their properties. This kind of assumption is not exclusive to physics; for instance, economical science relied for decades on the assumption that all economical agents are perfectly rational players operating on a completely transparent market. It took a long time to realize that, opposite to this highly idealized picture, human activities are characterized by the diversity of behaviors and the coexistence of various, more or less inhomogeneous, populations.

Allowing diversity of beheviors is of course particularly important in the context of economical science. While for microscopic agents constituting physical systems adopting such or such a behaviour makes no subjective difference, in economics the possibility to tune one's behaviour makes the full difference between gain and loss, success and failure. A well-known model featuring a clear-cut notion of success and failure is the so-called Minority Game introduced by [23], elaborating on the previously known *El Farol dilemma* [6]. El Farol is a saloon somewhere in New Mexico where every thursday evening a band plays irish folk music. Theses sessions are so appreciated that when

too many peoples attends brawls usually burst and attention is - sadly - diverted from the music. This painful situation thus rises a dilemma: *go to the pub or stay home ?* If too many people decide to get out, brawls are unvoidable and people get little satisfaction from the music (brawling in this idealized discussion is supposed to be no fun). If too many people decide to stay home, the happy few attendants will enjoy the music, but the majority will be bored to death. The optimal situation is when the attendance is just below the brawling threshold, but it is not obvious how to set up, for each player, a strategy allowing an optimal attendance (in a planified society on the other hand fun optimization is much easier). What is certain is that there can be no ideal strategy from the perspective of an individual agent, because otherwise all players would use this strategy and behave similarly, resulting in the disastrous issue mentioned above.

1.1 Classical Minority Game

The Minority Game introduced in [23] is essentially a reformulation of the El Farol dilemma in monetary terms. It consists of a set of N (chosen odd to ensure that a minority is well defined) agents taking binary decisions $+1$ and -1 and competing within an iterative betting game. After each iteration, a count is made of how many agents played $+1$ and how many played -1. Agents belonging to the minority are supposed to win, although in some cases a *majority* game would shape a more faithful picture of the reality[1]. Denoting by $a_i(t)$ the decision of agent i at round t, we define the *attendance* $A(t) = \sum_{i=1}^{N} a_i(t)$ and the *outcome* $O(t) = -sign(A(t))$. Both quantities represent macroscopic descriptions of the game at a given instant. O offers a particularly coarse-grained description since it describes only the winning (= minoritary) choice.

Agents base their decisions upon *strategies*. A strategy $s(t)$ is a mapping from a historical series of the M previous outcomes to a binary decision $a_{i,s_i(t)} = \{\pm 1\}$. A strategy is therefore a mapping from the vector of the last M outcomes to $\{+1, -1\}$. Each of the 2^M possible histories can be assigned two decisions, so that there are 2^{2^M} possible strategies. Instead of considering all possible strategies, each agent is assigned a limited number S of strategies picked at random. To each strategy is assigned a score by computing how many times over the historical record at disposal this strategy would have indicated the bet that actually happened to win. Among the strategies at hand one is then selected with a probability proportional to its score. Finally each player is assigned a wealth W_i which is updated as

$$W_i(t+1) = W_i(t) + a_i(t)O(t)\left(\frac{1+O(t)}{2}\left(\frac{N-A}{N+A}\right) + \frac{1-O(t)}{2}\left(\frac{N+A}{N-A}\right)\right). \quad (1.1)$$

This messy expression simply states that the loosers' money is shared among the winners. Since the total wealth in the game is defined as $W = \sum_{i=1}^{N} W_i$, it is easily verified

[1] In finance it seems that both attitudes coexist. Most actors are *trend followers*, but traders that deliberately bet against the current consensus form the non-neglegible *contrarian* faction. See also [2].

that $W(t+1) = W(t)$; the Minority Game is therefore a zero-sum game.

1.2 Informed Minority Game

In the classical game all agents are identical in the sense that they all have the same number of strategies and choose among them according to the same procedure. Our immediate purpose is to bring heterogeneity in this setup by introducing a variant of the game which involves two different populations, one regrouping usual players and another regrouping players possessing an extra information about the behaviour of the first population.

This can be achieved by defining the following two-step dynamics. Let N_{inf} among the N players get an extra information, while the $N_{non-inf} = N - N_{inf}$ other agents play according to the rules of the standard minority game at time $t - \epsilon$. At time t informed players are allowed to look at the decisions of a subgroup of standard players (informing pool) that just occurred at time $t - \epsilon$. Each informed player bets -1 if most players in its informing pool play $+1$, and $+1$ if most of its informers play -1. The pool changes from one smart player to another, but all pools will be assigned a fixed size B.

The outcome of the game at time t is therefore defined in terms of the set of decisions of naive players at $t - \epsilon$ and informed players at time t.

1.3 Wealth dynamics

Intuitively we might expect the extra information allowed to informed agents to give them a competitive advantage leading to larger average gains. Figure 1.1 displays the average gain per step g_{inf} of an informed player as well as the average gain per step $g_{non-inf}$ of a standard player, as a function of the size B of the informing pools. Gains are cumulated over $t_{max} = 500$ time steps. $M = 3$ bits of history are used for defining strategies and each naive agent has at hand a set of $S = 10$ strategies. The number of standard agents is set to $N_{non-inf} = 51$ while the informed population varies from $N_{inf} = 4$ to $N_{inf} = 20$. In order to reduce stochastic noise the results are averaged over 100 realizations. Informed players are signaled by $(\times)$, standard ones by $(+)$.

For $N_{inf} = 4$ (black curves) the gain of informed players grows steadily until $B \approx 40$ while standard players loose the corresponding amount. When pool sizes get closer to $B = 51$ the gain of informed players tends to decrease slightly. The informed agents therefore draw a considerable advantage from their privileged knowledge when the size of their informing networks stays small. Even though their gain starts decreasing when B becomes large, informed players eventually retain some advantage over standard players even when looking at the whole non-informed population (*i.e.* $B = 51$).

This is no longer the case when $N_{inf} = 8$ (blues curves), for there the curves of informed and standard players intersect around $B \approx 40$. Thus not only wider informing pools result in lesser competitive advantage, but it also results in an effect quite opposite

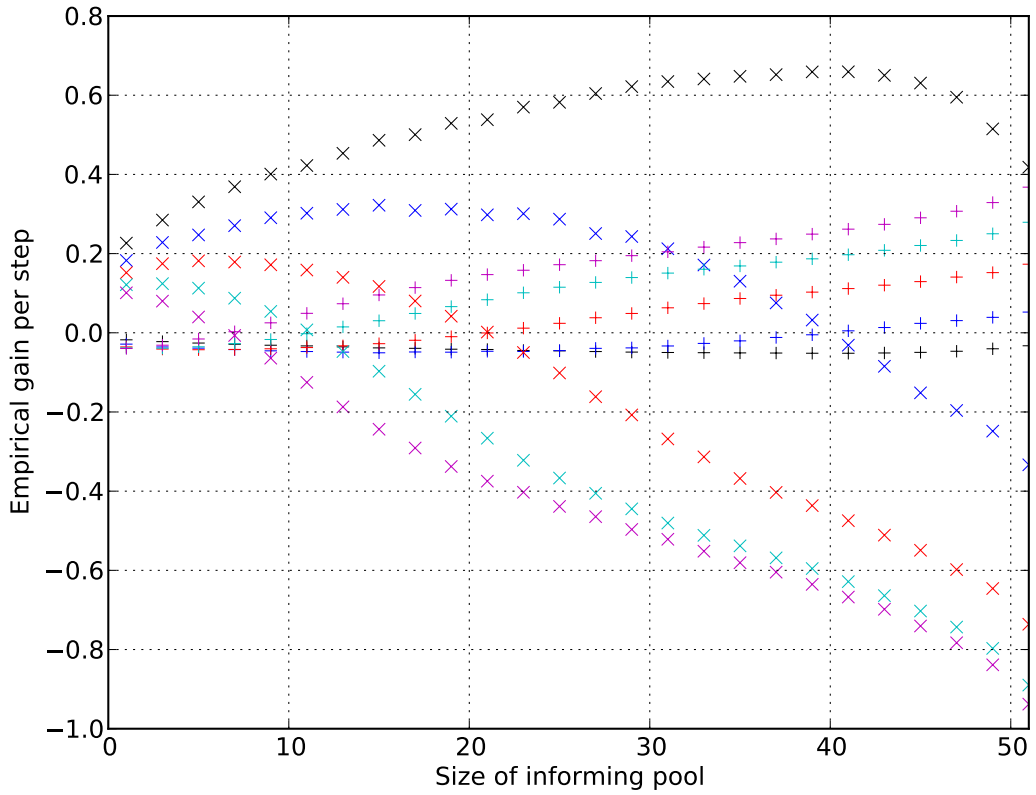


Figure 1.1: Average gain per step of informed (×) and non-informed (+) agents as a function of the size of informing pools. Average is taken over 100 simulations over 500 time steps each. Each simulation comprises $N_{non-inf} = 51$ naive players. Are shown $N_{inf} = 4$ (black), $N_{inf} = 8$ (blue), $N_{inf} = 12$ (red), $N_{inf} = 16$ (cyan) and $N_{inf} = 20$ (magenta).

to the expected one, since information happens in that case to turn into a competitve drawback.

The situation of smart players worsens for $N_{inf} = 12$ (red curves), $N_{inf} = 16$ (cyan curves) and $N_{inf} = 20$ (magenta curves). The gain of informed agents for small pools decreases gradually, while the threshold value B_c beyond which information results in a loss gets gradually smaller.

The behaviour observed in figure 1.1 can be ascribed to two distinct mechanisms at work here. As long as the parameter B stays small the informing pools are approximately independent of each other, while when B grows basins of different informed players tend to overlap significantly. Then the information collected by different players becomes more and more similar, leading these informed agents to take their decisions similarly. This behaviour might still result in a gain compared to naive playing, however the gain of the informed group will be shared among more players resulting in a smaller individual

gain.

The second effect is more dramatic. When the population of smart players is small compared to $N_{non-inf}$ its global decision has a limited impact of the game as a whole, but when the number of agents that play like the minority of standard agents becomes sufficiently large then they might lead the majority to swap. Those agents that played in accordance with the information they had at hand fall so doing in the majoritary group and eventually loose their bet.[2]

1.4 Random non-informed players: pool representativity

The probability that a given pool of informers provides a faithful information as to the actual behaviour of the non-informed population as a whole can be calculated exactly at the price of a simplifying assumption. Let us therefore modify the rules of the game by assuming naive players do not play according to strategies anymore, but place their bets randomly, while the rules stay the same for informed players. In figure 1.2 we display the counterpart of figure 1.1 with these modified rules. This modification does not alter much qualitatively speaking the features highlighted above; however these modified rules result in somewhat lesser gains for informed players, as well as smaller threshold values (see below).

The assumption of randomness allows us to obtain the probability distribution of the size m of the minority over the non-informed population, as follows. By a classical combinatorial argument, the probability of finding m agents playing -1 and $N_{non-inf}-m$ playing $+1$ is

$$\frac{1}{2^{N_{non-inf}}} \cdot \frac{N_{non-inf}!}{m!(N_{non-inf}-m)!}, \tag{1.2}$$

so that, multiplying this probability by two in order to account for the two possible outcomes (-1 wins; $+1$ wins), we find that the size of the minority is distributed as

$$p(m) = \begin{cases} \frac{1}{2^{N_{non-inf}-1}} \cdot \frac{N_{non-inf}!}{m!(N_{non-inf}-m)!} & \text{if } m < N_{non-inf}/2 \\ 0 & \text{if } m > N_{non-inf}/2 \end{cases} \tag{1.3}$$

Considering a pool of size B, we can deduce from (1.3) the joint probability that the size of the minority is m and that among this minority m_B are in the pool. Indeed $m - m_B$ players are out of the pool, so that the same combinatorial argument as above gives

$$p(m_B; m) = \frac{1}{2^{N_{non-inf}-1}} \cdot \frac{B!(N_{non-inf}-B)!}{m_B!(B-m_B)!(m-m_B)!(N_{non-inf}-B-m+m_B)!} \tag{1.4}$$

[2]Should we make a social analogy, the latter effect eventually favours (part of) the proletariat while the former effect amounts to an internal struggle among the upper class.

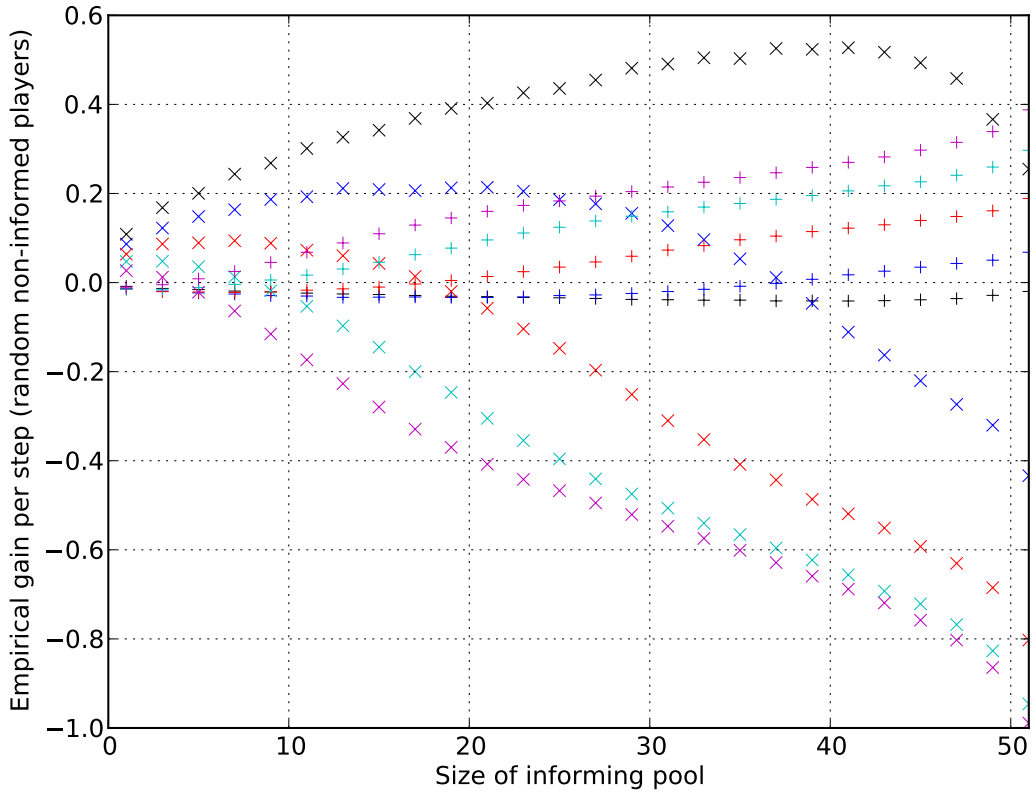


Figure 1.2: Average gain per step of informed (×) and non-informed (+) agents as a function of the size of informing pools. Here non-informed players are taken to place their bets randomly and not according to strategies. Other parameters stay the same as for figure 1.1.

if $m < N_{non-inf}/2$ and $m_B \leq m$ and $(N_{non-inf} - B) \geq (m - m_B)$.

Using eqs. (1.3), (1.4) we can write the condition probability to find m_B minoritary agents in a pool given that the overall minority counts m players as

$$\begin{aligned} p(m_B|m) &= \frac{p(m_B;m)}{p(m)} \\ &= \frac{B!m!(N_{non-inf}-m)!(N_{non-inf}-B)!}{m_B!(B-m_B)!N_{non-inf}!(m-m_B)!(N_{non-inf}-B-m+m_B)!}. \end{aligned} \tag{1.5}$$

Since the condition for the pool to be representative given that m players are minoritary is that m_B must be smaller than $(B-1)/2$, we get

$$p(\text{pool is representative } |m) = \sum_{m_B=0}^{(B-1)/2} p(m_B|m). \tag{1.6}$$

For instance in the case where $m = 0$ (all random players play the same), then by necessity $m_B = 0$. In that case (1.6), (1.5) yield

$$p(\text{pool is representative} \mid \text{all random players play the same}) = 1. \tag{1.7}$$

Of course in this case an informed player is perfectly informed !

From (1.3), (1.4), (1.5) and (1.6) we eventually obtain an expression for the probability for a pool of size B to be representative, namely to play like the whole non-informed group, as

$$\begin{aligned} p_B &= p(B\text{-pool is representative}) \\ &= \sum_{m=0}^{(N_{non-inf}-1)/2} p(B\text{-pool is representative } |m)p(m) \\ &= \sum_{m=0}^{(N_{non-inf}-1)/2} \sum_{m_B=0}^{(B-1)/2} \frac{B!(N_{non-inf}-B)!}{m_B!(B-m_B)!(m-m_B)!(N_{non-inf}-B-m+m_B)!}. \end{aligned} \tag{1.8}$$

This probability of representativity is plotted in figure 1.3.

1.5 Independent pools

1.5.1 Minority swap

As we pointed out above, if too many agents are well informed there is a risk that the overall minority will be reversed. This swap occurs when N_{inf} exceeds a threshold value

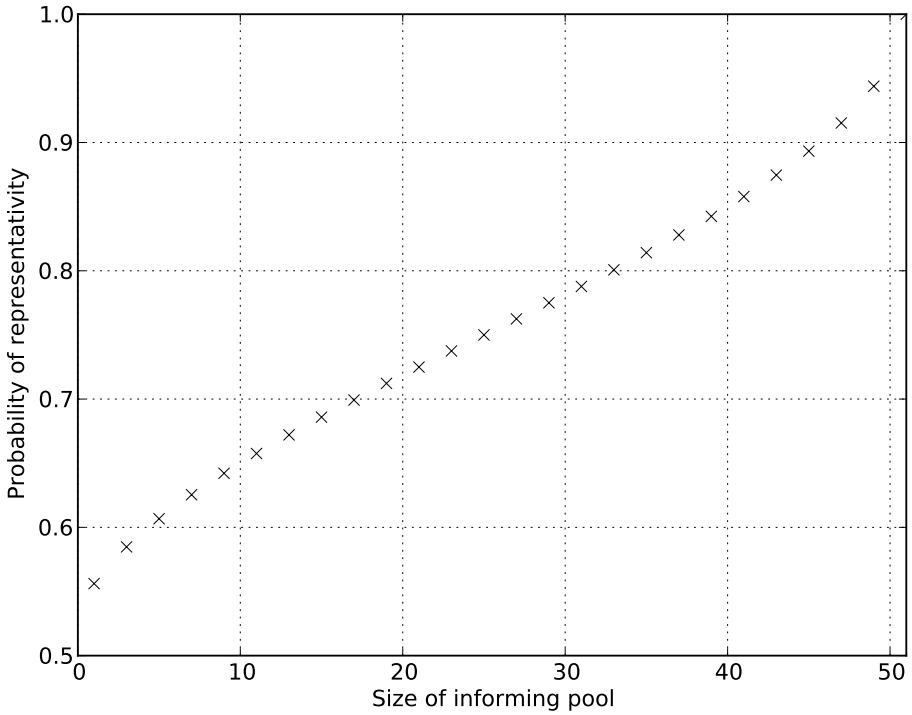


Figure 1.3: Probability for an informing pool to have its minority playing like the overall minority of standard players (computed from eq. (1.8)). The number of standard players is set to $N_{non-inf} = 51$.

N_{inf}^c. This critical value can be approximated when the informing pools of different smart players are assumed to be indepedendent of each other.

Under this assumption, the average number of informed agents that play according to the non-informed minority is $N_{inf} \cdot p_B$ while the average number of informed agents that are fooled is $N_{inf} \cdot (1 - p_B)$. Therefore the condition for making the minority swap is

$$N_{inf}^c p_B + \langle m \rangle = N_{inf}^c (1 - p_B) + N_{non-inf} - \langle m \rangle \tag{1.9}$$

where $\langle m \rangle$ denotes the average size of the minority. It can be deduced from (1.3) to be

$$\langle m \rangle = \sum_{m=0}^{(N_{non-inf}-1)/2} m p(m) = \sum_{m=0}^{(N_{non-inf}-1)/2} \frac{m}{2^{N_{non-inf}-1}} \frac{N_{non-inf}!}{m!(N_{non-inf}-m)!}. \tag{1.10}$$

which even though it cannot be put in closed form happens to be very well approximated as

$$\langle m \rangle \simeq \frac{N_{non-inf}}{2} - \sqrt{\frac{N_{non-inf}}{2\pi}}. \tag{1.11}$$

Rearranging (1.9) we find for the critical number of informed players

$$N_{inf}^c = \frac{N_{non-inf} - 2\langle m \rangle}{2p_B - 1} \simeq \frac{1}{2p_B - 1} \sqrt{\frac{2N_{non-inf}}{\pi}}. \tag{1.12}$$

1.5.2 Expected gain

Under the same hypothesis of pool independence we can also compute the gain per step of informed and non-informed players. When N_{inf} is smaller than N_{inf}^c, the average number of agents in the overall minority is $\langle m \rangle + N_{inf} p_B$. Each winner then increases its wealth by an amount

$$\frac{N_{inf} + N_{non-inf}}{\langle m \rangle + N_{inf} p_B} - 1. \tag{1.13}$$

It follows almost immediately that the average gain for each informed player is

$$g_{inf}^{N_{inf} < N_{inf}^c} = p_B \frac{N_{inf} + N_{non-inf}}{\langle m \rangle + N_{inf} p_B} - 1, \tag{1.14}$$

while the average gain for each non-informed player is

$$g_{non-inf}^{N_{inf}<N_{inf}^c} = \frac{\langle m \rangle}{N_{non-inf}} \cdot \frac{N_{inf} + N_{non-inf}}{\langle m \rangle + N_{inf} p_B} - 1. \tag{1.15}$$

When on the other side N_{inf} exceeds N_{inf}^c, then the average number of agents in the overall minority is $N_{non-inf} - \langle m \rangle + N_{inf}(1 - p_B)$. Each winner then increases its wealth by an amount

$$\frac{N_{inf} + N_{non-inf}}{N_{non-inf} - \langle m \rangle + N_{inf}(1 - p_B)} - 1, \tag{1.16}$$

and we get respectively

$$g_{inf}^{N_{inf}>N_{inf}^c} = (1 - p_B) \frac{N_{inf} + N_{non-inf}}{N_{non-inf} - \langle m \rangle + N_{inf}(1 - p_B)} - 1 \tag{1.17}$$

and

$$g_{non-inf}^{N_{inf}>N_{inf}^c} = \frac{N_{non-inf} - \langle m \rangle}{N_{non-inf}} \cdot \frac{N_{inf} + N_{non-inf}}{N_{non-inf} - \langle m \rangle + N_{inf}(1 - p_B)} - 1. \tag{1.18}$$

1.5.3 Importance of dependent pools

Theoretical gains obtained in equations (1.14), (1.15), (1.17) and (1.18) are plotted in figure 1.4 using the estimates (1.8), (1.10) obtained for p_B and $\langle m \rangle$. Again we set $N_{non-inf} = 51$ and $N_{inf} = \{4, 8, 12, 16, 20\}$. These theoretical results do not agree particularly well with those obtained numerically. This should not come as a surprise since the assumption of pool independence is not supported in the game defined here. In figure 1.5 we focus on the range $B = \{1, ..., 21\}$ and small values of N_{inf} (see caption). In that case the pools are almost independent and the agreement between empirical and theoretical gains is fairly good. It is actually almost perfect for $N_{inf} = 4$, but this happens to be a coincidence.

It must be emphasized that (1.8) and (1.10) are but approximations that hold for random non-informed players. In particular, eq. (1.10) overestimates significantly the size of the minority among non-informed players by around 25% compared to what can be observed in the simulations, which is not unexpected since assuming random playing leads to - on average - as many players betting $+1$ or -1, and thus to the largest possible minority. This overestimation in turns leads to underestimate N_{inf}^c as suggested by (1.12). Using empirical values drawn from simulations instead of (1.8), (1.10) allows us to reduce the discrepancy between theoretical and observed results, nonetheless the dependence between informing pools turns out to be an essential ingredient of the model.

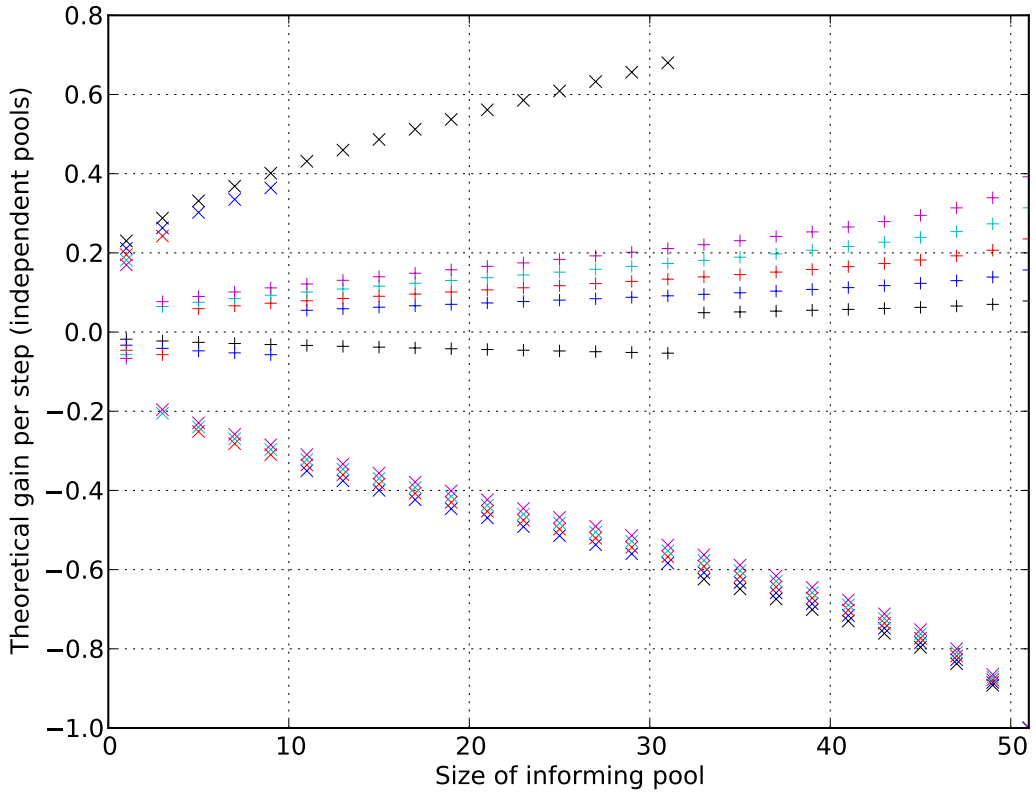


Figure 1.4: Theoretical gains for informed ($\times$) and non-informed ($+$) players according to (1.14), (1.15), (1.17) and (1.18). As previously we have $N_{non-inf} = 51$ and respectively $N_{inf} = 4$ (black), $N_{inf} = 8$ (blue), $N_{inf} = 12$ (red), $N_{inf} = 16$ (cyan) and $N_{inf} = 20$ (magenta).

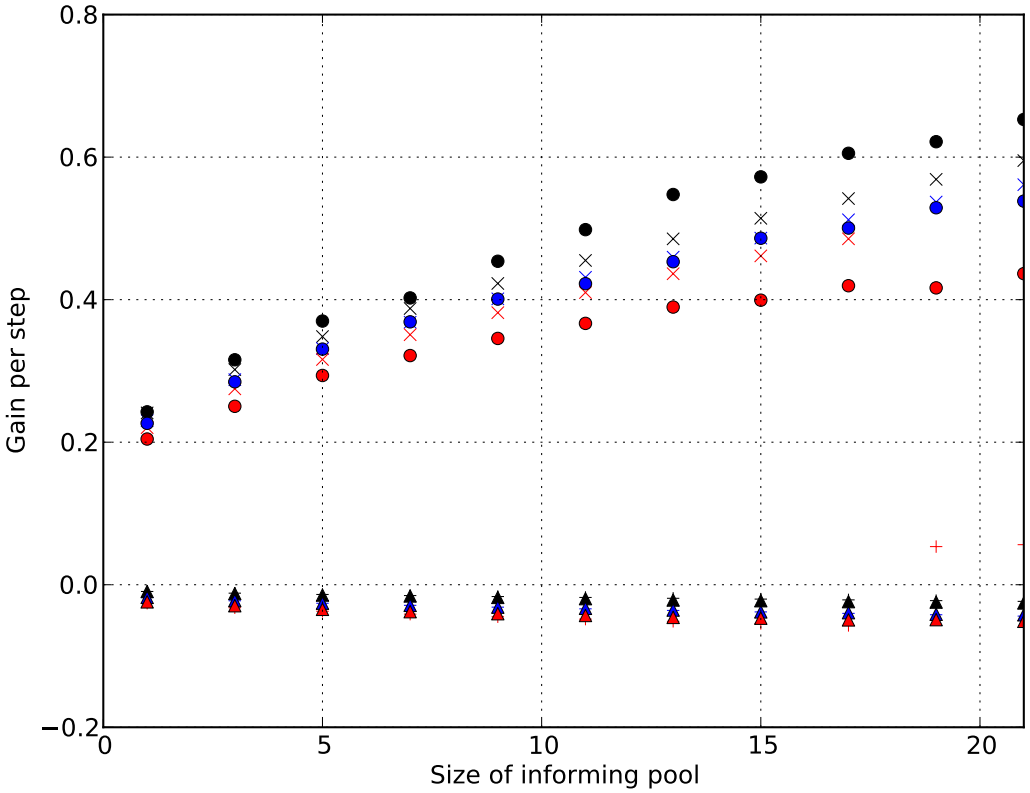


Figure 1.5: Theoretical gains for informed (×) and non-informed (+) players according to (1.14), (1.15), (1.17) and (1.18), compared to gains observed experimentally. Are shown $N_{inf} = 2$ (in black), $N_{inf} = 4$ (blue) and $N_{inf} = 6$ (red).

The informed Minority Game discussed in this chapter therefore displays non-trivial features as to the ambivalent effects of information on the dynamics of a complex systems. We shall now focus our interest on a particular notion of information which relies on a probabilistic description of the system under consideration.

Chapter 2

Elements of Information Theory

After the previous liminary chapter we now present the building blocks on which the subsequent material will be elaborated. We start by showing the existence of a functional that would quantify the amount of uncertainty characterizing a probability distribution. Any such functional should fulfill some intuitive requirements that can be stated in the form of three axioms, which lead to a unique solution known as *Shannon entropy*. Relaxing one of these requirements leads to less standard uncertainty measures, which for completeness are glanced through even though in what follows we shall make use of Shannon's entropy exclusively. Once the concept of entropy is introduced, several extensions flow naturally from it; in particular, mutual information provides an important measure of codependence that (opposite to correlation) depends on the joint distribution only. Mutual information in turn can be generalized to multivariate situations, providing a measure of the global amount of codependence in a set of variables.

2.1 Khinchin's axioms and Shannon entropy

2.1.1 The axioms

Let X be a random variable taking one of N possible values $x_1, x_2, ..., x_N$, and denote by $p(x_i)$ the probability that the event x_i is actually realized. We are asking whether it is possible to associate to the distribution $p(x)$ a functional $H(p(x))$ that would quantify the amount of "uncertainty" encoded in p. It is not an easy task to provide a precise and satisfying definition of uncertainty, but it is much easier to expose some intuitive properties that uncertainty should display. These properties are best encapsulated in a set of axioms introduced by Khinchin [59] in its rigorous re-exposition of Shannon's seminal paper [91]. These axioms are:

1. *The most uncertain scheme*[1] *occurs when all outcomes are equally probable, namely, for any* $p(x_1), ..., p(x_N)$,

[1]We adopt here Khinchin's convenient terminology, where a *probabilitic scheme* is meant to denote both a set of events and the associated probabilities of occurrence.

$$H(p(x_1), ..., p(x_N)) \leq H\left(\frac{1}{N}, ..., \frac{1}{N}\right). \tag{2.1}$$

2. *Extending a scheme by adding an impossible event has no consequence on the uncertainty, that is*

$$H(p(x_1), ..., p(x_N)) = H(p(x_1), ..., p(x_N), 0). \tag{2.2}$$

3. *If Y denotes another random variable and (X, Y) the composite scheme characterized by the joint probability $p(x_i, y_j)$, then the uncertainty of the composite system should be expressed as*

$$H(p(x, y)) = H(p(x)) + \sum_{i=1}^{N} p(x_i) H(p(y|x_i)). \tag{2.3}$$

In other words the uncertainty of the composite scheme is the uncertainty on X, plus the conditional uncertainty on Y weighted by the probability of occurrence of each outcome of X.

The first two axioms are quite intuitive and deserve no further explanation, while the third is of a more technical nature and can hardly be grasped without detailed exegesis[2]. It should first be noted that (2.3) could be written as well with X and Y interchanged; thinking of $-H$ as the information that can be extracted from a scheme, axiom 3 amounts to say that the total information that can be extracted about (X, Y) does not depend on the order in which it is collected[3]. Second, in the case where X and Y are independent we have $H(p(y|x_i)) = H(p(y))$ so that

$$H(p(x, y)) = H(p(x)) + H(p(y)). \tag{2.4}$$

In such a case axiom 3 therefore reduces to extensivity of H when bringing in contact (conceptually speaking) independent subsystems. However, (2.4) is a weaker requirement than (2.3), and as we shall see below requiring extensivity is not sufficient to single out Shannon entropy as the only possible form for H.

The relevance of Khinchin's axioms stems from the following theorem: *The only functional H that is continuous in its arguments and satisfies Khinchin's axioms is (up to a multiplicative constant)*

[2]This is quite reminiscent of Euclide's postulates in geometry, of which four are truly trivial while the fifth alone contains the (highly) non-trivial mathematics.

[3]As an instance where this is not verified, [13] discusses the case of a student trying to learn physics; he will obviously take out much more by attending the course on classical mechanics before the course on quantum mechanics than he would proceeding the other way round, even though technically speaking the material delivered is the same in both cases.

$$H(p(x)) = -\sum_{i=1}^{N} p(x_i) \log p(x_i). \tag{2.5}$$

H then receives the name of *Shannon entropy*; in addition to the properties stated in the axioms, entropy is always larger than or equal to 0 since all terms in the sum are negative. Note that $H(p(x))$ is usually simply written as $H(X)$ (a convention that we shall follow from chapter 3 on albeit it is slightly abusive).

Proof (Khinchin): We first prove the theorem for the case of equiprobable outcomes. Let us denote $f(N) = H(1/N, ..., 1/N)$. f is an increasing function since by axioms 1 and 2 we have

$$f(N) = H\left(\frac{1}{N}, ..., \frac{1}{N}, 0\right) \leq H\left(\frac{1}{N+1}, ..., \frac{1}{N+1}, \frac{1}{N+1}\right) = f(N+1). \tag{2.6}$$

Now consider M independent schemes $S_1, ..., S_M$ of which each contains K equiprobable events, so that $H(S_i) = f(K)$. Then $H(S_1...S_M) = \sum_{i=1}^{M} H(S_i) = Mf(K)$. However the product scheme $S_1...S_M$ consists of K^M equiprobable events, so that its entropy can also be computed as $H(S_1...S_M) = f(K^M)$. It results that

$$f(K^M) = Mf(K). \tag{2.7}$$

Now let K, W and N be chosen arbitrarily, and M such that by hypothesis

$$K^M \leq W^N \leq K^{M+1}. \tag{2.8}$$

Then taking the logarithm of each term in (2.8) we have $M \ln K \leq N \ln W \leq (M+1) \ln K$ or, rearranging,

$$\frac{M}{N} \leq \frac{\ln W}{\ln K} \leq \frac{M}{N} + \frac{1}{N}. \tag{2.9}$$

But letting act f on each term of (2.8) instead of the logarithm we can deduce similarly that

$$\frac{M}{N} \leq \frac{f(W)}{f(K)} \leq \frac{M}{N} + \frac{1}{N}. \tag{2.10}$$

It follows from the last two inequalities that

$$\left|\frac{f(W)}{f(K)} - \frac{\ln W}{\ln K}\right| \leq \frac{1}{N}. \tag{2.11}$$

Since the l.h.s. does not depend on M and since N can be chosen arbitrarily, it results that

$$\frac{\ln K}{f(K)} = \frac{\ln W}{f(W)} \tag{2.12}$$

and therefore $f(N) = \kappa \ln N$ (with $\kappa > 0$ by monotonicity). We have proved the theorem for equiprobable schemes.

The case of probabilities taking rational values is dealt with by rewriting $p(x_i) = q_i/q$, where the q_i's are integers and $q = \sum_{i=1}^{N} q_i$. We now define an auxiliary random variable Y such that the scheme of Y contains q events that are dispatched among N groups containing q_1, ..., q_N events respectively. If for the variable X the event x_k is realized, then for Y all q_k events of the kth group have the same probability q_k, while events in other groups are impossible. Thus once conditioned on x_k the scheme Y reduces to q_k equiprobable events, so that

$$H(p(y|x_k)) = f(q_k) = \kappa \ln q_k \tag{2.13}$$

and

$$\sum_{k=1}^{N} p(x_k) H(p(y|x_k)) = \kappa \sum_{k=1}^{N} p(x_k) \ln q_k = \kappa \sum_{k=1}^{N} p(x_k) \ln p(x_k) + \kappa \ln q. \tag{2.14}$$

As to the composite scheme (X, Y), it should be noted that the event (x_i, y_j) is possible only when y_j belongs to the ith group. Thus for a given i the total number of possible events is q_i so that the total number of outcomes of (X, Y) is $\sum_{i=1}^{N} q_i = q$. Therefore the outcome (x_i, y_j) occurs with probability $p(x_i)/q_i = 1/q$, *i.e.* it is the same for all outcomes. Therefore

$$H(p(x, y)) = f(q) = \kappa \ln q. \tag{2.15}$$

Using axiom 3 it follows that

$$H(p(x)) = \kappa \ln q - \kappa \sum_{k=1}^{N} p(x_k) \ln p(x_k) - \kappa \ln q = -\kappa \sum_{k=1}^{N} p(x_k) \ln p(x_k). \tag{2.16}$$

Since we have proved that H assumes this form for rational probabilities, requiring continuity implies that this extends to any probability distribution, which proves the theorem.

2.1.2 Continuous variables

For continuous distributions we shall define similarly the *Shannon (differential) entropy* by replacing the sum by an integral:

$$H(p(x)) = -\int dx p(x) \ln p(x). \tag{2.17}$$

This extension is not as trivial as expected, for (2.17) is not invariant under change of variables (note that mutual information to be defined below, on the other side, is). Some further technical points are commented in detail in [31] but they will play no role in our investigations.

2.2 Exotic entropies

2.2.1 Rényi entropy

The fact that Khinchin's third axiom is not particularly intuitive has led to consider that it was actually too stringent a requirement. If only extensivity (2.4) is imposed then a measure of uncertainty more general than Shannon's is singled out. This measure is known as *Rényi entropy* [86] and takes the form

$$H_R^\alpha(p(x)) = \frac{1}{1-\alpha} \ln \left(\sum_{i=1}^N p(x_i)^\alpha \right). \tag{2.18}$$

This actually defines a class of entropies depending on the parameter α. H_R itself is not well defined in the limit $\alpha \to 1$, but the limit can nonetheless be computed using L'Hôpital's rule. Since

$$\frac{d}{d\alpha} (1-\alpha) \,|_{\alpha=1} = -1 \tag{2.19}$$

and

$$\frac{d}{d\alpha} \left(\ln \left(\sum_{i=1}^N p(x_i)^\alpha \right) \right) |_{\alpha=1} = \frac{\sum_{i=1}^N p(x_i)^\alpha \ln p(x_i)}{\sum_{i=1}^N p(x_i)^\alpha} |_{\alpha=1} = \sum_{i=1}^N p(x_i) \ln p(x_i) \tag{2.20}$$

it follows that

$$H_R^1(p(x)) = -\sum_{i=1}^N p(x_i) \ln p(x_i) = H(p(x)). \tag{2.21}$$

In other words the Rényi entropies form a family which includes Shannon entropy in the limit $\alpha \to 1$. Though more general than Shannon's, this form has two drawbacks that make it unconvenient in the context of statistical mechanics. First, the entropy of a composite system cannot be expressed easily in terms of entropies of subsystems (*i.e.* by an expression like (2.3)); moreover it lacks the convexity property on which relies much of the connection between statistical mechanics and thermodynamics [49].

2.2.2 Tsallis entropy

Another important measure is known as *Tsallis entropy* [96]

$$H_T^\alpha(p(x)) = \frac{1}{\alpha - 1}\left(1 - \sum_{i=1}^{N} p(x_i)^\alpha\right). \tag{2.22}$$

Again it defines a parametrized family of entropies, which also includes Shannon entropy for $\alpha \to 1$. There is actually a close relationship between Rényi and Tsallis entropies for

$$H_T^\alpha = \frac{1}{\alpha - 1}\left(1 - \exp\left((1-\alpha)H_R^\alpha\right)\right). \tag{2.23}$$

Therefore H_T^α grows monotonously with H_R^α, so that any maximum of one is bound to be a maximum of the other. Tsallis version has the advantage of being strictly concave for $\alpha > 0$ and stricly convex for $\alpha < 0$. However since Rényi entropies are the only measures satisfying extensivity (2.4), this requirement is not fulfilled by (2.22). Considering two independent systems X and Y, we find indeed that their joint entropy is given in this case by

$$\begin{aligned} H_T^\alpha(p(x,y)) &= \frac{1}{\alpha-1}\left(1 - \sum_{i,j=1}^{N} p(x_i,y_j)^\alpha\right) \\ &= \frac{1}{\alpha-1}\left(1 - \sum_{i=1}^{N} p(x_i)^\alpha \sum_{j=1}^{N} p(y_j)^\alpha\right) \\ &= \frac{1}{\alpha-1}\left(1 - (1-(\alpha-1)H_T^\alpha(p(x)))\,(1-(\alpha-1)H_T^\alpha(p(y)))\right) \\ &= H_T^\alpha(p(x)) + H_T^\alpha(p(y)) - (\alpha-1)H_T^\alpha(p(x))H_T^\alpha(p(y)). \end{aligned} \tag{2.24}$$

The Tsallis entropy of the composite system is therefore extensive up to a corrective term, which vanishes in the limit case where Tsallis entropy reduces to Shannon entropy. However, it has been shown [97] that for some correlated systems H_T^α become asymptotically extensive, while this is not the case of Shannon entropy.

2.2.3 Kaniadakis entropy

The last exotic form of entropy that we shall discuss here is *Kaniadakis entropy*, defined as

$$H_K^\kappa(p(x)) = \sum_{i=1}^{N} \frac{p(x_i)^{1+\kappa} - p(x_i)^{1-\kappa}}{2\kappa}. \tag{2.25}$$

Kaniadakis' idea [57, 58] is actually to generalize usual operators in terms of a deformation parameter. Defining for instance the deformed logarithm as

$$\ln_\kappa(x) = \frac{x^\kappa - x^{-\kappa}}{2\kappa}, \tag{2.26}$$

eq. (2.25) can be rewritten as

$$H_K^\kappa(p(x)) = \sum_{i=1}^{N} p(x_i) \ln_\kappa p(x_i). \tag{2.27}$$

Since the deformed logarithm reduces to the usual logarithm when $\kappa \to 0$, the deformed entropy as well reduces to Shannon entropy in this limit. The rationale for introducing such κ-deformed operators stems from the observation that the law of composition of relativistic momenta reduces to simple addition when expressed in terms of κ-deformed sum.

Such modified operators thus appear naturally in a relativistic context. In particular it happens that deformed canonical distributions describe accurately spectral properties of cosmic rays. While it might be thought at first that Kaniadakis entropies can find their use in astrophysical settings exclusively, it has been pointed out that human ecosystems can actually be considered as relativistic inasmuch as information does not propagate immediately. The possiblity of arbitrage opportunities resulting from such relativistic effects was recently investigated in [107].

In what follows we shall keep aside the exotic measures of uncertainty sketched in this section and focus our attention on Shannon entropy. It should however be kept in mind that, at least at a conceptual level, the forthcoming chapters depend little on the choice of entropy itself and the arguments presented therein could possibly be applied with minor modifications to other measures as well. However in the context of the later chapters it is not obvious to us that switching from Shannon's to any other kind of entropy could be beneficial.

2.3 Mutual information

Coming back to Shannon's entropy, we have seen that the entropy of a joint distribution over two variables is given by

$$\begin{aligned} H(p(x,y)) &= -\sum_{i,j} p(x_i, y_j) \ln p(x_i, y_j) \\ &= H(p(y)) + \sum_j p(y_j) H(p(x|y_j)) \\ &= H(p(y)) + H(p(x|y)) \end{aligned} \tag{2.28}$$

where in the last line we defined the *conditional entropy of X knowing Y* as the conditional entropy of X knowing that some particular y_j is realized, averaged over all realizations. Moreover, noting that $f(t) = t \ln t$ is convex, we can use the general property of convex functions that

$$\sum_k a_k f(t_k) \geq f(\sum_k a_k t_k) \tag{2.29}$$

(which holds for non-negative coefficients such that $\sum_k a_k = 1$). Indeed, setting $a_k = p(y_k)$ and $t_k = p(x_i|y_k)$ we can write

$$\left(\sum_k p(y_k)p(x_i|y_k)\right) \ln \left(\sum_k p(y_k)p(x_i|y_k)\right) \leq \sum_k p(y_k)p(x_i|y_k) \ln p(x_i|y_k) \tag{2.30}$$

that is

$$p(x_i) \ln p(x_i) \leq \sum_k p(x_i, y_k) \ln p(x_i|y_k). \tag{2.31}$$

Summing over i yields

$$H(p(x|y)) \leq H(p(x)), \tag{2.32}$$

which means that, in accordance with intuition, knowledge reduces uncertainty. Therefore besides of providing a measure of the uncertainty encoded in a probability distribution, entropy also suggests a characterization of the decrease of uncertainty on a variable brought by the knowledge of another. This is achieved by introducing the *mutual information between X and Y* defined as

$$I(X, Y) = H(p(x)) - H(p(x|y)), \tag{2.33}$$

which, as we said, quantifies to what extent knowledge of a variable Y impacts knowledge on a variable X. Though it is not obvious at first glance that mutual information is symmetric, we can use once more that $H(p(x, y)) = H(p(y)) + H(p(x|y)$ so that we can rewrite (2.33) as

$$I(X, Y) = H(p(x) + H(p(y) - H(p(x, y), \tag{2.34}$$

which displays symmetry explicity.

It might be convenient having at our disposal a normalized version of mutual information in order to make easier comparison between different systems. Such a goal can be achieved by defining

$$i(X,Y) = \frac{I(X,Y)}{\sqrt{H(p(x))H(p(y))}} = \frac{H(p(x) - H(p(x|y)}{\sqrt{H(p(x))H(p(y))}} = \frac{H(p(y) - H(p(y|x)}{\sqrt{H(p(x))H(p(y))}}. \quad (2.35)$$

Indeed, the case where the knowledge of Y determines X completely, and vice versa, should correspond to the largest possible normalized mutual information. Then we have by definition that $H(p(x|y) = H(p(y|x) = 0$, so that (2.35) becomes

$$i(X,Y) = \frac{H(p(x)}{\sqrt{H(p(x))H(p(y))}} = \frac{H(p(y)}{\sqrt{H(p(x))H(p(y))}}, \quad (2.36)$$

whence $i(X,Y)^2 = 1$ and $i(X,Y) = 1$.

2.4 Kullback-Leibler divergence and total information

2.4.1 Mutual information as a Kullback divergence

Mutual information can of course be defined for two groups of variables as well as for two primitive variables; however in any case it remains a fundamentally bivariate measure (albeit it can be generalized in several other respects, one of the best known generalization being the *transfer entropy* proposed by [89] as an oriented measure of information flows in data). When more than two variables (or groups of variables) are involved, it is not immediately obvious what would be the most appropriate generalization and different proposals have been made in the literature. A possible generalization flows naturally from rewriting (2.33) directly in terms of probability distributions as

$$I(X,Y) = \sum_{i,j} p(x_i, y_j) \ln \frac{p(x_i, y_j)}{p(x_i)p(y_j)}. \quad (2.37)$$

This expression happens to be a particular instance of a tool which is widely used in probability theory, known as *Kullback-Leibler (KL) divergence*. The Kullback-Leibler divergence between two distributions $p(x)$ and $q(x)$ is defined as

$$D(p,q) = \sum_{i,j} p(x_i) \ln \frac{p(x_i)}{q(x_i)}. \quad (2.38)$$

The KL divergence has the property of being positive except in the case where $p = q$ (then $D(p,p) = 0$), so that it is sometimes employed as a measure of distance over the space of probability distributions. However this analogy should not be pushed too far since it must be noted that $D(p,q) \neq D(q,p)$ and that it does not satisfy the triangular inequality; thus it cannot qualify as a *distance* properly said[4].

[4]We shall therefore refrain from calling it *KL distance* as is often the case in the litterature. Needless to say it is not a divergence in the sense of vector analysis either.

From (2.37) the mutual information between two variables can therefore be envisaged as the Kullback-Leibler divergence between the joint distribution describing the variables and the product of marginal distributions, or in other words as their "distance" from statistical independence. When $I(X,Y) = 0$ we have $p(x,y) = p(x)p(y)$ and the variables are independent of each other.

2.4.2 Total information

Letting $X^1, ..., X^n$ be n random variables, we are thus led to introduce a statistical measure that quantifies the divergence between their joint distribution and statistical independence, namely

$$T(X^1, ..., X^n) = \sum_{i_1,...,i_n} p(x^1_{i_1}, ..., x^n_{i_n}) \ln \frac{p(x^1_{i_1}, ..., x^n_{i_n})}{\prod_{k=1}^n p(x^k_{i_k})}. \tag{2.39}$$

This quantity is known as *total information* or *multi-information* [105]. It shares with mutual information the property of dealing with all variables on the same footing. In terms of entropy we can rewrite the total information as

$$T(X_1, ..., X_n) = \sum_{i=1}^n H(p(x_i)) - H(p(x_1, ..., x_n)). \tag{2.40}$$

Like for the bivariate mutual information, it might be convenient having a normalized version of total information. The reasoning leading to the normalized mutual information (2.35) can fortunately be generalized as follows. Let us rewrite first (a bracketed object means it is excluded from the summation/sequence in which it appears)

$$\begin{aligned} T(X^1, ..., X^n) &= \sum_{i_1,...,i_n} p(x^1_{i_1}, ..., x^n_{i_n}) \ln \frac{p(x^1_{i_1}, ..., [x^j_{i_j}], ..., x^n_{i_n} | x^j_{i_j})}{\prod_{k \neq j} p(x^k_{i_k})} \\ &= \sum_{i=1 \neq j}^n H(p(x^i)) - H(p(x^1, ..., [x^j], ..., x^n | x^j)). \end{aligned} \tag{2.41}$$

If the knowledge of X^j determines all other variables then the last term on the r.h.s. vanishes and $T(X^1, ..., X^n) = \sum_{i=1 \neq j}^n H(p(x^i))$. Let us now introduce the notation $G_j = \sum_{i=1 \neq j}^n H(p(x^i))$. We define the *normalized total information* as

$$\tau = \frac{T}{(\prod_{i=1}^n G_i)^{1/n}}. \tag{2.42}$$

In the case where X^j fully characterizes all others variables the expression (2.41) for T allows to rewrite

$$\tau = \frac{G_j}{(\prod_{i=1}^n G_i)^{1/n}}. \tag{2.43}$$

Now consider the situation where *any* variable determines all others, which obviously has the largest information content. In that case the above relationship holds for any j, and multiplying together the n such expressions yields

$$\tau^n = \frac{G_1 ... G_n}{\prod_{i=1}^n G_i} = 1 \qquad \rightarrow \qquad \tau = 1. \tag{2.44}$$

Therefore the normalized total information of the most fully inter-correlated network of variables does not exceed 1 as required, while the lower bound at 0 is obvious.

2.5 Mutual information versus correlation

2.5.1 Correlation

We have seen that mutual information flowed naturally from entropy as a measure of codependence. However, codependence is traditionally measured by covariance and correlation. The *covariance of two random variables X, Y* is defined as

$$C_{X,Y} = \langle XY \rangle - \langle X \rangle \langle Y \rangle = \sum_{i,j} x_i y_j p(x_i, y_j) - \left(\sum_i x_i p(x_i) \right) \left(\sum_j y_j p(y_j) \right). \tag{2.45}$$

Covariance is usually replaced by the so-called *(coefficient of) correlation*

$$\rho_{X,Y} = \frac{C_{X,Y}}{\sigma(X)\sigma(Y)} = \frac{\langle XY \rangle - \langle X \rangle \langle Y \rangle}{\sqrt{\langle X^2 \rangle - \langle X \rangle^2} \sqrt{\langle Y^2 \rangle - \langle Y \rangle^2}} \tag{2.46}$$

which is normalized in the sense that it takes its values in $[-1, 1]$. The first important thing to note is that $\rho_{X,Y}$ lets intervene the values of the variables themselves. This is an important difference with mutual information, which depends only on the probability distribution characterizing the variables.

Two random variables are completely (anti-)correlated (*i.e.* their correlation coefficient takes value ± 1) if, and only if, there is a linear relationship between them. The proof relies on the Cauchy-Schwarz inequality

$$\langle XY \rangle^2 \leq \langle X^2 \rangle \langle Y^2 \rangle \tag{2.47}$$

which is proved as follows : for a, b in $\mathbb{R}$, we have

$$0 \leq \langle (aX - bY)^2 \rangle = a^2 \langle X^2 \rangle - 2ab \langle XY \rangle + b^2 \langle Y^2 \rangle. \tag{2.48}$$

The r.h.s. being a polynomial quadratic in a, it vanishes at most in one point, from

what we can conclude that its discriminant has to be smaller than or equal to zero, that is

$$\Delta = 4b^2\langle XY\rangle^2 - 4b^2\langle X^2\rangle\langle Y^2\rangle \leq 0$$

which is precisely (2.47). In order for the inequality to get saturated, we have necessarily $\langle (aX - bY)^2\rangle = 0$, hence $aX = bY$.

The proof of the proposition for given variables X, Y now follows easily applying the Cauchy-Schwarz inequality to the centered variables $W = X - \langle X\rangle$ and $Z = Y - \langle Y\rangle$. Indeed,

$$\langle WZ\rangle^2 \leq \langle W^2\rangle\langle Z^2\rangle.$$

As shown above, this inequality is saturated if and only if there exists a linear relationship between W and Z, and thus between X et Y. But saturation implies

$$\langle XY - Y\langle X\rangle - X\langle Y\rangle + \langle X\rangle\langle Y\rangle\rangle^2 = \langle X^2 - 2X\langle X\rangle + \langle X\rangle^2\rangle\langle Y^2 - 2Y\langle Y\rangle + \langle Y\rangle^2\rangle$$

or

$$\langle XY\rangle^2 - 2\langle X\rangle\langle Y\rangle\langle XY\rangle + \langle X\rangle^2\langle Y\rangle^2 = (\langle X^2\rangle - \langle X\rangle^2)(\langle Y^2\rangle - \langle Y\rangle^2)$$

which amounts exactly to $\rho^2_{X,Y} = 1$. It would be too strong asserting that correlation is blind to any dependence other than linear since two variables that are not linearly related can be strongly correlated; nonetheless the above shows that *maximal* correlation requires linear dependence.

A standard example will illustrate this fact: let $\omega \in \{-1, 0, 1\}$ uniformly distributed and consider the two random variables $X = \omega$ and $Y = |\omega|$. Of course $\langle X\rangle = 0$ and $\langle Y\rangle = 2/3$, while $\langle XY\rangle = 0$, and therefore $\rho_{X,Y} = 0$ even though these variables are tightly dependent. On the other side mutual information information can be computed using that $H(X) = -\log(1/3)$, $H(Y) = -\log(1/3) - 2/3$, and $H(X, Y) = -\log(1/3)$, so that $I(X, Y) = -\log(1/3) - 2/3 \sim 0.92 \neq 0$. Mutual information is thus able to capture the non-linear dependence relating X and Y.

The second issue about correlation we would like to point out arises when the variables considered are identically distributed. The correlation coefficient becomes

$$\rho = \frac{1}{\sigma^2}\left(\sum_{i,j=1}^{n} x_i x_j p(x_i, x_j) - \left(\sum_{i=1}^{n} x_i p(x_i)\right)^2\right). \tag{2.49}$$

We can cast (2.49) in matrix form defining $\mathbf{x} = (x_1, ..., x_n)$ and $\mathbf{C}$ such that $C_{ij} \equiv p(x_i, x_j) - p(x_i)p(x_j)$. Eq. (2.49) becomes then

$$\rho = \frac{1}{\sigma^2} \cdot \mathbf{x}\mathbf{C}\mathbf{x}^T. \tag{2.50}$$

Splitting now $\mathbf{C} = \mathbf{S} + \mathbf{A}$ into its symmetric and antisymmetric parts, it is fairly obvious that [65]

$$\rho = \frac{1}{\sigma^2} \cdot \mathbf{x}(\mathbf{S} + \mathbf{A})\mathbf{x}^T = \frac{1}{\sigma^2} \cdot \mathbf{x}\mathbf{S}\mathbf{x}^T$$

In other words, although the $\mathbf{C}$ matrix encodes all there is to know regarding the dependance between the variables, in the case of identically distributed variables the autocorrelation coefficient is blind to that amount of information which is displayed by the antisymmetric part of $\mathbf{C}$. This issue does not arise in general since $\mathbf{x}\mathbf{A}\mathbf{y}^T$ does not vanish generally speaking.

From the properties discussed above, it therefore appears that in spite of its wide use, correlation might not be the most appropriate tool in the aforementioned situations. Although the second point might seem of narrow relevance, it should be paid attention in the context of complex systems where we usually deal with large ensembles of essentially identical entities.

2.5.2 The Gaussian case

For completeness it might be useful to calculate explicitely correlation and mutual information for the archetypal case of a (two-dimensional) gaussian vector $\mathbf{X}$. Let us remind that we call so a vector formed by n random variables if any linear combination of these variables follows a gaussian distribution. Such a vector is said to be distributed with $\mathcal{N}(\mu, \mathbf{C})$ and its joint distribution law is (we put $C \equiv \det \mathbf{C}$)

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n C}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \tag{2.51}$$

The covariance of such a vector is given by $Cov(X_i, X_j) = C_{ij}$. For instance in the case $n = 2$ we have

$$\begin{aligned} p(x,y) &= \frac{1}{2\pi\sqrt{C}} \exp\left(-\frac{1}{2}\sum_{i,j=1}^{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\ &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)\right) \end{aligned} \tag{2.52}$$

where we used that

$$\mathbf{C} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

and thus immediately

$$\mathbf{C}^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_x^2} & -\frac{\rho}{\sigma_x \sigma_y} \\ -\frac{\rho}{\sigma_x \sigma_y} & \frac{1}{\sigma_y^2} \end{pmatrix}.$$

The Shannon entropy of $p(\boldsymbol{x})$ can easily be calculated in the general case to give

$$\begin{aligned} H(p(\mathbf{x})) &= -\int p(\mathbf{x}) \ln p(\mathbf{x}) \\ &= \ln \sqrt{(2\pi)^n C} \int p(\mathbf{x}) + \frac{1}{2} \int p(\mathbf{x}) (\mathbf{x}-\mu)^T \mathbf{C}^{-1} (\mathbf{x}-\mu) \\ &= \frac{1}{2} \ln \left((2\pi)^n C\right) + \frac{1}{2} \langle \sum_{ij} (X_i - \mu_i)(C^{-1})_{ij} (X_j - \mu_j) \rangle \\ &= \frac{1}{2} \ln \left((2\pi)^n C\right) + \frac{1}{2} \sum_{ij} \langle (X_j - \mu_j)(X_i - \mu_i) \rangle (C^{-1})_{ij} \\ &= \frac{1}{2} \ln \left((2\pi)^n C\right) + \frac{1}{2} \sum_{ij} C_{ji} (C^{-1})_{ij} \\ &= \frac{1}{2} \ln \left((2\pi)^n C\right) + \frac{1}{2} \sum_{j} (C C^{-1})_{jj} \\ &= \frac{1}{2} \ln \left((2\pi)^n C\right) + \frac{n}{2} \\ &= \frac{1}{2} \ln \left((2\pi e)^n C\right) \end{aligned} \tag{2.53}$$

In the case $n = 2$ that we shall be chiefly interested in, this result becomes in particular $H(p(x,y)) = \ln(2\pi e\sqrt{C}) = \ln(2\pi e \sigma_X \sigma_Y \sqrt{1-\rho^2})$.

The calculation of the mutual information between the two components of our vector still requires the individual distribution laws. They would be obtained in full generality by marginalizing the joint probability, but in this case it is much easier to note that each component of a Gaussian vector is itself Gaussian, as follows immediately from the definition. Thus

$$H(p(x)) = \frac{1}{2} \ln(2\pi e \sigma_X^2)$$

$$H(p(y)) = \frac{1}{2} \ln(2\pi e \sigma_Y^2)$$

and we get finally

$$
\begin{aligned}
I(X,Y) &= H(p(x)) + H(p(y)) - H(p(x,y)) \\
&= \frac{1}{2}\ln(2\pi e\sigma_X^2) + \frac{1}{2}\ln(2\pi e\sigma_Y^2) - \ln(2\pi e\sigma_X\sigma_Y\sqrt{1-\rho^2}) \\
&= -\frac{1}{2}\ln\left(1-\rho^2\right). \qquad (2.54)
\end{aligned}
$$

So it appears that in the case of a Gaussian vector, there is a direct relationship between correlation and mutual information. This should not come as a surprise since it is well known that in the case of components of a Gaussian vector there is a prefect overlap between (un)correlation and (in)dependance, of which our example provides an illustration.

We shall see in the next chapter that given a random vector $\mathbf{X}$, the distribution that maximizes entropy among all centered distributions with fixed covariance matrix is the Gaussian distribution. The calculation above therefore has the side benefit of providing and upper bound $H(p(\mathbf{x})) \leq \frac{1}{2}\ln((2\pi e)^n C)$.

2.6 Information geometry and Fisher information

Entropy, total information and Kullback-Leibler divergence are suited to non-parametric as well as to parametric distributions. Some concepts of information theory however make sense exclusively in the context of parametric distributions, as is the case of *Fisher information* [31]. Let $p(x|\theta)$ denote a probability distribution for a random variable X conditional on a parameter θ. The density defines a likelihood function $\mathcal{L}$ for θ such that $\mathcal{L}(\theta|x) = p(x|\theta)$ is the likelihood that θ rules the distribution p, given the fact that the outcome x is observed. It is then usual to define the *score* as

$$
S = \frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta|x).
$$

S characterizes the sensitivity of the likelihood function to the parameter. The expectation value of the score is easily shown to vanish since

$$
\begin{aligned}
\mathbb{E}[S|\theta] &= \int dx \left(\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta|x)\right) p(x|\theta) \\
&= \int dx \left(\frac{1}{p(x|\theta)}\frac{\partial p(x|\theta)}{\partial\theta}\right) p(x|\theta) \\
&= \frac{\partial}{\partial\theta}\int dx p(x|\theta) \\
&= 0.
\end{aligned}
$$

Fisher information is then defined as the variance of the score:

$$J(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta|x)\right)^2|\theta\right].$$

The relevance of Fisher information to statistical inference stems from the Cramér-Rao bound [31], that states that the inverse of $J(\theta)$ provides a lower bound on the variance of any unbiased estimator of θ.

In the following our focus is strictly on non-parametric distributions and we shall not comment further on Fisher information. The information theory of parametric inference however gives rise to rich mathematical developments since considering parameters as coordinates in the space of distributions allows bringing in the full apparatus of differential geometry; Fisher information can then be envisaged as a metric on this space (see [1] for an exhaustive introduction, or [18] for a physics-related perspective; the relevance of Fisher information to reaction-diffusion systems was investigated within the Sophocles project in [47]).

Chapter 3

Maximum Entropy principle(s)

Armed with the notion of entropy as a measure of uncertainty, we now turn our attention to the *principle of maximum entropy* (MEP) around which revolves the remaining of this thesis. The term *criterion of maximum entropy* could actually be preferred since it states clearly what the MEP precisely is - a heuristic criterion selecting as least biased distribution the one having the largest uncertainty - and we shall use this term occasionally. However, the word *principle* carries a somewhat holistic and imprecise meaning that matches particularly well the ambiguous epistomological status of the MEP. On the one side, the MEP is subjective inasmuch as it relies on a notion of uncertainty, which itself, as we have explained in chapter 2, is built axiomatically. On the other side, we shall see in section 3.4 that the maximum entropy distribution can be justified by an objective argument about the most probable distribution. This versatility may be a source of confusion that we shall try to dissipate when possible.

Technically speaking, we draw a distinction between maximum entropy distributions that are inferred from empirical constraints on expectation values of observables (section 3.1), and distribution that are inferred from constraints on marginal distributions (section 3.2). The former are used in chapter 4, while the latter lie at the ground of chapters 5 and 6. In section 3.3 we touch on the case of constrained pairwise mutual information. This hybrid case is interesting from an academic standpoint, although it raises technical issues that make it unlikely to be of any practical use in a nearby future.

3.1 MEP based on expectation values

3.1.1 The idea

Entropy has been a key concept of statistical physics, if not from the beginning of thermodynamics, at least from Clausius' discovery of the second principle of thermodynamics. It was however understood as a quantity whose existence was rooted in physical reality. Around the turn of the 20th century the pioneers of statistical mechanics had made clear that entropy could be related to the distribution function describing a system, but it was Shannon's achievement (1948) to highlight its intrinsic relevance to

probability theory. It is therefore no coincidence that within a few years (1957) Jaynes proposed to unstrip statistical mechanics from its mechanical clothes so as to make it an essentially statistical theory [36].

The philosophy underlying his approach is opposite to the standard way of thinking, where a structured model is built and tuned to match observed properties. In the maximum entropy approach on the other side, we proceed by seeking the least structured model compatible with a given set of observations. This is done by noting that this least structured probability density is the one which has the largest entropy while still satisfying observational constraints, which are usually provided by some set of observables f_k $(k = 1, 2, ..., K)$ the average values of which are known, $\langle f_k \rangle = \mu_k$.

Assume we are looking for a probability distribution p on a set of N variables $X_1, ..., X_N$ collectively denoted by $\mathbf{X}$, such that $H(\mathbf{X}) = -\sum_{\mathbf{x}} p(\mathbf{x}) \ln p(\mathbf{x})$ is maximal, while making sure that the constraints $\langle f_k \rangle := \sum_{\mathbf{x}} f_k(\mathbf{x}) p(\mathbf{x}) = \mu_k$ and $\sum_{\mathbf{x}} p(\mathbf{x}) = 1$ are enforced. Using Lagrange's multipliers, the problem can be formulated as finding a solution to the equation

$$\begin{aligned} 0 &= \frac{\partial}{\partial p(\mathbf{y})} \left(-\sum_{\mathbf{x}} p(\mathbf{x}) \ln p(\mathbf{x}) + \lambda_0 \left(\sum_{\mathbf{x}} p(\mathbf{x}) - 1 \right) + \sum_{k=1}^{K} \lambda_k \left(\sum_{\mathbf{x}} f_k(\mathbf{x}) p(\mathbf{x}) - \mu_k \right) \right) \\ &= -\sum_{\mathbf{x}} \delta_{\mathbf{x},\mathbf{y}} \ln p(\mathbf{x}) - \sum_{\mathbf{x}} \delta_{\mathbf{x},\mathbf{y}} + \lambda_0 \sum_{\mathbf{x}} \delta_{\mathbf{x},\mathbf{y}} + \sum_{k=1}^{K} \lambda_k \sum_{\mathbf{x}} f_k(\mathbf{x}) \delta_{\mathbf{x},\mathbf{y}} \\ &= -\ln p(\mathbf{y}) - 1 + \lambda_0 + \sum_{k=1}^{K} \lambda_k f_k(\mathbf{y}), \end{aligned} \tag{3.1}$$

where the λ's are the multipliers. Therefore the sought-after distribution may be written as

$$p(\mathbf{x}) = \frac{1}{Z} \exp \left(\sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}) \right), \tag{3.2}$$

where the multipliers have to be chosen to match the constraints imposed on normalization and averages. Dividing by the partition function $Z = \sum_{\mathbf{x}} \exp \left(\sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}) \right)$ ensures that p is properly normalized.

3.1.2 Examples

1. *Uniform distribution.* In the trivial case where no constraint besides normalization is imposed, we have $\lambda_k = 0 \ \forall k$ and the maximum entropy distribution is nothing but the uniform distribution $p(\mathbf{x}) = 1/Z$.

2. *Gaussian distribution.* When considering unbounded continuous variables it is mandatory to impose constraints besides normalization in order to avoid that the distribution vanishes everywhere. Considering one such variable x, we can

constrain its mean and variance so that $\langle x \rangle = \mu$ and $\langle (x-\mu)^2 \rangle = \sigma^2$. From (3.2) follows that

$$p(x) = \frac{1}{Z} \exp\left(\lambda_1 x + \lambda_2 (x-\mu)^2\right) \tag{3.3}$$

in which after some manipulations we can recognize a Gaussian distribution. Since the entropy of a Gaussian distribution is easily computed (see chapter 2) this provides us with an upper bound on the entropy of any distribution satisfying given mean and variance.

3. *Gibbs canonical distribution.* The archetypal case that led to the maximum entropy formulation of statistical mechanics is the case where an energy $E(\mathbf{x}_i)$ is associated to each state $\mathbf{x}_i$ of a physical system formed by an assembly of N particles, and where the only constraint at hand is on the average energy $\langle E \rangle$ of the system. Then (3.2) becomes

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left(\lambda E(\mathbf{x})\right), \tag{3.4}$$

which, once solved for λ, reduces to Gibbs' distribution[1] on which is build most of statistical mechanics. Therefore while Gibbs distribution and all subsequent results can be obtained in the standard approach from non-trivial dynamical considerations, things are much simpler in Jaynes formalism which can be summarized in two steps: 1) the observational quantity one is chiefly interested in is usually the total energy of a system; 2) knowing the average value of this quantity - and nothing else - one is led to assume that the distribution characterizing the system is given by (3.4); any other guess would bring unfounded biases into the play. In Jaynes' own words [52]:

> *(...) there is nothing in the general laws of motion that can provide us with any additional information about the state of a system beyond what we have obtained from measurement.*

This quote provides an occasion to underline that the maximum entropy principle has nothing to say about the number of measurements that are necessary to exhaust the underlying richness of the system investigated. Of course if other observables are considered (3.4) has to be amended so as to take them into account using (3.2). This point will be discussed further in chapter 6, where we investigate the consequences of considering bilocal observables in a gas of corpuscules undergoing binary collisions.

[1] Gibbs distribution is meant here in the sense of classical statistical mechanics; we do not make reference here to the more sophisticated notion of Gibbs-Bowen measures used in mathematical statistical mechanics [30].

3.1.3 Relation to Ising model

The maximum entropy principle based on averages is made extremely appealing to the physicists by the following fact: assume not only the mean $\langle x_i \rangle$ of each variable is known, but also the pairwise correlation $\langle x_i x_j \rangle$ for each pair of variables. Then (3.2) becomes

$$p(\mathbf{x}) = \frac{1}{Z} \exp \left(\sum_{k=1}^{N} \lambda_k x_k + \sum_{i,j=1}^{N} J_{ij} x_i x_j \right). \tag{3.5}$$

When - whether we assume homogeneity or not - only the nearest neighbour correlation function is known and not the detailed structure of correlations (and similarly for the mean), eq. (3.5) in turn reduces to

$$p(\mathbf{x}) = \frac{1}{Z} \exp \left(\lambda \sum_{k=1}^{N} x_k + J \sum_{\langle i,j \rangle = 1}^{N} x_i x_j \right) \tag{3.6}$$

which is none but the canonical distribution of an Ising model. A bit more daring statement would be that any system which is known only through its empiral correlation function can be expressed as a generalized Ising model, and this analogy prompted widespread applications of this result to neural networks [87], bird flocks [14], and economics (see the recent detailed study of the structure of stock markets in [20]) among others. In particular, [87] highlighted that maximum entropy distributions inferred from weak pairwise correlations are actually quite different from factorized distributions when considered globally, so that observing non-correlation is - generally speaking - as relevant as observing correlation.

However, this analogy between (3.5), (3.6) and Ising models should not be taken soo seriously, since the fact that a system is described at equilibrium by the canonical distribution of an Ising model does not necessarily imply that it actually obeys an Ising dynamics; moreover (3.6) depends on our subjective knowledge of the system, which is irrelevant for the dynamics itself. The couplings inferred from (3.6) or (3.5) do not in general reflect faithfully the actual couplings. We shall face similar issues in chapter 5 when trying to infer orders of interaction in a system from its marginal distributions.

3.1.4 Systems out of equilibrium

The maximum entropy principle in its original formulation applies only to systems at equilibrium, since measured quantities that serve as constraints need to be deduced from some historical sample of finite length. However, the assumption of stationarity in this formulation of statistical mechanics is imposed by 'statistics' (or rather by the kind of data used for inference) and not by 'physics' itself. It was therefore natural to try to extend the maximum entropy approach to non-equilibrium. Early attempts tend to 'temporalize' equilibrium distributions by considering a state space enlarged so as to encompass the full orbit of a system [54], and this approach is still very much in fashion [22, 38, 39, 37, 41]. Another approach [100] is to implement a maximum entropy criterion directly in the dynamics generating the process. The inherent technical difficulties

will be discussed in chapter 4.

3.2 MEP based on marginals

Although maximum entropy distributions were initially derived for constraints given on expectation values of some set of observables, this is not the only kind of observational constraint one can think of. In chapters 5 and 6 below the main role is taken over by constraints put on marginal distributions characterizing subsets of our N variables. While both cases seem to be quite different from each other, it is fortunate that the former is general enough to englobe the latter. Actually marginals can be seen as averages of counting operators, at the price of some δ-functions gymnastics.

Let us take for illustration the case of four variables (*i.e.* $\mathbf{x} = (w, x, y, z)$), and assume the tri-variate marginal $p_{123}(a, b, c)$ is known. Putting $f(\mathbf{x}) = \delta(w, a)\delta(x, b)\delta(y, c)$ allows writing

$$\begin{aligned}\langle f \rangle &= \sum_{\mathbf{x}} f(\mathbf{x})p(\mathbf{x}) \\ &= \sum_{w,x,y} \delta(w, a)\delta(x, b)\delta(y, c) \sum_{z} p(\mathbf{x}) \\ &= \sum_{w,x,y} \delta(w, a)\delta(x, b)\delta(y, c)p_{123}(w, x, y) \\ &= p_{123}(a, b, c). \end{aligned} \tag{3.7}$$

Applying the general result (3.2) to all possible values of the arguments then yields

$$\begin{aligned}p(\mathbf{x}) &= \frac{1}{Z} \exp\left(\sum_{a,b,c} \lambda(a, b, c)f(\mathbf{x})\right) \\ &= \frac{1}{Z} \exp\left(\sum_{a,b,c} \lambda(a, b, c)\delta(w, a)\delta(x, b)\delta(y, c)\right) \\ &= \frac{1}{Z} \exp\left(\lambda(w, x, y)\right) \end{aligned} \tag{3.8}$$

where λ now denotes a well-chosen multiplying *function*. The result (3.8) generalizes straightforwardly to any number of marginals; for instance, if besides p_{123} the marginals p_{124} and p_{34} are fixed we get

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left(\lambda_1(w, x, y) + \lambda_2(w, x, z) + \lambda_3(y, z)\right) \tag{3.9}$$

for some functions $\lambda_1, \lambda_2, \lambda_3$. The general expression is maybe easier to state in plain

words than to write down symbolically: *knowing a set of any number of marginals of any order (whether the same for each marginal considered or not), the maximum entropy estimate of the joint distribution is a product of functions each taking as arguments the arguments of the corresponding marginal.*

In spite of its elegance, this result is sadly of little practical use since the determination of the multiplying functions is a difficult problem. There is however an important exception in the simple case when constrained marginals are univariate. Then $\lambda(w) = \ln p_1(w)$ (and $\lambda(x) = \ln p_2(x)$, *etc.*) obviously satisfies the requirements; in other words under constrained univariate marginals the distribution having the largest entropy is the factorized distribution. In all other cases, we have to resort to the so-called *iterative scaling algorithm* [19] which allows numerical calculations. Iterative scaling will be exposed in chapter 5 where it is used extensively.

3.3 Constraining general codependences

We explained in the previous chapter that correlation as a measure of codependence was not adequate for assessing non-linear dependences. One might therefore think that in defining maximum entropy models based on constrained correlations we give up in a sense the flexibility of the tools provided by information theory. In particular, it seems natural to specify the mutual information between variables instead of their correlation. Curiously, such models have never been investigated previously to the best of our knowledge, so that it might be the place to make some remarks.

Mutual information expressed as (2.37) can be considered as an average in a generalized sense since

$$I(X,Y) = \langle \ln \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \rangle. \tag{3.10}$$

However this particular kind of average lets intervene p itself, so that the result (3.2) cannot be carried over as such and it is necessary to revert to the original minimization problem. Assuming for definiteness that the mutual information between variables X_i and X_j is fixed to $I(X_i, X_j) = I_{ij}$, the functional to minimize becomes

$$-\sum_{\mathbf{x}} p(\mathbf{x}) \ln p(\mathbf{x}) + \lambda_0 \left(\sum_{\mathbf{x}} p(\mathbf{x}) - 1 \right) + \lambda \left(\sum_{x_i, x_j} p(x_i, x_j) \ln \frac{p(x_i, x_j)}{p(x_i)p(x_j)} - I_{ij} \right). \tag{3.11}$$

The minimization problem is therefore

$$
\begin{aligned}
0 &= \frac{\partial}{\partial p(\mathbf{y})}\left(-\sum_{\mathbf{x}} p(\mathbf{x})\ln p(\mathbf{x}) + \lambda_0\left(\sum_{\mathbf{x}} p(\mathbf{x}) - 1\right)\right) \\
&\quad + \lambda\frac{\partial}{\partial p(\mathbf{y})}\left(\sum_{x_i,x_j} p(x_i,x_j)\ln\frac{p(x_i,x_j)}{p(x_i)p(x_j)} - I_{ij}\right) \\
&= -\ln p(\mathbf{y}) - 1 + \lambda_0 + \lambda\frac{\partial}{\partial p(\mathbf{y})}\left(\sum_{x_i,x_j} p(x_i,x_j)\ln\frac{p(x_i,x_j)}{p(x_i)p(x_j)} - I_{ij}\right). \qquad (3.12)
\end{aligned}
$$

Developing the last term we get

$$
\begin{aligned}
&\frac{\partial}{\partial p(\mathbf{y})}\left(\sum_{x_i,x_j} p(x_i,x_j)\ln\frac{p(x_i,x_j)}{p(x_i)p(x_j)}\right) \\
&= \sum_{x_i,x_j}\left(\frac{\partial p(x_i,x_j)}{\partial p(\mathbf{y})}\ln\frac{p(x_i,x_j)}{p(x_i)p(x_j)} + p(x_i)p(x_j)\frac{\partial}{\partial p(\mathbf{y})}\left(\frac{p(x_i,x_j)}{p(x_i)p(x_j)}\right)\right) \\
&= \sum_{x_i,x_j}\left(\frac{\partial p(x_i,x_j)}{\partial p(\mathbf{y})}\ln\frac{p(x_i,x_j)}{p(x_i)p(x_j)} + \frac{\partial p(x_i,x_j)}{\partial p(\mathbf{y})} - \frac{p(x_i,x_j)}{p(x_i)}\frac{\partial p(x_i)}{\partial p(\mathbf{y})} - \frac{p(x_i,x_j)}{p(x_j)}\frac{\partial p(x_j)}{\partial p(\mathbf{y})}\right) \\
&= \sum_{x_i,x_j}\left(\delta_{x_i,y_i}\delta_{x_j,y_j}\ln\frac{p(x_i,x_j)}{p(x_i)p(x_j)} + \delta_{x_i,y_i}\delta_{x_j,y_j} - \frac{p(x_i,x_j)}{p(x_i)}\delta_{x_i,y_i} - \frac{p(x_i,x_j)}{p(x_j)}\delta_{x_j,y_j}\right) \\
&= \ln\frac{p(y_i,y_j)}{p(y_i)p(y_j)} + 1 - \sum_{x_j}\frac{p(y_i,x_j)}{p(y_i)} - \sum_{x_i}\frac{p(x_i,y_j)}{p(y_j)} \\
&= \ln\frac{p(y_i,y_j)}{p(y_i)p(y_j)} - 1 \qquad (3.13)
\end{aligned}
$$

so that (3.12) becomes

$$
-\ln p(\mathbf{y}) - 1 + \lambda_0 + \lambda\left(\ln\frac{p(y_i,y_j)}{p(y_i)p(y_j)} - 1\right) = 0, \qquad (3.14)
$$

which is the condition that any solution of the problem has to fulfill. Exponentiating both sides and rearranging, we get that

$$
p(\mathbf{y})\left(\frac{p(y_i,y_j)}{p(y_i)p(y_j)}\right)^{-\lambda} = e^{\lambda_0-\lambda-1}. \qquad (3.15)
$$

The case where the mutual information between each pair of variables is constrained is deduced similarly. Equation (3.14) becomes then

$$-\ln p(\mathbf{y}) - 1 + \lambda_0 + \sum_{i,j>i} \lambda_{ij} \left(\ln \frac{p(y_i, y_j)}{p(y_i)p(y_j)} - 1 \right) = 0 \tag{3.16}$$

leading to

$$p(\mathbf{y}) \prod_{i,j>i} e^{\lambda_{ij}} \left(\frac{p(y_i, y_j)}{p(y_i)p(y_j)} \right)^{-\lambda_{ij}} = e^{\lambda_0 - 1}. \tag{3.17}$$

We can recognize here some similarities with the case of constrained marginals, raising (at least) equivalent technical difficuties. This probably explains why, as far as we know, this approach has not been developed so far.

3.4 Wallis' argument

In section 3.1 we based our justification of the maximum entropy criterion on entropy envisaged as a measure of uncertainty. Would the reader be reluctant to rely on arguments letting subjectivity come into the play, we would like to present in conclusion of this chapter an alternative justification known as *Wallis' argument* [55].

Suppose an operator has to assign K quanta (balls) of probability (each worth $1/K$) to N possible outcomes (buckets) in order to produce *some* probability distribution that agrees with a set of constraints at his disposal. Let the operator throw the balls randomly, and k_i denote the number of balls falling in bucket i (thus the i-th outcome is assigned probability $p_i = k_i/K$). If the resulting distribution disagrees with the constraints it is rejected and the operator tries again; if the distribution agrees with the contraints it is saved in memory, after what the operator starts again.

Among the possible distributions generated by this experiment, some are more or less likely depending on their multiplicity m, given by

$$m = \frac{K!}{\prod_{i=1}^{N} k_i!}. \tag{3.18}$$

In particular, the most probable distribution is the one maximizing m or, equivalently, maximizing any monotonic function of m, like $(\ln m)/K$. But we have

$$\begin{aligned} \frac{\ln m}{K} &= \frac{1}{K} \left(\ln K! - \sum_{i=1}^{N} \ln(Kp_i)! \right) \\ &\approx \frac{1}{K} \left(K \ln K - \sum_{i=1}^{N} (Kp_i) \ln(Kp_i) \right) \\ &= -\sum_{i=1}^{N} p_i \ln p_i, \end{aligned} \tag{3.19}$$

where the approximation results from Stirling's approximation for $K \to \infty$. Therefore, it happens that when the number of quanta of probability to be attributed is large, the distribution having the largest entropy is also the most probable one.

Wallis' argument thus provides a justification that has the merit of being more objective than the classical one based on uncertainty. It also has the merit of being an adaptation of a well-known argument in classical statistical mechanics used for deriving the most probable distribution of corpuscules in a gas [49].

Chapter 4

Maximum entropy reconstruction of time series

We mentioned in the previous chapter that the maximum entropy principle in its original formulation applies only to systems at equilibrium, since measured quantities implemented as constraints need to be deduced from a historical sample of finite length. However, the assumption of stationarity in this formulation of statistical mechanics is imposed by 'statistics' (or rather by the kind of data used for inference) and not by 'physics' itself. It was therefore natural to try to extend the maximum entropy approach to non-equilibrium. The obvious way to address the question is to enlarge the state space under consideration by encompassing not only configurations at a given time, but trajectories of the system themselves [54]. This direct approach has met a considerable popularity [22, 38, 39, 37, 41], possibly due to its ressemblance with the path integral formulation of quantum (field) theory [40, 60]. However a drawback of this approach is that it requires 'freezing' samples of a given length so as to compute the associated entropy.

In this chapter we consider another approach due to Van der Straten [100] which implements a maximum entropy criterion directly in the dynamics generating the process. The major technical difference compared to the material presented in chapter 3 is that we now consider the *entropy rate* of stochastic processes, which is essentially the entropy per symbol of the process as introduced in section 4.1. This makes the corresponding minimization problem considerably more involved; it is solved in section 4.3 under simplifying assumption time reversibility, while particular cases are considered in sections 4.4 and 4.5.

The key intuition behind this chapter is that since the maximum entropy criterion allows us to build probability distributions from observables whose measurement is not demanding in terms of the quantity of data, it should also, when applied to temporal processes, allow us to build stochastic processes from observables whose measurement is not demanding in terms of length of samples. This is the main point addressed in sections 4.6 to 4.8, the latter dealing with processes that *a priori* do not fulfill our hypotheses. We conclude in section 4.9 by sketching an alternative approach proposed by [29] that allows overcoming the assumption of detailed balance.

4.1 Entropy rate of Markov chains

Let $X_1, X_2, ..., X_n, ...$ denote a stochastic process, namely an infinite sequence of random variables living in a state space S. For any subsequence can be defined a probability distribution $p(x_1, x_2, ..., x_n)$ denoting the joint probability for X_1 to take value x_1, for X_2 to take value x_2, and so on.[1] Such a process is said *stationary* if $p(x_1, ..., x_n)$ is invariant with respect to shifts in the temporal index, that is

$$p(x_1, ..., x_n) = p(x_{1+\tau}, ..., x_{n+\tau}) \tag{4.1}$$

$\forall \tau \in \mathbb{Z}$. Once a probability measure is specified over finite subsequences, the Shannon entropy of such sequences is calculated from (2.5). To understand how the entropy of sequences grows with the length of sequences, we have however to consider the limit case of an infinite subsequence by defining the *entropy rate* of the process by

$$h(\mathbf{X}) = \lim_{n\to\infty} \frac{H(X_1, ..., X_n)}{n}. \tag{4.2}$$

This expression is not very convenient for computational purposes, but it happens that for stationary processes entropy rate can be re-expressed more conveniently using the chain rule[2]

$$H(X_1, ..., X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, ..., X_1). \tag{4.3}$$

Then we have

$$h(\mathbf{X}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i|X_{i-1}, ..., X_1). \tag{4.4}$$

But reduction of uncertainty through conditioning (equation (2.32)) and stationarity (equation (4.1)) imply that

$$\begin{aligned} H(X_i|X_{i-1}, ..., X_1) &\leq H(X_i|X_{i-1}, ..., X_2) \\ &= H(X_{i-1}|X_{i-2}, ..., X_1), \end{aligned} \tag{4.5}$$

so that the terms of the sum are actually decreasing. It follows that the entropy rate of stationary processes assumes the form

$$h(\mathbf{X}) = \lim_{n\to\infty} H(X_i|X_{i-1}, ..., X_1). \tag{4.6}$$

[1] For an introduction to the subject – and much more – we refer to [85, 92, 30], among others.

[2] The chain rule follows immediately from introducing in the definition of Shannon entropy the decomposition $p(x_1, ..., x_n) = p(x_n|x_1, ..., x_{n-1})p(x_{n-1}|x_{n-2}, ..., x_1)...p(x_2|x_1)p(x_1)$.

We can actually get rid of the limit in the special case of a *Markov process* [32]. The key feature of Markov stochastic processes are that they are memoryless in the sense that the probability of transitioning from one state to another does not depend on the previous states visited by the process, namely

$$p(x_i|x_{i-1}, ..., x_1) = p(x_i|x_{i-1}). \tag{4.7}$$

the same holds true for entropy, so that we can re-express (4.6) as

$$h(\mathbf{X}) = \lim_{n\to\infty} H(X_i|X_{i-1}) = H(X_2|X_1). \tag{4.8}$$

The last equality results from assuming *time invariance*, that is

$$p(x_{i+\tau}|x_{i-1+\tau}) = p(x_i|x_{i-1}) \tag{4.9}$$

$\forall\tau \in \mathbb{Z}$. Equation (4.8) can now be expressed in terms of probabilities as

$$\begin{aligned} h(\mathbf{X}) &= H(X_2|X_1) \\ &= \sum_{x_1} p(x_1) H(X_2|x_1) \\ &= -\sum_{x_1,x_2} p(x_1)p(x_2|x_1) \ln p(x_2|x_1). \end{aligned} \tag{4.10}$$

Time invariance is a stronger requirement than stationarity; its importance for Markov processes lies in the fact that a time invariant Markov process over a finite-dimensional state space can be expressed as an $N \times N$ matrix $\mathbf{w}$ whose elements are defined as transition probabilities, such that[3]

$$w(x, y) = p(y|x). \tag{4.11}$$

We shall also reserve the notation $\boldsymbol{\pi}$ for the stationary distribution such that $\pi(y) = \sum_x \pi(x)w(x, y)$. With these notations, the entropy rate takes its final form

$$h(\mathbf{X}) = -\sum_{x,y} \pi(x)w(x, y) \ln w(x, y). \tag{4.12}$$

[3]Writing w_{xy} could be more appropriate, but we shall stick instead to a notation allowing an easy switch to continuous states if necessary.

4.2 Maximizing the entropy rate

Entropy rate (4.12) is therefore the quantity we need to maximize in order to create a dynamics producing minimally informative sequences. Compared to the derivation leading to (3.2), the present situation is much more involved since the stationary distribution $\boldsymbol{\pi}$ itself depends on the transition matrix $\mathbf{w}$ in a non-trivial way, for the former has to be the left eigenvector associated with the unitary eigenvalue of the latter. The minimization procedure exposed in chapter 2 therefore cannot be applied as such. A way out, suggested by [100], is to maximize instead

$$\eta = -\sum_{x,y} p(x)w(x,y)\ln w(x,y), \tag{4.13}$$

where $\boldsymbol{p}$ is *any* distribution over S. The maximization is thus carried through both for $\boldsymbol{p}$ and $\mathbf{w}$, specifying separately that $\boldsymbol{p}$ has to be stationary. Stationarity can be guaranteed *a posteriori* by enforcing the *condition of detailed balance*

$$p(x)w(x,y) = p(y)w(y,x). \tag{4.14}$$

Though stationarity can be recovered by summing (4.14) over y, assuming detailed balance is a much stronger requirement; while stationarity can be visualized as the fact that the in- and outcoming fluxes of probability through a state have to be equal, detailed balance implies that the flux received by x from y is equal to the flux sent by x to y.[4] Among the far-reaching implications of this assumption is the fact that Markov processes satisfying detailed balance are reversible. Though there may be physical arguments supporting such an assumption, this is not the case when dealing with more diversified situations; we shall therefore content ourselves for the moment by considering detailed balance as a simplifying mathematical device.

This assumption provides the side benefit that it yields a convenient expression for the stationary probabilities in terms of the transition probabilities; indeed

$$\begin{aligned} p(x) &= 1 - \sum_{y\neq x} p(y) \\ &= 1 - \sum_{y\neq x} \frac{p(x)w(x,y)}{w(y,x)} \end{aligned} \tag{4.15}$$

so that

$$p(x) = \left(1 + \sum_{y\neq x} \frac{w(x,y)}{w(y,x)}\right)^{-1}. \tag{4.16}$$

[4]Note that 2-state processes at equilibrium necessarily satisfy detailed balance.

The maximum entropy problem for Markov processes can therefore be formulated by finding the couple $(\boldsymbol{p}, \mathbf{w})$ which maximizes η while satisfying

$$\sum_x p(x) = 1 \tag{4.17}$$

$$\sum_y w(x,y) = 1 \tag{4.18}$$

$$p(x)w(x,y) = p(y)w(y,x) \tag{4.19}$$

$$C(\boldsymbol{p}, \mathbf{w}) = 0. \tag{4.20}$$

The first three constraints are structural constraints ensuring that $\boldsymbol{p}$ is a well-defined distribution, that $\mathbf{w}$ is row-stochastic, and that detailed balance (hence stationarity) is satisfied. The term $C(\boldsymbol{p}, \mathbf{w})$ expresses the observational constraints imposed to the model (see (4.33) for an example) and will be our principal object of interest in this chapter.

4.3 General solution

We now derive the general solution of the problem defined by (4.17), (4.18), (4.19) and (4.20). Introducing Lagrange multipliers Λ, $\lambda(u)$ and $\theta(u,v)$, the function to maximize is

$$\begin{aligned} L = &- \sum_{u,v} p(u)w(u,v) \ln w(u,v) + \Lambda \left(\sum_u p(u) - 1 \right) + \sum_u \lambda(u) \left(\sum_v w(u,v) - 1 \right) \\ &+ \sum_{u,v<u} \theta(u,v) \left(p(u)w(u,v) - p(v)w(v,u) \right) + C(\boldsymbol{p}, \mathbf{w}). \end{aligned} \tag{4.21}$$

Deriving L with respect to $\boldsymbol{p}$ and $\mathbf{w}$, we get

$$0 = \frac{\partial L}{\partial w(x,x)} = -p(x) \ln w(x,x) - p(x) + \lambda(x) + \frac{\partial C}{\partial w(x,x)} \tag{4.22}$$

$$0 = \frac{\partial L}{\partial w(x, y<x)} = -p(x) \ln w(x,y) - p(x) + \lambda(x) + \theta(x,y)p(x) + \frac{\partial C}{\partial w(x, y<x)} \tag{4.23}$$

$$0 = \frac{\partial L}{\partial w(x, y>x)} = -p(x) \ln w(x,y) - p(x) + \lambda(x) - \theta(y,x)p(x) + \frac{\partial C}{\partial w(x, y>x)} \tag{4.24}$$

$$0 = \frac{\partial L}{\partial p(x)} = -\sum_v w(x,v)\ln w(x,v) + \Lambda + \sum_{v<x}\theta(x,v)w(x,v) - \sum_{u>x}\theta(u,x)w(x,u) + \frac{\partial C}{\partial p(x)}. \tag{4.25}$$

The strategy now is to substitute the expression for $\lambda(x)$ provided by (4.22) into (4.23) and (4.24); so doing these equations become

$$\ln\frac{w(x,x)}{w(x,y)} + \theta(x,y) + \frac{1}{p(x)}\left(\frac{\partial C}{\partial w(x,y<x)} - \frac{\partial C}{\partial w(x,x)}\right) = 0 \tag{4.26}$$

$$\ln\frac{w(x,x)}{w(x,y)} - \theta(y,x) + \frac{1}{p(x)}\left(\frac{\partial C}{\partial w(x,y>x)} - \frac{\partial C}{\partial w(x,x)}\right) = 0. \tag{4.27}$$

Swapping arguments $x \leftrightarrow y$ in the last equation and adding the modified equation to the first yields

$$\begin{aligned}\ln\frac{w(x,x)w(y,y)}{w(x,y)w(y,x)} &+ \frac{1}{p(x)}\left(\frac{\partial C}{\partial w(x,y<x)} - \frac{\partial C}{\partial w(x,x)}\right)\\ &+ \frac{1}{p(y)}\left(\frac{\partial C}{\partial w(y,x>y)} - \frac{\partial C}{\partial w(y,y)}\right) = 0.\end{aligned} \tag{4.28}$$

As to (4.25), let us split the equation into diagonal, upper and lower triangular terms as

$$\begin{aligned}0 = &-w(x,x)\ln w(x,x) + \Lambda + \frac{\partial C}{\partial p(x)}\\ &+ \sum_{v<x} w(x,v)\left(\theta(x,v) - \ln w(x,v)\right)\\ &+ \sum_{v>x} w(x,v)\left(-\theta(v,x) - \ln w(x,v)\right).\end{aligned} \tag{4.29}$$

Eliminating θ by means of equations (4.26) and (4.27), we get

$$\begin{aligned}0 = &-\ln w(x,x) + \Lambda + \frac{\partial C}{\partial p(x)} + \frac{1-w(x,x)}{p(x)}\frac{\partial C}{\partial w(x,x)}\\ &- \frac{1}{p(x)}\sum_{v<x} w(x,v)\frac{\partial C}{\partial w(x,v<x)}\\ &- \frac{1}{p(x)}\sum_{v>x} w(x,v)\frac{\partial C}{\partial w(x,v>x)}.\end{aligned} \tag{4.30}$$

Different observable quantities can be chosen as constraints in (4.20). Even before considering their physical relevance, two crucial criteria that should guide our choice are, first, that these quantities should be easy to measure; and second, that the resulting constraints should be reasonably easy to implement theoretically or numerically. The philosophy behind the maximum entropy approach consisting in bypassing brute-force estimation of a probability distribution by making a guess based on simple observables, this approach would be spoilt altogether if the retained observables turned out to be awfully difficult to measure, hence the first point. As to the second, there is no need for comment here.

Keeping these guidelines in mind, we consider here constraints on the average square value of a symbol and the average product of two consecutive symbols. Denoting respectively by σ^2 and A the measured values of these quantites[5], we constrain

$$\sum_x x^2 p(x) = \sigma^2 \tag{4.31}$$

and

$$\sum_{x,y} xyp(x,t;y,t+1) = \sum_{x,y} xyw(x,y)p(x) = A. \tag{4.32}$$

In other words we have

$$C(\boldsymbol{p},\mathbf{w}) = \alpha\left(\sum_u u^2 p(u) - \sigma^2\right) + \beta\left(\sum_{u,v} uvw(u,v)p(u) - A\right). \tag{4.33}$$

It follows that

$$\frac{\partial C}{\partial p(x)} = \alpha x^2 + \beta \sum_v xvw(x,v)$$

$$\frac{\partial C}{\partial w(x,x)} = \beta x^2 p(x)$$

$$\frac{\partial C}{\partial w(x,y)} = \beta xyp(x).$$

We can now rewrite our general equations as

$$\ln \frac{w(x,x)w(y,y)}{w(x,y)w(y,x)} - \beta\,(x-y)^2 = 0 \tag{4.34}$$

[5]For convenience we shall refer to (4.31) and (4.32) as "variance" and "1-step autocorrelation" since the shortcut is not usual in the field of signal processing.

$$-\ln w(x,x) + (\alpha+\beta)\,x^2 + \Lambda = 0 \qquad \rightarrow \qquad \ln \frac{w(x,x)}{w(y,y)} = (\alpha+\beta)\left(x^2 - y^2\right). \quad (4.35)$$

Equations (4.34) and (4.35) therefore provide relations among the transition probabilities of the Markov chain which maximizes the entropy rate. When no constraint beyond (4.17), (4.18) and (4.19) is introduced, we have $\alpha = \beta = 0$; it is then easily verified that $w(x,y) = \exp\Lambda \ \forall x, y$ satisfies (4.34), (4.35), with $\Lambda = -\ln|S|$.

4.4 Analytically solvable 2-state processes

Although the system (4.34), (4.35) has in general to be solved numerically, the case of a system of two states $\{-1,+1\}$ provides an excellent benchmark for analytical work. Indeed we get

$$\ln \frac{w(-1,-1)w(+1,+1)}{w(-1,+1)w(+1,-1)} = 4\beta \quad (4.36)$$

$$\ln \frac{w(-1,-1)}{w(+1,+1)} = 0. \quad (4.37)$$

The multiplier α associated to the constraint on the variance of the process disappears since in this particular case $\sigma^2 = 1$ whatever the transition matrix $\mathbf{w}$. Using (4.37) and row-stochasticity, (4.36) can be solved to give

$$w(-1,-1) = w(+1,+1) = 1 - w(-1,+1) = 1 - w(+1,-1) = \frac{1}{1+\exp(-2\beta)}. \quad (4.38)$$

The multiplier β has finally to be related to the observed 1-step autocorrelation A. An easy calculation yields $\beta = \frac{1}{2}\ln\left(\frac{1+A}{1-A}\right)$, hence the transition matrix generating the process having the largest entropy rate can be written in the 2-state case as

$$\mathbf{w} = \frac{1}{2}\begin{pmatrix} 1+A & 1-A \\ 1-A & 1+A \end{pmatrix}. \quad (4.39)$$

This simple result illustrates the squeezing of the space of independant transition coefficients (here 2-dimensional) onto a submanifold (here 1-dimensional) due to the imposition of a constraint. As we shall see later, enforcing multiple constraints would result in a submanifold of larger dimensionality.

Remark: In [26], [25] we characterized this situation as a system of two states *encoded* as $\{-1,+1\}$, but this formulation is deeply misleading. Consider a process generating

a sequence of abstract symbols (alphabetical for instance), and endow these states with numerical values so as to be able to compute their autocorrelation (and variants such as (4.32)); it is obvious from (4.36), (4.37), and more generally (4.34), (4.35), that the transition probabilities of the maximum entropy rate Markov chain will depend on the 'encoding' of the states, namely the numerical values with which the abstract states are put in correspondence. Said otherwise, our maximum entropy rate transition matrix will characterize not only the process considered abstractly, but also the encoding (of course as long as this encoding of the state space is physically motivated the issue vanishes - this might actually be true even up to a rescaling).

Though this idea sounds disturbing at first, it could not be otherwise since the autocorrelation itself depends on the values of the variables, and autocorrelation is the only thing we are allowed to observe. This might actually have bothered Van der Straeten since in [100] he considers observables that are postulated to be combinations of *transition records*; these records are in our opinion so tightly related to the matrix elements we are seeking that the calculations presented there are likely to contain a good deal of circularity.

4.5 Larger state spaces

For larger state spaces equations (4.34) and (4.35) have to be solved numerically. Let us work out in detail the case of a 3-state process with states $\{-1, 0, +1\}$. Then (4.35) becomes

$$\frac{w(-,-)}{w(0,0)} = e^{\alpha+\beta} \tag{4.40}$$

$$\frac{w(0,0)}{w(+,+)} = e^{-\alpha-\beta} \tag{4.41}$$

$$\frac{w(-,-)}{w(+,+)} = 1, \tag{4.42}$$

while (4.34) gives

$$\frac{w(-,-)w(0,0)}{w(-,0)w(0,-)} = e^{\beta} \tag{4.43}$$

$$\frac{w(-,-)w(+,+)}{w(-,+)w(+,-)} = e^{4\beta} \tag{4.44}$$

$$\frac{w(0,0)w(+,+)}{w(0,+)w(+,0)} = e^{\beta}. \tag{4.45}$$

The row normalization provides three equations as well:

$$w(-,-) + w(-,0) + w(-,+) = 1 \tag{4.46}$$

$$w(0,-) + w(0,0) + w(0,+) = 1 \tag{4.47}$$

$$w(+,-) + w(+,0) + w(+,+) = 1. \tag{4.48}$$

Writing $w(-,-) = k$ for convenience, we can fill the diagonal of the maximum entropy rate transition matrix $\mathbf{w}_{ME}$, so that

$$\mathbf{w}_{ME} = \begin{pmatrix} k & * & * \\ * & \frac{k}{e^{\alpha+\beta}} & * \\ * & * & k \end{pmatrix}. \tag{4.49}$$

Putting $w(-,0) = m$ and using the condition (4.43) relating $w(0,-)$ and $w(-,0)$, we have

$$\mathbf{w}_{ME} = \begin{pmatrix} k & m & * \\ \frac{k^2}{e^{\alpha+2\beta}m} & \frac{k}{e^{\alpha+\beta}} & * \\ * & * & k \end{pmatrix}. \tag{4.50}$$

Then we deal with the couple $w(-,+), w(+,-)$, assigning $w(-,+) = 1-k-m$ in order to compel with the normalization condition on the first row. Using (4.44) yields

$$\mathbf{w}_{ME} = \begin{pmatrix} k & m & (1-k-m) \\ \frac{k^2}{e^{\alpha+2\beta}m} & \frac{k}{e^{\alpha+\beta}} & * \\ \frac{k^2}{e^{4\beta}(1-k-m)} & * & k \end{pmatrix}. \tag{4.51}$$

Carrying through the same procedure for the couple $w(0,+), w(+,0)$ by means of (4.45), and using the normalization condition on the second line gives

$$\mathbf{w}_{ME} = \begin{pmatrix} k & m & (1-k-m) \\ \frac{k^2}{e^{\alpha+2\beta}m} & \frac{k}{e^{\alpha+\beta}} & \left(1 - \frac{k^2}{e^{\alpha+2\beta}m} - \frac{k}{e^{\alpha+\beta}}\right) \\ \frac{k^2}{e^{4\beta}(1-k-m)} & \frac{k^2 m}{e^{\alpha+2\beta}m - k^2 - e^{\beta}km} & k \end{pmatrix}. \tag{4.52}$$

Enforcing the normalization condition on the third line eventually gives that

$$m = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{4.53}$$

with

$$a := (e^{5\beta} - e^{4\beta})k^2 - (e^{6\beta+\alpha} + e^{5\beta})k + e^{6\beta+\alpha} \tag{4.54}$$

$$b := (e^{5\beta} - e^{\beta})k^3 - (e^{6\beta+\alpha} + 2e^{5\beta} - e^{2\beta+\alpha})k^2 + (2e^{6\beta+\alpha} + e^{5\beta})k - e^{6\beta+\alpha} \tag{4.55}$$

$$c := (e^{4\beta} - 1)k^4 - 2e^{4\beta}k^3 + e^{4\beta}k^2. \tag{4.56}$$

We have not yet at this stage used the normalization over p. Using (4.16), it is now necessary, for given values of the multipliers α, β, to scan possible value of k in order to find the one such that the probabilistic normalization is enforced. Once this is done, we have to proceed similarly for adjusting the values of α and β so that the constraints are satisfied. A possible implementation is sketched in the pseudo-code 1 below.

4.6 Accuracy of reconstruction

Our task now is to estimate the quality of the maximum entropy rate reconstruction. This can be done analytically for the 2-state process with states $\{\pm 1\}$. Letting A denote the autocorrelation of the process, we have seen that for such processes (4.34) and (4.35) could be solved to give the maximum entropy transition matrix

$$\mathbf{w}_{ME} = \begin{pmatrix} \frac{1+A}{2} & \frac{1-A}{2} \\ \frac{1-A}{2} & \frac{1+A}{2} \end{pmatrix}, \tag{4.57}$$

with A the measured autocorrelation. Our purpose is to investigate if there exists a subset of the space of 2×2 stochastic matrices for which the maximum entropy method is more efficient than sampling in estimating $\mathbf{w}$ when we only have short samples at our disposal. We detail the calculations for the coefficient $w(-,-)$, the other three being similar.

Since the sample autocorrelation of a well-behaved process obeys a central limit theorem [17], we can make the assumption that the sample autocorrelation $A^{(n)}$ measured from a sample of length n is distributed normally according to $\mathcal{N}(A, n^{-1})$. According to (4.57), it follows that the error made on the estimation of $w(-,-)$ using the MaxEnt method is distributed as $\mathcal{N}(\frac{1+A}{2} - w(-,-), (4n)^{-1})$. The absolute value of this error thus obeys a folded normal distribution, which has mean and standard deviation given by (see [63])

$$\langle |\Delta_{ME} w(-,-)| \rangle^{(n)} = \frac{e^{-2n\mu_{--}^2}}{\sqrt{2\pi n}} + \mu_{--} \left(1 - 2\Phi\left(-2\sqrt{n}\mu_{--}\right)\right) \tag{4.58}$$

$$\sigma^{(n)}(|\Delta_{ME} w(-,-)|) = \sqrt{\mu_{--}^2 + \frac{1}{4n} - \left(\langle |\Delta_{ME} w(-,-)| \rangle^{(n)}\right)^2}, \tag{4.59}$$

Algorithm 1 Estimation of the 3-state transition matrix $\mathbf{w}_{ME}$

Specify empirical variance σ^2 and autocorrelation A
// Set arbitrary initial values to Lagrange multipliers
$\alpha = 1$, $\beta = 1$
// Loop initialization
$\epsilon_\alpha = 1$, $\epsilon_\beta = 1$, $n = 0$
// Loop on multipliers α, β. The loop ends when two successive iterates of both multipliers do not differ by more than $\epsilon = 10^{-5}$.
while $\epsilon_\alpha > 10^{-5}$ AND $\epsilon_\beta > 10^{-5}$ AND $n < 1000$ **do**
  // Loop initialization
  $k_0 = 0$, $dk = 0.1$, $s_{max} = 0$
  // Loop on k, minimizing $|1 - \sum p|$. The domain of k is explored by successive refinements of dk. An accuracy of $dk = 10^{-4}$ is targeted here.
  while $dk > 10^{-5}$ **do**
    for $k = \max(k_0 - 10dk, 0)$; $k < k_0 + 10dk$; $k+ = dk$ **do**
      compute a from (4.54), b from (4.55), c from (4.56)
      compute m from (4.53) *// $\pm$ to select $0 \leq m \leq 1$*
      compute $\mathbf{w}_{ME}$ from (4.52)
      compute p from (4.16)
      compute $s = |1 - \sum p|$
      if $s < s_{max}$ AND $0 \leq$ elements of $\mathbf{w}_{ME} \leq 1$ **then**
        $s_{max} = s$, $k_1 = k$
      end if
    end for
    // Step refinement
    $k_0 = k_1$, $dk = dk/10$
  end while
  $k = k_0$
  compute a,b,c ; m ; $\mathbf{w}_{ME}$
  compute the variance $\hat{\sigma}^2$ from $\mathbf{w}_{ME}$ with (4.31)
  $\alpha_1 = \alpha + \sigma^2 - \hat{\sigma}^2$
  $\epsilon_\alpha = |\alpha - \alpha_1|$
  // Update α
  $\alpha = \alpha_1$
  compute the autocorrelation $\hat{A}$ from $\mathbf{w}_{ME}$ with (4.32)
  $\beta_1 = \beta + A - \hat{A}$
  $\epsilon_\beta = |\beta - \beta_1|$
  // Update β
  $\beta = \beta_1$
  $n = n + 1$
end while

where $\mu_{--} = \left(\frac{1+A}{2} - w(-,-)\right)$ and Φ denotes the normal cumulative distribution function.

Similarly an estimate of the error commited when estimating $w(-,-)$ by frequency sampling can be provided. It can be shown [16] that the coefficient sampled from a window of size n is distributed normally according to $\mathcal{N}(w(-,-), \frac{w(-,-)(1-w(-,-))}{np(-)})$ where $p(-)$ denotes the stationary probability of being in state -1, which using (4.16) is given by $p(-) = \frac{1-w(+,+)}{2-w(-,-)-w(+,+)}$. Following the same steps as previously, the sampled absolute error on $w(-,-)$ is found to have its mean and deviation given by

$$\langle|\Delta_S w(-,-)|\rangle^{(n)} = \sqrt{\frac{2w(-,-)(1-w(-,-))}{\pi np(-)}} \tag{4.60}$$

$$\sigma^{(n)}(|\Delta_S w(-,-)|) = \sqrt{\left(1-\frac{2}{\pi}\right)\frac{w(-,-)(1-w(-,-))}{np(-)}}. \tag{4.61}$$

The estimates (4.58), (4.59), (4.60), (4.61) can be compared to the errors actually obtained when generating a sequence from a known transition matrix $\mathbf{w}$ and sampling it in order to reconstruct $\mathbf{w}$ using either maximum entropy or histogram sampling. Figure 4.1 shows this for two different processes. Obviously for small sample length the quality of our estimates depends on the transition matrix and is lesser when the autocorrelation is large; when n grows the central limit theorem ensures that our estimates match empirical errors perfectly.

Equipped with estimates (4.58) and (4.60) we are now able to define an *accuracy gain* as

$$\Delta_{--}^{(n)} = \langle|\Delta_S w(-,-)|\rangle^{(n)} - \langle|\Delta_{ME} w(-,-)|\rangle^{(n)}. \tag{4.62}$$

This accuracy gain is positive when, for samples of size n, the maximum entropy reconstruction provides a better estimation of $w(-,-)$ than frequency sampling does. Though $\Delta_{ij}^{(n)}$ $(i,j \in \{\pm 1\})$ depends on the coefficient, one may wish to define a global $\Delta^{(n)}$ for the matrix $\mathbf{w}$ considered. A conservative option is to choose the minimum over all coefficients, but we shall rather tolerate a poor estimation of one of the coefficients as long as the corresponding transitions occur scarcely and therefore define $\Delta_{\mathbf{w}}^{(n)}$ as the sum of all $\Delta_{ij}^{(n)}$'s weighted by the stationary distribution, namely

$$\Delta_{\mathbf{w}}^{(n)} = \sum_i p(i)\Delta_{ij}^{(n)}. \tag{4.63}$$

From our experience, the definition of $\Delta_{\mathbf{w}}^{(n)}$ does not alter qualitatively the forthcoming results (see figure 4.2). We now let $n_c(\mathbf{w})$ denote the value of n above which $\Delta_{\mathbf{w}}^{(n)}$ becomes negative. In other words, a non-negative n_c means that for historical samples

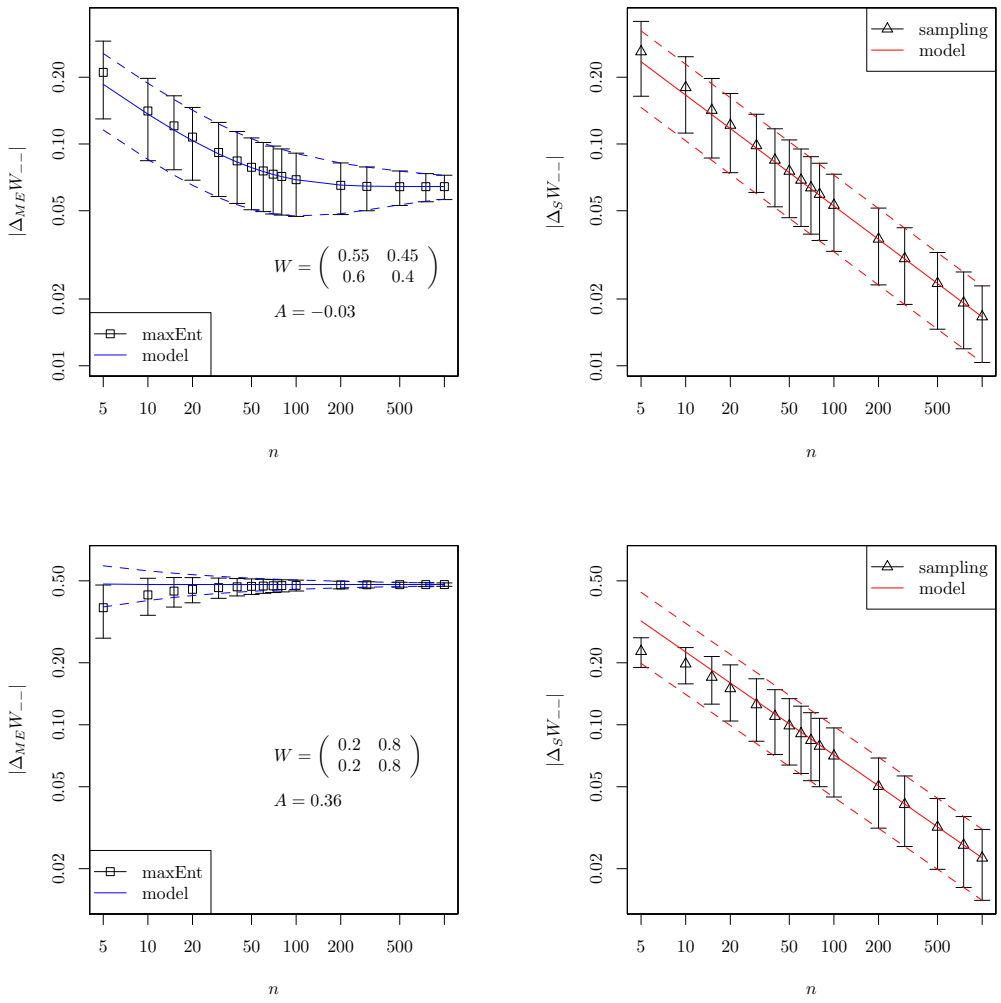
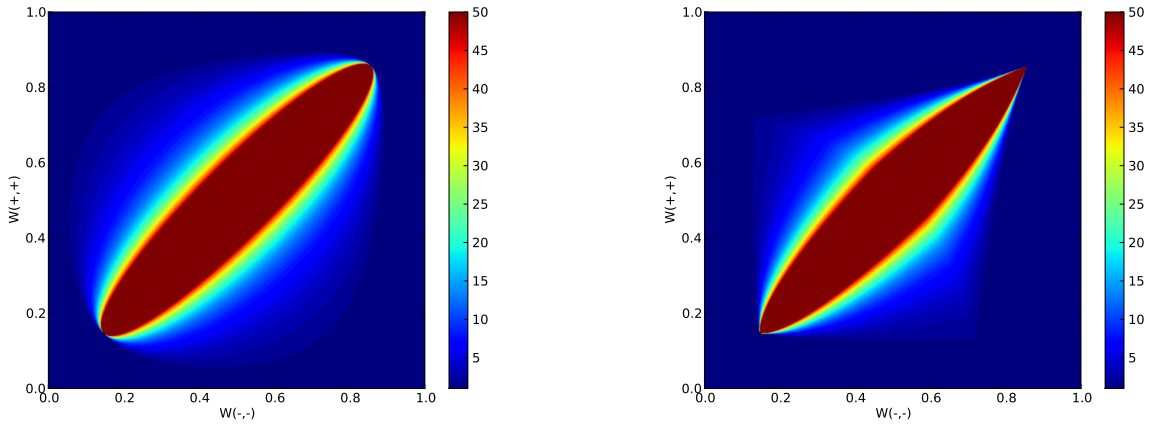


Figure 4.1: Comparison between empirical mean and standard deviation of data (bars) and estimates (4.58), (4.59), (4.60), (4.61) derived from the central limit assumption (continuous lines: mean; dashed lines: standard deviation), for transition matrices with autocorrelation $A = -0.03$ (top) and $A = 0.36$ (bottom).



(a) (b)

Figure 4.2: $n_c(\mathbf{w})$ plotted over the space of 2-state stochastic matrices parametrized by $w(-,-)$, $w(+,+)$, for $\Delta_{\mathbf{w}}^{(n)}$ chosen as (a) the weighted sum of individual coefficients (b) the minimum over coefficients.

shorter than n_c, the MaxEnt method gives better results when estimating the transition matrix underlying the observed process.

The quantity $n_c(\mathbf{w})$ found from (4.62) is plotted in figure 4.2 over the space of 2×2 stochastic matrices parametrized by $(w(-,-), w(+,+))$. Note that n_c is large near the diagonal but decays when one moves away from it, which means that a matrix which is "compatible" with the structure (4.57) is better estimated using the maximum entropy scheme while matrices that do not fit in (4.57) are not adequately grasped by this approach.

4.7 The curse of dimensionality

Denoting $M(n)$ the set of matrices such that $n_c(\mathbf{w}) \geq n$ and $\mu(n)$ the relative size of $M(n)$ compared to $M(0)$ (the space of all 2×2 stochastic matrices), then the relevance of the maximum entropy approach for a given state space will depend critically on the function $\mu(n)$. In the 2-state case, one can read from figure 4.2 that $M(50)$ is concentrated in a neighbourhood around the diagonal so that $\mu(50) \approx 0.15$. In other words for samples of size smaller than $n = 50$ the maximum entropy estimate is better than the frequency sampling estimate for about 15% of all possible processes. One should however note that processes on which one might want to apply the method are unlikely to be scattered randomly over $[0,1]^2$, but will rather be processes having a large entropy rate, that is low predictability. This tends to focus our interest on the central area of $[0,1]^2$ and increase the effective $\mu(n)$.

One might expect the maximum entropy rate estimate to outperform sampling when

the dimensionality of the state space increases since an efficient frequency sampling in high-dimensional spaces requires very long samples. Unfortunately, there is another effect to take into account. We have seen that the maximum entropy criterion squeezes the space of independent coefficients onto a space whose dimensionality is equal to the number of constraints implemented (for instance (4.57) defines a one-dimensional manifold). The net result of these competing effects is therefore to penalize the maximum entropy approach for large state spaces. This can of course be alleviated by considering suplementary constraints, each one increasing the dimensionality of the maximum entropy submanifold, at the expense of an extra computational cost.

To illustrate these points let us consider 3-state processes taken randomly in regions characterized by a given range of entropy rate h. In practice, we dissected the matrix space into five regions C_i defined by the conditions $h_i < h < h_{max}$ where $h_{max} = \ln 3$ corresponds to the maximally entropic process such that $w_{ij} = 1/3\ \forall i, j$ and h_i is specified in figure 4.3; this figure shows the effectiveness of our approach by highlighting that processes having a large entropy rate are more suited to it. On figure 4.3 are also displayed the cases where two constraints are enforced (blue curves), and where the constraint on the variance of the process is relaxed (red curves). We observe that, for short samples, going from one to two constraints results in a loss of performance or at best a marginal gain as estimation errors of constraints tend to accumulate. However, when the sampling window is long enough to allow for an accurate estimation of all constraints, adding constraints results in a clear improvement of the maximum entropy estimate.

4.8 Non-stationary processes

As long as stationary processes only are considered, the maximum entropy estimate is actually essentially of academical interest since nothing there precludes the use of (almost) arbitrarily long samples. Things are very different when the dynamics itself changes over time, for then a quick estimation of dynamical parameters becomes necessary. The crucial point, which follows immediately from our previous results, is that if the coefficients $w_{ij}(t)$ evolve within $M(\tau)$, where τ is the typical time scale on which the parameters of the dynamics change, then the maximum entropy criterion should be able to provide a better on-the-fly estimate of the instantaneous dynamics than sampling does.

The change of perspective should be emphasized: while the maximum entropy rate approach is stationary in essence, we now apply it on non-stationary processes by approximating them locally in time by "effective" Markov processes.

4.8.1 Toy model

Let us illustrate this by reconsidering the 2-state process for which we derived the estimates (4.58)-(4.61), but now in the case when the process is generated from a transition matrix whose coefficients vary sinusoidally over time.

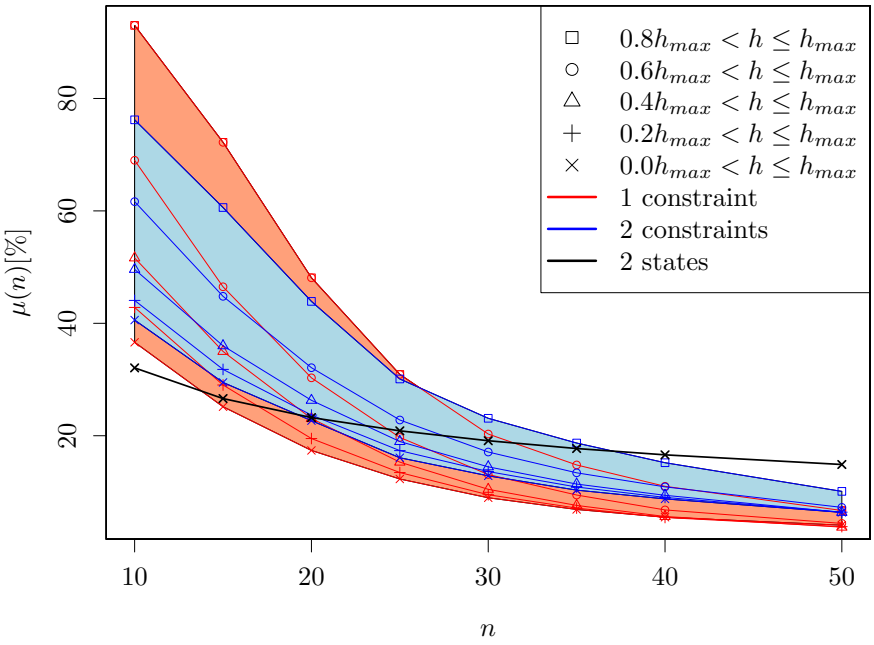


Figure 4.3: Success rate as a function of the sampling window, for 2-state and 3-state processes involving one or two constraints. Cumulated quintiles of the entropy rate are displayed separately for 3-state processes. 1000 processes are taken in each quintile.

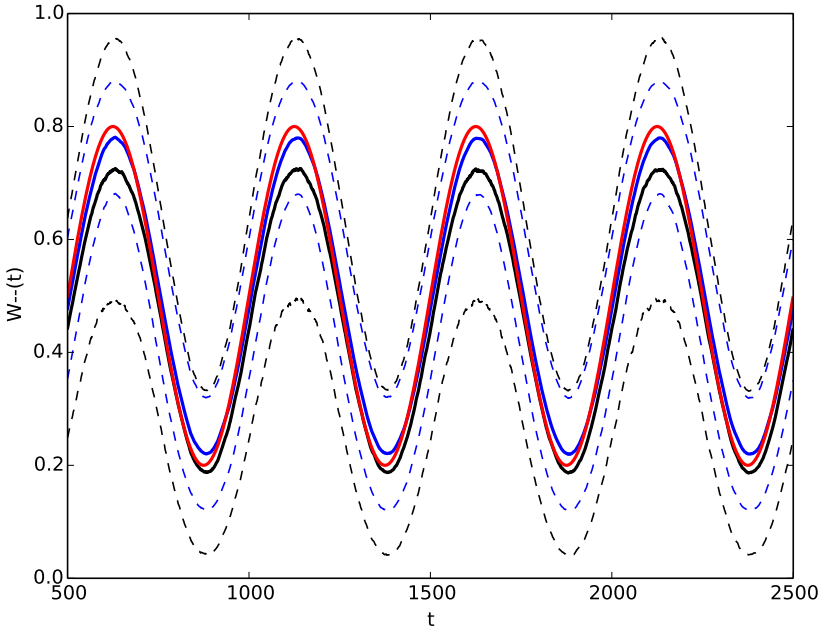


Figure 4.4: The process (4.64) for a sampling window of size $n = 15$. The actual coefficient $w_{--}(t)$ (red) is compared to its maximum entropy (blue) and sampling (black) estimates averaged over 10000 realizations. Standard deviations are shown by dotted channels.

Figures 4.4, 4.5 and 4.6 show the case of a process generated from a transition matrix

$$\mathbf{w}(t) = \begin{pmatrix} 0.5 + 0.3\sin\left(\frac{2\pi t}{T}\right) & 0.5 - 0.3\sin\left(\frac{2\pi t}{T}\right) \\ 0.5 - 0.3\sin\left(\frac{2\pi t}{T}\right) & 0.5 + 0.3\sin\left(\frac{2\pi t}{T}\right) \end{pmatrix}, \tag{4.64}$$

where $T = 500$. Since the generating matrix is diagonal with oscillating coefficients, the dynamics is always of the type 4.57 that in the static case is adequately captured by the maximum entropy estimate. Figure 4.4 displays the actual transition coefficient $w_{--}(t)$ given by (4.64) (in red), its maximum entropy estimate computed from a historical window of length $n = 15$ (in blue) and its sampling estimate from the same window (in black). Both estimates are averaged over 10000 realizations, and the corresponding standard deviations are depicted as dotted channels. The average quality of the reconstruction is comparable for both methods, but as expected from our calculations in the static case, the deviation is considerably reduced using the maximum entropy estimate. Figures 4.5 and 4.6 are similar to 4.4, but when the sampling window is lengthened to $n = 40$ and $n = 100$ respectively. Both estimates are still equivalent in terms of averages, with an increasing lag when n gets larger. When $n \simeq 100$, standard deviations become comparable for both approaches.

Figure 4.7 presents the same plot (for $n = 15$) for a case where the generating transition matrix changes as

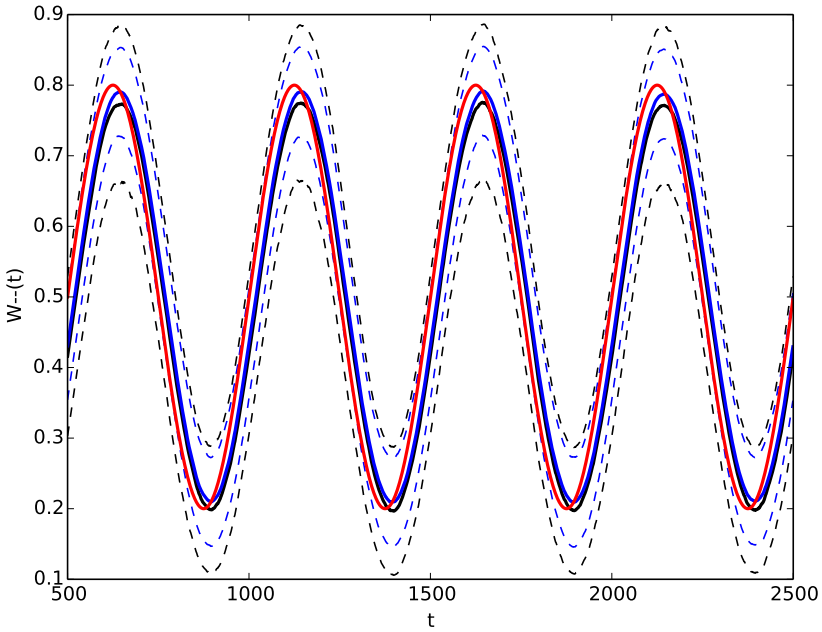


Figure 4.5: The process (4.64) as in figure 4.4, now for $n = 40$.

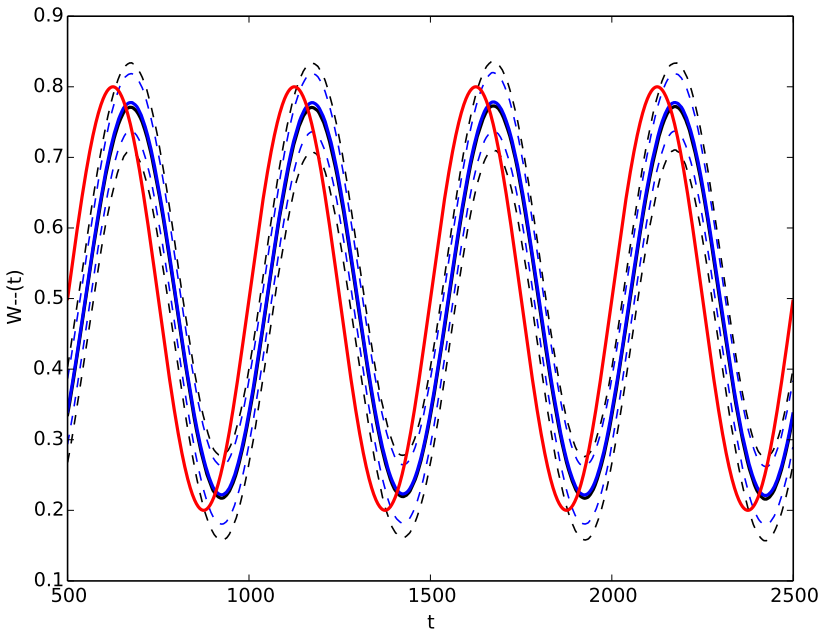


Figure 4.6: Same figure as 4.5, now for $n = 100$.

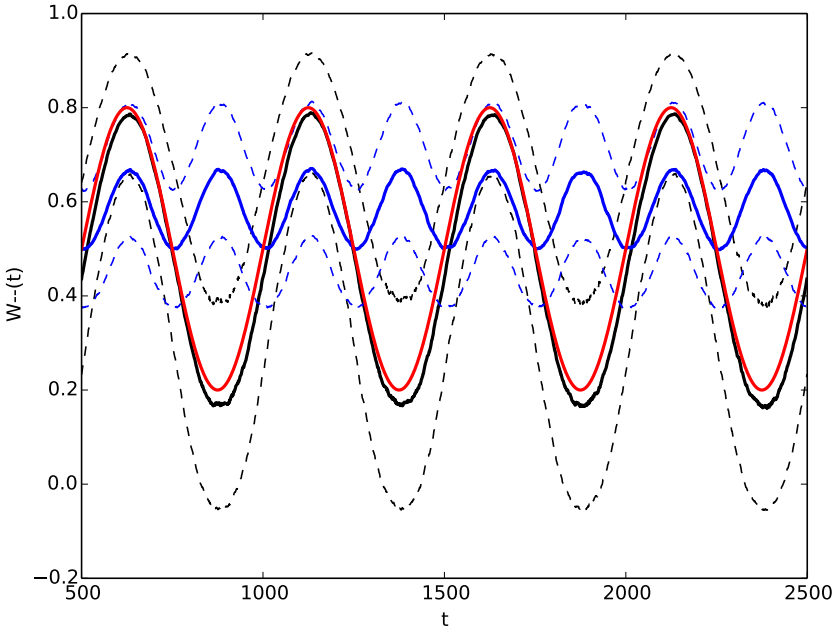


Figure 4.7: The process (4.65) for sampling window of size $n = 15$.

$$\mathbf{w}(t) = \begin{pmatrix} 0.5 + 0.3\sin\left(\frac{2\pi t}{T}\right) & 0.5 - 0.3\sin\left(\frac{2\pi t}{T}\right) \\ 0.5 + 0.3\sin\left(\frac{2\pi t}{T}\right) & 0.5 - 0.3\sin\left(\frac{2\pi t}{T}\right) \end{pmatrix}, \tag{4.65}$$

that is the system oscillates along a diagonal perpendicular to the previous one. Our expectation is that the maximum entropy estimate should miss its target altogether, which is what can be actually observed.

The most interesting case is of course when the dynamics does not evolve exactly on the diagonal where the maximum entropy approximation turns exact, but still evolve in a region of the space of processes where this approximation can perform well. Figure 4.8 illustrates this intermediate case when

$$\mathbf{w}(t) = \begin{pmatrix} 0.5 + 0.3\sin\left(\frac{2\pi t}{T}\right) & 0.5 - 0.3\sin\left(\frac{2\pi t}{T}\right) \\ 0.5 - 0.3\sin\left(\frac{2\pi t}{1.02T}\right) & 0.5 + 0.3\sin\left(\frac{2\pi t}{1.02T}\right) \end{pmatrix}. \tag{4.66}$$

The process is tracked on a time interval $t = [0, 2500]$, during which it covers more or less the oval region apprearing in red in figure 4.2. Again we can note that the quality of the maximum entropy estimate is good, and in particular that the deviation is much smaller than for the sampling estimate.

4.8.2 Empirical time series

Let us finally move one more step further from the original setup in which the maximum entropy rate criterion was formulated, and consider an empirical time series which is

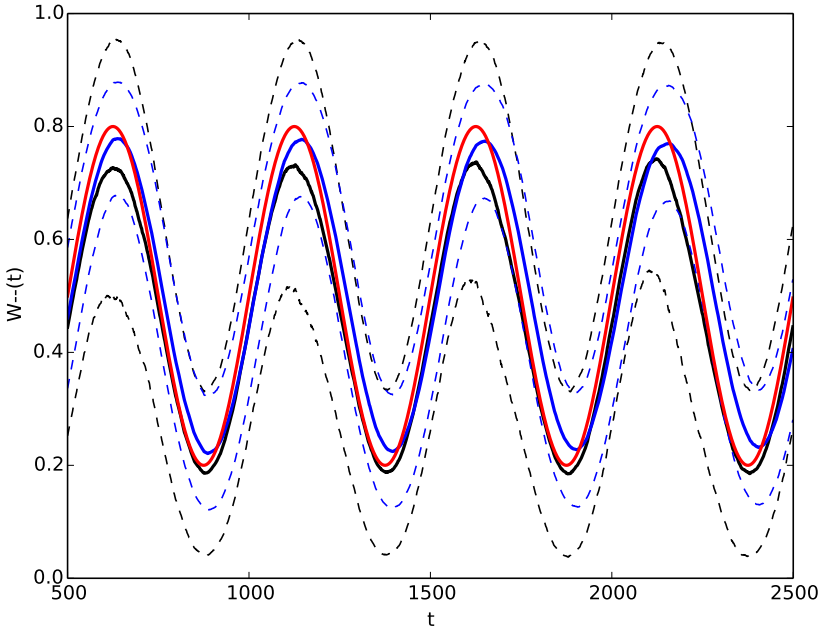


Figure 4.8: The process (4.66) for sampling window of size $n = 15$.

driven by an unknown underlying dynamics that, contrarily to our toy model, is unlikely to change smoothly. The underlying hope is that real processes take place in a region of the process space where it is possible to take advantage of the maximum entropy approach.

Our aim now is to estimate the return tail distributions of a financial asset. For convenience, let us consider the EUR/USD price series over a 3-year period extending from Jan. 1, 2009 up to Jan. 1, 2012. A series of returns is established by defining $r_t = (p_t - p_{t-1})/p_{t-1}$ where p_t is the price at time t and where data are picked up every 15 minutes. A 3-state time series x_t is built by discretizing the returns according to

$$x_t = \begin{cases} -1 & \text{if } r_t < -0.01\% \\ +1 & \text{if } r_t > 0.01\% \\ 0 & \text{otherwise.} \end{cases} \tag{4.67}$$

Modelling the discretized time series with a 3-states Markov chain, we are in a position to estimate the maximum entropy rate transition matrix at each time using the running sampling window.

Our strategy for assessing the relevance of this estimate is to compute the forward distribution of prices and compare it to the prices effectively observed. This is done in two steps:

1. Using the transition matrix $\mathbf{w}^{ME}$ reconstructed from the maximum entropy criterion, we can determine the distribution of prices s steps ahead of current time t,

under the assumption that $\mathbf{w}^{ME}$ still holds over that range. Denoting $q^{ME}_{t+s}(k|i_0)$ the probability to find the process in k at time $t+s$, we have

$$q^{ME}_{t+s}(k|i_0) = \sum_{i_1,\cdots,i_s} w^{ME}_{i_0,i_1} \cdots w^{ME}_{i_{s-1},i_s} \tag{4.68}$$

with $k = i_1 + \cdots + i_s$ and where $i_\tau \in \{-1, 0, +1\}$ is the state at time $t+\tau$.

2. In order to assess the quality of the tails of the distribution q^{ME}, we define its first and last ten symmetrized[6] percentiles π^{ME}_k, $k = 1, ..., 10$. Coming back to the real process, we measure the fraction of times ϕ_k it actually falls in π^{ME}_k. The error for percentile k is then given by

$$\Delta_k = \frac{|0.02 - \phi_k|}{0.02} \tag{4.69}$$

(remind that percentiles are symmetrized) and we can define an average error $\Delta = \langle \sum_k \Delta_k \rangle$ for the tails of the distribution.

Figure 4.9 shows the average error Δ as a function of the sample length n, for the maximum entropy (squares) as well as the sampling (circles) estimate. According to our expectations, we observe that maximum entropy outperforms sampling for sample length shorter than $n = 40$ (*i.e.* 10 hours). Interestingly, it can be noted that for longer samples, sampling errors seem to grow and depart from maximum entropy errors, which we believe to be an over-average effect as discussed above. The non-informed guess, assuming all transitions equiprobable, is also displayed on fig. 4.9 (triangles), and is always outperformed by the maximum entropy estimate.

It is quite remarkable that the maximum entropy method turns out to be relevant in this context, since *a priori* none of the assumptions is satisfied. Referring back to the discussion above, it is likely that processes driving financial data such as the one considered here show very low predictability, hence high entropy, which makes them well suited to the estimation scheme considered in this chapter.

4.9 Reconsidering the problem: an algebraic approach

Arguably the weakest hypothesis in our analysis lies in our assumption that the Markov process to reconstruct obeys the detailed balance condition (4.14), which amounts to assume that the dynamics is invariant under time reversal [101, 80, 99]. An alternative approach based on the Perron-Frobenius theorem [44] has been proposed recently [102, 29] that bypasses this restrictive condition. The idea of this method is to build from the observables a transfer matrix $\mathcal{L}$, which falls into the realm of the Perron-Frobenius theorem. This ensures the existence of a dominant eigenvalue λ larger in modulus than

[6] That is we aggregate the first and the last percentile, and so on.

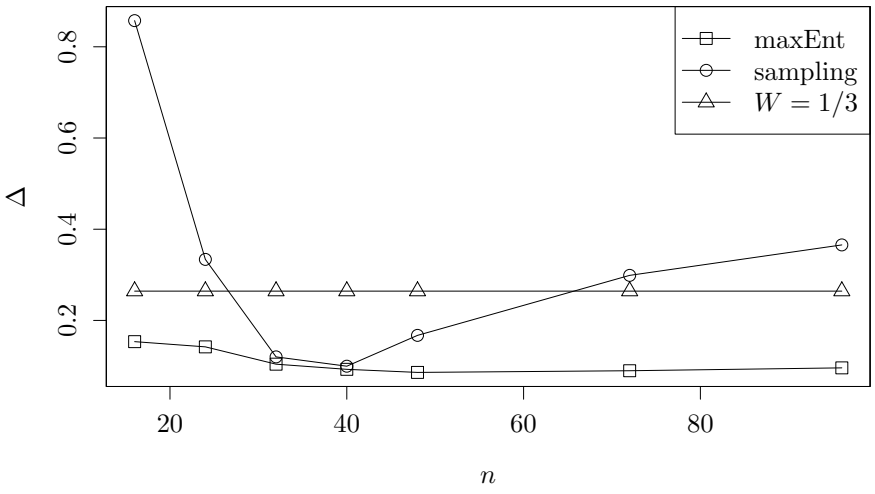


Figure 4.9: The average error of tail distribution estimations, as a function of the sample size. Maximum entropy (squares) and sampling (circles) estimates are used, as well as the naive guess assuming all equiprobable transitions, $w_{ij} = 1/3 \ \forall i, j$ (triangles).

any other eigenvalue, and to which are associated left- and right- eigenvectors L_λ and R_λ. The key property on which this method relies is that the maximum entropy transition matrix and associated stationary density can be recovered directly from λ, L_λ, R_λ and $\mathcal{L}$ [24]. As a complement to the method exposed above, our purpose in this section is to put that approach at work on a tractable case and check its ability to deal with non-reversible dynamics.

The method was first devised in the context of inference of spiking patterns in neural networks[7], so that we shall stick to some terminology in usage there. A neuron is typically modelled by a binary variable $\sigma_i \in \{0, 1\}$. A network of N such neurons at discrete time t is then represented by a vector $\boldsymbol{\sigma}^t = (\sigma_1^t, ..., \sigma_N^t)$, and a spiking pattern of length T by an array $(\boldsymbol{\sigma}^0, \boldsymbol{\sigma}^1, ..., \boldsymbol{\sigma}^T)$. Observables are now defined as products of neurons possibly distant in space but considered at same time or successive times; possible observables are for instance:

$$O(\boldsymbol{\sigma}^t, \boldsymbol{\sigma}^{t+1}) = \sigma_i^t \sigma_j^{t+1} \qquad O(\boldsymbol{\sigma}^t) = \sigma_i^t \qquad O(\boldsymbol{\sigma}^t) = \sigma_i^t \sigma_j^t. \tag{4.70}$$

Let us illustrate this by considering the case of two neurons, namely a system $\boldsymbol{\sigma}$ with $2^2 = 4$ states whose states are

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{4.71}$$

[7]We mean here real neural networks, in contradistinction with highly idealized networks used in deep learning, for instance.

We consider the situation in which we retain as observables the one-timestep-correlation in both directions (from σ_1 to σ_2 and vice-versa), that is

$$O_1(\boldsymbol{\sigma}^t, \boldsymbol{\sigma}^{t+1}) = h_1 \sigma_1^t \sigma_2^{t+1} \qquad O_2(\boldsymbol{\sigma}^t, \boldsymbol{\sigma}^{t+1}) = h_2 \sigma_2^t \sigma_1^{t+1}. \tag{4.72}$$

The coefficients in front of each observable play a role entirely similar to Langrange multipliers, and altough they are not part of the observable properly said it is more convenient for what follows to include the multiplier in the definition, as above.

Evaluating (4.72) on our states is straightforward; we get for instance

$$O_1\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = 0 \qquad O_1\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = h_1 \tag{4.73}$$

$$O_2\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = 0 \qquad O_2\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) = h_2, \tag{4.74}$$

the other twelve elements being evaluated similarly. For each observable these elements can be arranged in a 4×4 matrix $\mathbf{O}_i$, and adding $\mathbf{O}_1$ and $\mathbf{O}_2$ we get the resulting matrix of observables, in our case

$$\mathbf{O} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & h_1 & h_1 \\ 0 & h_2 & 0 & h_2 \\ 0 & h_2 & h_1 & h_1 + h_2 \end{pmatrix}. \tag{4.75}$$

The next step is to introduce a transfer matrix $\mathcal{L}$ defined as the element-wise exponential of $\mathbf{O}$, that is $\mathcal{L}_{ij} = \exp(\mathbf{O}_{ij})$. We get

$$\mathcal{L} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & e^{h_1} & e^{h_1} \\ 1 & e^{h_2} & 1 & e^{h_2} \\ 1 & e^{h_2} & e^{h_1} & e^{h_1+h_2} \end{pmatrix}. \tag{4.76}$$

The key point of this method is to note that the transfer matrix satisfies the hypotheses of the Perron-Frobenius theorem [44], and therefore has a unique dominant eigenvalue λ_{max}. This dominant eigenvalue has to be related to the empirical average values of the observables through

$$\frac{\partial \ln \lambda_{max}}{\partial h_1} = \langle \sigma_1^t \sigma_2^{t+1} \rangle \qquad \frac{\partial \ln \lambda_{max}}{\partial h_2} = \langle \sigma_2^t \sigma_1^{t+1} \rangle. \tag{4.77}$$

To this eigenvalue are associated a left eigenvector $\mathbf{e}_L$ and right eigenvector $\mathbf{e}_R$, which are both non-negative. Then we can build from λ_{max}, $\mathcal{L}$ and $\mathbf{e}_R$ a matrix

$$\mathbf{W} = \frac{1}{\lambda_{max}} \cdot diag(\mathbf{e}_R)^{-1} \cdot \mathcal{L} \cdot diag(\mathbf{e}_R) \tag{4.78}$$

where $diag(\mathbf{e}_R)$ is the diagonal matrix built from the elements of $\mathbf{e}_R$. This transition matrix $\mathbf{W}$ happens to be precisely the matrix maximizing the entropy rate under constrained observed average values of $\sigma_1^t \sigma_2^{t+1}$ and $\sigma_2^t \sigma_1^{t+1}$. The method also provides the stationary distribution, deduced from the eigenvectors as

$$\boldsymbol{\pi}_k = \frac{(\mathbf{e}_R)_k (\mathbf{e}_L)_k}{\mathbf{e}_R \cdot \mathbf{e}_L}. \tag{4.79}$$

As a practical example, let us consider a random process generated from an arbitrary transition matrix

$$\mathbf{W} = \begin{pmatrix} 0.05 & 0.05 & 0 & 0.9 \\ 0.1 & 0.1 & 0.4 & 0.4 \\ 0.4 & 0.1 & 0.4 & 0.1 \\ 0 & 0.1 & 0.6 & 0.3 \end{pmatrix}. \tag{4.80}$$

Measuring the correlations we find $\langle \sigma_1^t \sigma_2^{t+1} \rangle \simeq 0.3758$ and $\langle \sigma_2^t \sigma_1^{t+1} \rangle \simeq 0.2135$. Carrying through the procedure above numerically according to the algorithm 2 below we find for the maximum entropy reconstruction

$$\mathbf{W}^{(ME)} = \begin{pmatrix} 0.2046 & 0.4549 & 0.1056 & 0.2349 \\ 0.0920 & 0.2046 & 0.2182 & 0.4852 \\ 0.3961 & 0.2633 & 0.2046 & 0.1360 \\ 0.1782 & 0.1184 & 0.4225 & 0.2809 \end{pmatrix} \tag{4.81}$$

whose stationary distribution is

$$\boldsymbol{\pi}^{(ME)} = (0.2167, 0.2488, 0.2488, 0.2857). \tag{4.82}$$

We can deduce from (4.81), (4.82) that for instance

$$\mathbf{W}^{ME}_{(01)\to(10)} \boldsymbol{\pi}^{(ME)}(01) = 0.2182 \cdot 0.2488 \simeq 0.0543 \tag{4.83}$$

while

$$\mathbf{W}^{ME}_{(10)\to(01)} \boldsymbol{\pi}^{(ME)}(10) = 0.2633 \cdot 0.2488 \simeq 0.0655, \tag{4.84}$$

so that in this approach the detailed balance is indeed not enforced !

Algorithm 2 Estimation of the maximum entropy rate transition matrix for two binary neurons under constrained correlations $\langle \sigma_1^t \sigma_2^{t+1} \rangle$ and $\langle \sigma_2^t \sigma_1^{t+1} \rangle$

Specify empirical correlations $\langle \sigma_1^t \sigma_2^{t+1} \rangle$, $\langle \sigma_2^t \sigma_1^{t+1} \rangle$
// Set perturbation parameter ϵ
$\epsilon = 10^{-5}$
// Initialize stopping criterion
$\delta = 10^{-6}$
// Loop initialization
$k_1^0 = 0$, $k_2^0 = 0$, $dk = 0.1$
// Loop on k*. The domain of* k *is explored by successive refinements of* dk*. An accuracy of* 10^{-5} *is targeted here.*
while $dk > 10^{-6}$ **do**
  for $k_1 = k_1^0 - 10dk$; $k_1 < k_1^0 + 10dk$; $k_1 += dk$ **do**
    for $k_2 = k_2^0 - 10dk$; $k_2 < k_2^0 + 10dk$; $k_2 += dk$ **do**
      Compute unperturbed transfer matrix $\mathcal{L}(k_1, k_2)$ from (4.76)
      Compute perturbed transfer matrix $\mathcal{L}(k_1 + \epsilon, k_2)$ from (4.76)
      Compute perturbed transfer matrix $\mathcal{L}(k_1, k_2 + \epsilon)$ from (4.76)
      Compute dominant eigenvalue λ_{max} of $\mathcal{L}(k_1, k_2)$, $\lambda_{max}^{k_1+\epsilon}$ of $\mathcal{L}(k_1+\epsilon, k_2)$ and $\lambda_{max}^{k_2+\epsilon}$ of $\mathcal{L}(k_1, k_2 + \epsilon)$
      Compute $\Delta = |\frac{1}{\epsilon} \ln \frac{\lambda_{max}^{k_1+\epsilon}}{\lambda_{max}} - \langle \sigma_1^t \sigma_2^{t+1} \rangle| + |\frac{1}{\epsilon} \ln \frac{\lambda_{max}^{k_2+\epsilon}}{\lambda_{max}} - \langle \sigma_2^t \sigma_1^{t+1} \rangle|$
      if $\Delta < \delta$ **then**
        $s_1 = k_1$, $s_2 = k_2$
      end if
    end for
  end for
  // Step refinement
  $k_1^0 = s_1$, $k_2^0 = s_2$, $dk = dk/10$
end while
$h_1 = s_1$, $h_2 = s_2$
Compute $\mathcal{L}(h_1, h_2)$ from (4.76)
Compute right and left dominant eigenvectors
Compute maximum entropy transition matrix from (4.78)
Compute stationay distribution from (4.79)

Though the less sophisticated approach presented earlier in this chapter is more general in scope, this algebraic approach seems very promising for systems that can be expressed in terms of a transfer matrix. Note that in spite of its anlytical flavour, this approach also raises computational challenges since it makes necessary to evaluate eigenvalues of potentially large matrices and approximate numerically partial derivatives of these eigenvalues.

Chapter 5

Complexity in elementary cellular automata

Our approach in the previous chapter was rather phenomenological since we were concerned with the sequence of states a system goes through over time. As we saw, a sequence can be more or less difficult to predict, but in any case the underlying process from which the sequence is generated is always hidden and inaccessible from the data themselves. In this chapter our perspective will be quite different since we shall focus on the complexity of the dynamics itself and the way it can be grasped through information theory.

5.1 Elementary cellular automata

Due to the diversity of systems considered in studying complexity, the difficulty is to find a dynamics that can serve as a touchstone for the measures of complexity we want to investigate in this chapter. Here we choose to concentrate our attention on a class of dynamics that are discrete both in space and time, known as *one-dimensional elementary cellular automata* (ECA). A one-dimensional ECA is a chain of units such that each unit has two possible states $\{0, 1\}$. At each time step a unit looks at its neighbourhood constituted by its left and right nearest neighbours and itself, and updates its state according to a prescribed rule. This rule is deterministic and is the same for all units of the chain (which is usually taken to be periodic). Since there are $2^3 = 8$ possible configurations of the neighbourhood, and each configuration should lead the central cell to go to one of two possible states, there are in this setting $2^8 = 256$ distinct evolution rules overall. Some among these 256 rules are actually identical, so that it is sufficient to consider 88 of them [110].

Besides their extreme simplicity, ECA have been thoroughly investigated (see [28] for an introduction) because it can be observed empirically that they can display a limited number of different behaviours. This observation led to a classification scheme due to Wolfram [109], who assigned them to four different classes according to their behaviour as follows:

CLASS	RULES
I	0, 8, 32, 40, 128, 136, 160, 168
II	1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 19, 23, 24, 25, 26, 27, 28, 29, 33, 34, 35, 36, 37, 38, 42, 43, 44, 46, 50, 51, 56, 57, 58, 72, 73, 74, 76, 77, 78, 104, 108, 130, 131, 132, 133, 134, 138, 140, 142, 152, 154, 156, 162, 164, 170, 172, 178, 184, 200, 204, 232
III	18, 22, 30, 45, 60, 90, 105, 129, 146, 150, 161
IV	41, 54, 106, 110

Table 5.1: A compendium of the 88 inequivalent one-dimensional ECA.

1. Class I regroups ECA converging to a homogeneous pattern;
2. Class II displays an inhomogeneous stable pattern or periodic behaviour;
3. Rules in class III display pseudo-random patterns;
4. Class IV, which supposedly is the closest to our intuitive conception of complexity, regroups automata which exhibit slowly building up and decaying patterns.

For reference, table 5.1 lists the 88 inequivalent elementary rules and the Wolfram class they fall in.

This classification scheme has the drawback that the class of an automaton cannot in general be determined from the evolution rule itself; the determination requires instead to let some arbitrary initial condition generate a pattern from which the class can eventually be deduced, essentially on a visual basis. Our purpose in this chapter is to confront this standard classification scheme to two alternative schemes suggested by information theory.

The first approach, which is the object of section 5.2, is global and relies on maximum entropy models based on constraints imposed on marginal distributions. It aims at making more precise the claim, made in the introduction, that there was a relation between complexity and entropy. So far this connection is somewhat loose, since in particular we have not explained if, and how, the idea that "complexity arises when a system is more than the collection of its parts" could be made mathematically sound using the tools of information theory introduced in the previous chapters. We explain here how the principle of maximum entropy provides a mathematically sound interpretation, first proposed by Schneidman & al. [88], of what "being more than a collection of parts" possibly means and allows setting up a precise definition of complexity. We then apply the notion of complexity so defined to ECA and check to what extent it overlaps with Wolfram's classification.

In a second part, dealt with in section 5.3, we adopt a more 'local' viewpoint developed in [83], which focusses exclusively on the statistical dependences of neighbouring cells and in particular on the transfer of information from the neighbourhood to the target cell. This allows to define *information processing features* on the basis of which ECA rules can be compared.

5.2 Maximum entropy characterization of complexity

5.2.1 Decomposing multi-information using the MEP

Our first task is to understand how the fuzzy intuition that some systems should be "more than the collection of their parts" (in a modern formulation to be found for instance in [11]) has to go through several reformulations paving the way to a mathematical treatment. Such a reformulation would be that "a complex system cannot be fully understood by looking separately at its elements" [12]; the difficulty is however only shifted from the meaning of "being" to that of "understanding". Things become clearer when working in a probabilistic framework, because then the understanding of the system is equalled to the probability distribution describing the system's behaviour under repeated trials. Though much less ambitious than a detailed knowledge of the entities constituting the system and the relationships they entertain with each other, this approach lies nonetheless at the foundation of statistical mechanics (moreover all concepts introduced in the previous chapter of course make sense only as long as a probability distribution is available).

If understanding a system amounts to knowing the probability distribution describing it, then the sentence above can be rephrased by saying that *the joint distribution characterizing a complex system cannot be reconstructed from the distributions characterizing its subparts*. This statement is almost mathematically rigorous except for the fact that the reconstruction could be carried through in different ways. It was suggested by [88, 87, 7] to let the principle of maximum entropy come into the play at this stage as the appropriate way to reconstruct the full distribution from incomplete knowledge, namely knowledge of marginals of lower order.

To assess the discrepancy between a maximum entropy reconstruction and the target distribution the most direct way to proceed is to use the Kullback-Leibler divergence introduced in chapter 1. Another approach to compare distributions would be to compare their moments, but in the context of maximum entropy models this approach is bound to fail to a large extent since by construction many of the low-order moments of the maximum entropy reconstruction are equal to those of the targeted distribution. The KL approach seems therefore more appropriate and has been widely employed (for instance in [94] in a similar context), though it does not give a truly complete picture.

Let us denote by p a joint distribution describing a system of N variables, and by $p^{(k)}_{ME}$ the maximum entropy reconstruction based on marginals of a given order k (*i.e* marginals over k variables). The error commited so doing can be estimated by computing the Kullback-Leibler divergence $D(p, p^{(k)}_{ME})$. The larger the subsets considered, the

more accurate the resulting approximation will be since smaller-order marginals may be recovered from larger-order ones, whence the inequality $D(p, p_{ME}^{(k-1)}) \geq D(p, p_{ME}^{(k)})$. We can therefore define the difference

$$C_k := D(p, p_{ME}^{(k-1)}) - D(p, p_{ME}^{(k)}) \geq 0 \tag{5.1}$$

as the gain in quality of the reconstructed distribution when basing our guess on subsets of size k instead of $k-1$, and therefore quantifies specifically the role played by interactions of order k. This quantity is sometimes referred to as the *connected multi-information of order k* [88], but this name will not be used here (the reference to "connectedness" being misleading in our opinion).

Performing the telescopic sum of all coefficients gives

$$\sum_{k=2}^{N} C_k = D(p, p_{ME}^{(1)}) - D(p, p_{ME}^{(N)}). \tag{5.2}$$

We explained in section 3.2 that the maximum entropy estimate based on univariate marginals was nothing but the factorized distribution, that is $p_{ME}^{(1)} = \prod_{i=1}^{N} p(x_i)$. Moreover, $p_{ME}^{(N)}$ is trivially p itself, hence $D(p, p_{ME}^{(N)}) = 0$. It follows that we get finally[1]

$$\begin{aligned} \sum_{k=2}^{N} C_k &= D\left(p(\mathbf{x}), \prod_{i=1}^{N} p(x_i)\right) \\ &= \sum_{\mathbf{x}} p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{\prod_{i=1}^{N} p(x_i)} = T. \end{aligned} \tag{5.3}$$

We can recognize in the last expression the definition (2.39) of total information. Therefore the introduction of the coefficients defined by (5.1) allows us to decompose the total informational content in a system as a sum of terms each quantifying the role played by a certain order of interaction. As could have been expected, the total interdependence in a system is built by addition of pair-wise, triplet-wise, and so on, interactions.

An important question is whether all subsets of a given order should be treated on the same footing when there exists a notion of distance between variables, as would be the case for instance if the system were put on a graph and each variable assigned to a node (more generally, such a distance *has* to exist in all cases where a focus is put on "multi-scaleness"). Considering for instance the case of pairs, co-dependence between variables remote from each other will intuitively be smaller than between two neighbouring variables, so that it seems acceptable to discard pairs constituted by distant elements. On another side, these loose pairs are by far more numerous than tightly-bound ones, so that although their contribution may be weak when envisaged individually, it cannot

[1] Note that the sum could be started from $k = 1$, in which case it would add up to the KL divergence between p and the uniform distribution. This would make necessary to assign null interdependence to uniform distributions only, and not to factorizable ones.

be neglected anymore when considered globally. But this fact that the number of pairs (more generally n-tuples) grows combinatorially implies that considering them all becomes numerically difficult. Since in the experiments to follow we focus on systems that do display such a notion of distance, provided by the topology of the graph on which our variables are put, our viewpoint will be to retain as admissible n-tuples only those consisting of connected variables, *i.e.* n-tuples whose restricted graph is connected.

We should emphasize that the idea of quantifying the role played by successive orders of interaction in the informational content of cellular automata has already been adressed long ago by Lindgren and Nordahl [67, 68]. However, these authors did not resort to MaxEnt methods but use instead a decomposition of the entropy rate (see also chapter 4), which makes this tool restricted to one-dimensional topologies (this limitation of their work was actually an important motivation for undertaking the study exposed in the present chapter).

5.2.2 Computation of decomposition coefficients

In the context of a numerical study, theoretical results like (3.9) are of little use and it becomes necessary to resort to numerical procedures to estimate the successive maximum entropy distributions. One such procedure is the so-called *iterative scaling algorithm* [19] whose principle is, starting from some initial distribution, to consider all possible n-tuples of variables in sequence, each time adjusting the corresponding marginal. Denoting by $p^{(k)}$ the distribution obtained after k such adjustments, S_k the subset considered at the k-th step and p_{S_k} the marginal distribution of this set, then the procedure is defined by

$$p^{(k)} := p^{(k-1)} \frac{p_{S_k}}{p_{S_k}^{(k-1)}}. \tag{5.4}$$

The order in which the n-tuples are examined does not alter the distribution we converge to. As a rule of thumb, it seems that - at least in the setup considered here - going twice through each n-tuple is enough to reach a satisfying solution. Proofs of convergence can be found in [19].

Since the computation of $p_{ME}^{(k)}(\mathbf{x})$ requires, for all the 2^N values of its arguments, to visit at least twice each of the $\frac{N!}{k!(N-k)!}$ k-subsets and compute marginals involving $O(2^{N-k})$ summations, the total cost of this operation is of order

$$O\left(\frac{2^{2N-k}N!}{k!(N-k)!}\right). \tag{5.5}$$

The whole procedure leading to the determination of all maximum entropy distributions and C_k coefficients therefore consumes a time of order

$$O\left(\sum_{k=0}^{N} \frac{2^{2N-k}N!}{k!(N-k)!}\right) = O\left(2^{2N}\left(\frac{3}{2}\right)^N\right) = O(6^N). \tag{5.6}$$

In the case where only adjacent subgroups are retained the complexity reduces to $O(\sum_{k=0}^{N} 2^{2N-k}N) = O(N4^N)$. The algorithm leading to the determination of the coefficients is therefore exponentially demanding. Our investigations will therefore be limited to systems tractable with a standard desktop computer, namely $N = 14$ when adjacent subsets only are considered and $N = 10$ when all subsets are. An obvious issue with such small systems is that they eventually reach a periodic regime with period at most 2^N for binary agents; however the systems we investigate here reach equilibrium on a much shorter time scale, so that the periodic behaviours we observe can hardly be ascribed to finite-size effects.

We give in an Appendix a commented code performing the total information decomposition (5.3) by means of the iterative scaling procedure (5.4). Though the code provided is not particularly performant (being written in Python), it has the advantage of being relatively self-contained (except for the use of a built-in function returning the power set of an ensemble).

5.2.3 Computational framework

Before moving on, we should briefly address how the temporal evolution of the probability density is handled. Often this is done by means of Monte-Carlo simulations [76], letting evolve copies of the system and reconstructing the probabilities by sampling trajectories. Here we follow an alternative approach which is allowed by the fact that the dynamics of ECA is memoryless, whereby we mean that the possible states (here to be understood as the states of the system as a whole) of an ECA at some given instant t depend only on its previous state at time $t-1$. The dynamics can therefore be described adequately in terms of a *(memoryless) Markov chain*. The joint probability characterizing the system at time t is therefore determined in terms of the probability at time $t-1$ and of the transition matrix $\mathbf{w}$ such that w_{ij} denotes the probability of transitioning from state i to state j, as

$$p(s,t) = \sum_{s'} p(s', t-1) w_{s's}. \tag{5.7}$$

Apart from allowing a more straightforward transition from numerical exploration to theoretical investigation, this approach allows to follow simultaneously all trajectories down to the least probable ones since it handles all initial conditions as long as they are not explicitly assigned zero probability; this will turn out to be a crucial feature.

Note that while the probability vector and the transition matrix are heavy to handle, this does not imply that Monte-Carlo sampling would be more suited to our purpose, for two reasons. First, in the context of ECA the sparsity of $\mathbf{w}$ allows an efficient computation of (5.7), so that this step is negligible compared to the determination of the C_k coefficients; the way we get the probabilities thus does not impact much the algorithm as a whole. Second, sampled distributions provide a poor estimator of the total information, and *a fortiori* of its decomposition.

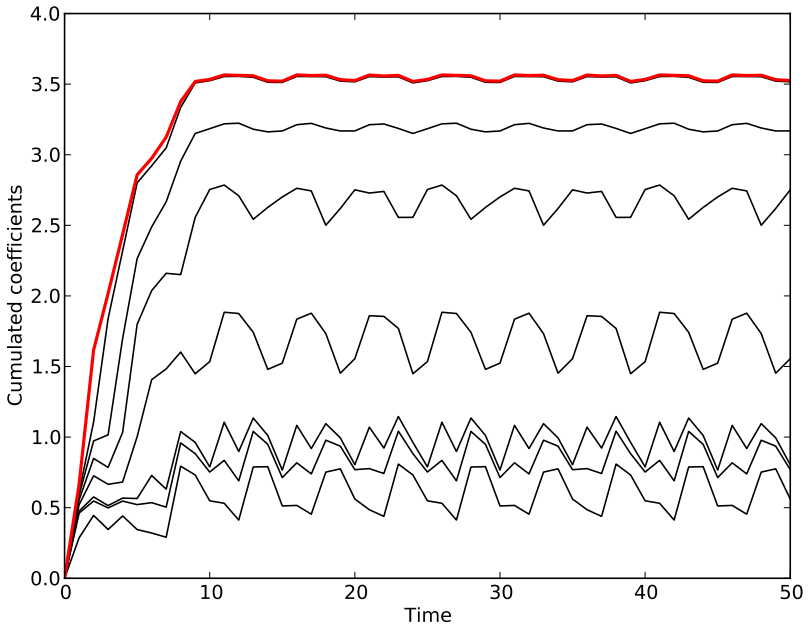


Figure 5.1: Time evolution of C_k coefficients in rule 110 (adjacent subsets). Coefficients are displayed cumulatively, *i.e.* we show successively (from bottom to top) C_2, $C_2 + C_3$, *etc.* The sum converges to M (red curve).

5.2.4 Total information decomposition in ECA

Decomposition coefficients

As a first step we compute the time evolution of C_k coefficients for ECA of size $N = 10$ put on a periodic string-like topology, as well as the behaviour of the total information. Figure 5.1 shows the result for rule 110, which is a typical instance of an automaton displaying complex (in the sense of Wolfram) behaviour. All coefficients except C_9 and C_{10} provide a significant contribution to the total information (as explained in the caption the coefficients are displayed cumulatively). After a transient phase of around ten steps, the system enters a periodic regime. While the total information is almost constant in this regime, the respective contributions show a much stronger variability. C_4, for instance, oscillates wildly, periodically decaying close to zero. On the contrary, the contribution of C_8 is nearly constant in the stationary regime.

For comparison we show in figure 5.2 the corresponding picture when all subsets of a given order, instead of adjacent ones only, are taken into account. The main difference is that here five coefficients are enough to reconstruct T, so that C_7 and C_8 play no significant role. On the whole the behaviour of the remaining coefficients is not qualitatively altered. Note however how our decision to rule out non-adjacent subsets introduced spurious high-order dependences.

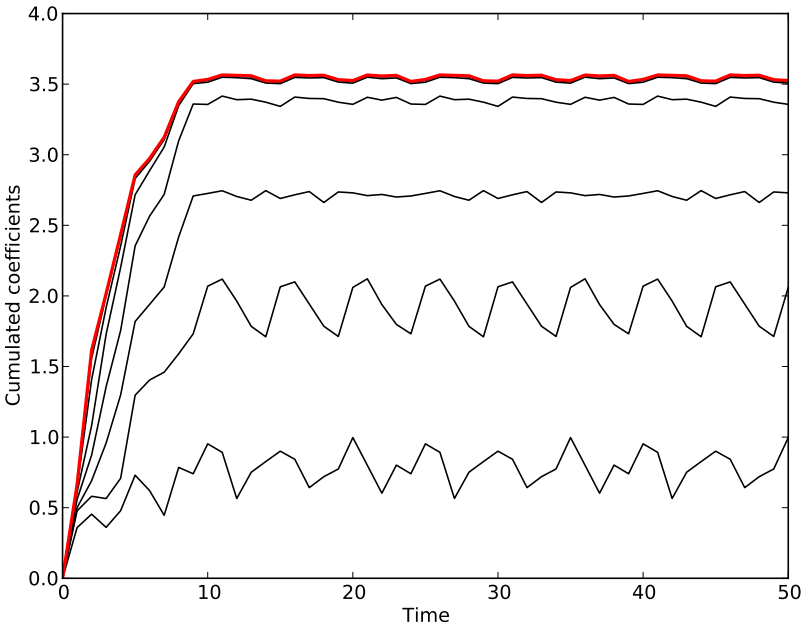


Figure 5.2: Time evolution of C_k coefficients in rule 110 (all subsets)

This behaviour constrasts sharply with figure 5.3, which displays the same plot for rule 90. This rule is characterized by the fact that the sole contribution to the total information is provided by the coefficient C_9, all other orders of interaction vanishing. Interestingly, rule 90 is nothing but the dynamics obtained by applying the XOR operator on the two neighbours of a variable and assigning the result to the variable itself. This highlights the fact that the notion of "order of interaction" as employed in the current context does not quite overlap with what we could expect from, for instance, classical kinetic theory [2]. There, "interaction of order n" would be understood in terms of the functional form of the Hamiltonian (in the sense that the latter could not be decomposed as a sum of functions involving less than n variables). The case of rule 90 would thus correspond to interactions of order two, and ECA in general to interactions of order three. It would then be difficult to justify the appearance of higher orders of interaction. An interesting open question would be to find a dynamics such that all informational content is brought in by interactions of order k in the sense of this decomposition, *i.e.* $C_k = M$ and $C_l = 0 \; \forall l \neq k$.

Averaged coefficients in the stationary regime

We shall now get rid of the transient phase, details of which depend on the probabilities we initially assigned to each configuration (each cell is initially given equal probability to take value 0 or 1). Moreover, in order to simplify our analysis, we shall discard temporal

[2] We face here an issue quite similar to the one pointed in chapter 3, that the coupling deduced from (3.6) are different from actual couplings.

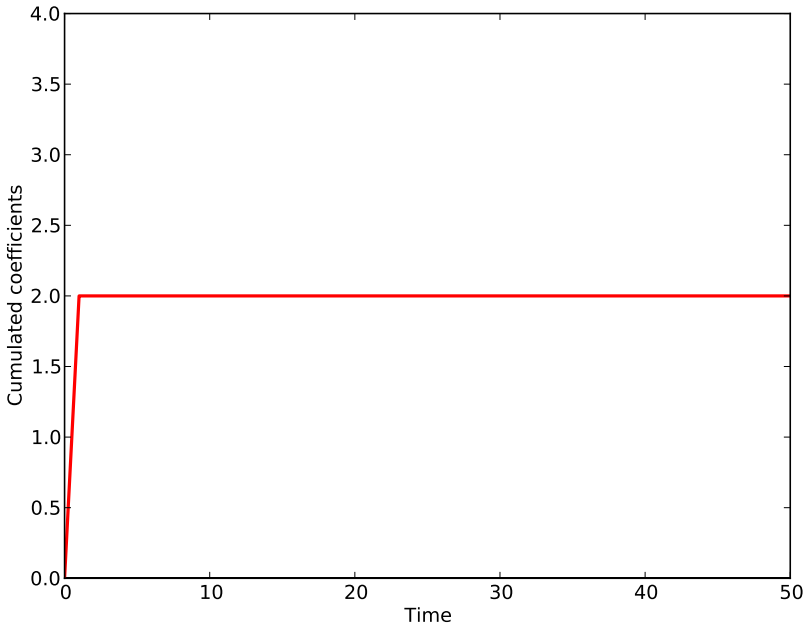


Figure 5.3: Time evolution of C_k coefficients (adjacent subsets) in rule 90. Only C_9 contributes.

variations of C_k coefficients by considering their average value in the stationary regime. From now on, coefficients will therefore be averaged when necessary over the stationary phase, so that all forthcoming statements about C_k coefficients will be statements about the average value $\langle C_k \rangle$ of these coefficients in this regime. In particular, we shall focus on the size $\langle \Sigma \rangle$ up to which the subsets should be considered in order to recover 90% of the total information in the stationary regime and investigate how $\langle \Sigma \rangle$ varies with the size of the system.

As shown in figure 5.4, two behaviours can be identified at first glance. First, in some rules $\langle \Sigma \rangle$ does not show any dependance on N. This is for instance the case of R14, for which knowing 5-variable marginals is always sufficient to reconstruct the full joint distribution up to 10%. A variation on this theme is shown by R32, where $\langle \Sigma \rangle$ oscillates between $\langle \Sigma \rangle = 2$ and $\langle \Sigma \rangle = 3$ depending on whether N is odd or even respectively. This may be explained by noting that there exists one initial configuration which does not converge to a stationary homogeneous pattern; namely, the configuration 01010101010101 gets replicated over time except for a one-cell shift at each iteration. Such an initial pattern is however possible only for even values of N, so a periodic configuration is possible only in this case, while for N odd the automaton necessarily settles down to a uniform configuration. This example provides a first clue that pathological initial conditions are important in our context.

At the opposite, other rules display a $\langle \Sigma \rangle$ that grows linearly with N, as is nicely illustrated by rules 73 and 106. This means that any attempt to infer global properties from subsystems is condemned to give intrinsically flawed results.

Besides constancy and linear growth, one more intermediate behaviour is instanciated by R54 and R110, where neither linear growth nor stabilization is obvious. The limited size range at our disposal does not allow us to conclude about the behaviour of such cases, which can be interpreted either as rules where $\langle\Sigma\rangle$ stabilizes at some high value, meaning that inference is conceptually possible but technically difficult, or as rules where $\langle\Sigma\rangle$ grows in a sublinear way (but again, this precludes inference at all).

Let us briefly digress about the situation where all subsets are considered instead of adjacent ones only. There, as expected, the size of subsets to be considered in order to reconstruct a specified fraction of M is almost always smaller than when unconnected ones are to be discarded. However the distinctive features underlined above are preserved.

Relation to Wolfram classes

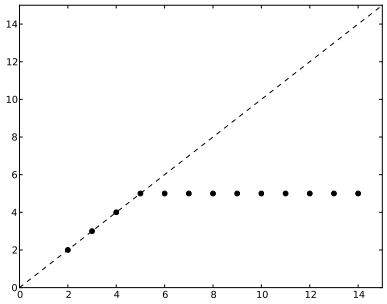
Our primary purpose of this chapter was to put alongside the notion of complexity stemming from the maximum entropy reconstruction and complexity as displayed in Wolfram's classification scheme. We should therefore examine whether or not some behaviours can be found to be common to all rules pertaining to a given Wolfram class.

We found that for all 8 rules in class I $\langle\Sigma\rangle$ takes a constant value, though this value varies from one rule to the other (the tricky case of R32 has already been discussed above, and R160 is very similar). This homogeneity of behaviours is in agreement with the simplicity of configurational patterns converged to in the stationary regime, whatever the initial conditions.

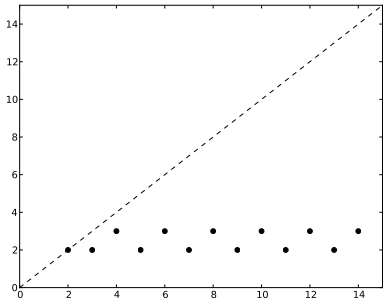
At the other end of the spectrum, several among the 11 rules belonging to class III display growing $\langle\Sigma\rangle$. None of these rules shows the simple behaviour encountered in class I. Again, this is in agreement with the chaoticity of rules pertaining to class III. However, a peculiarity of class III is that for some rules the total information goes to 0 as N grows, so that $\langle\Sigma\rangle$ is actually ill-defined. An instance thereof is provided by R30, which is occasionally used as a pseudo-random number generator, and R90, illustrated in figure 5.4.

Unfortunately, things are no longer that clear when we come to considering classes II and IV. Class II regroups no less than 65 of the 88 inequivalent rules. At least 41 of them display the same simple type of behaviour already encountered in class I, which is fine since rules in class II are not expected to present a high level of complexity. Nonetheless, the remaining 24 rules exhibit growing $\langle\Sigma\rangle$. The choice of initial configuration plays an important role in this respect. We already met above a rule (R32) whose classification depended tightly on the initial configuration chosen, and, implicitly, on the size of the system. Another such instance is provided by rule 73, which is classified as class II due to the appearance of "walls" splitting the configurations into sub-configurations which, being of finite size, will necessarily repeat themselves, hence the attribution to class II. It may however happen that the inital configuration is chosen in such a way as to forbid the appearance of such separating walls, in which case the dynamics should better be classified as chaotic. This is this chaoticity that shows up in the behaviour of $\langle\Sigma\rangle$.

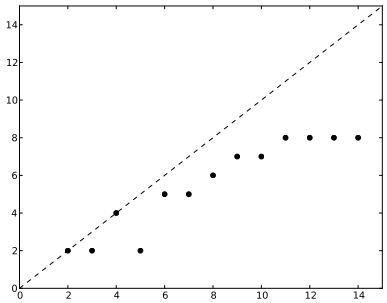
Class IV, finally, regroups only four inequivalent rules. Among these, three of which are shown above, one (R106) shows growing $\langle\Sigma\rangle$ while another (R54) seems to converge



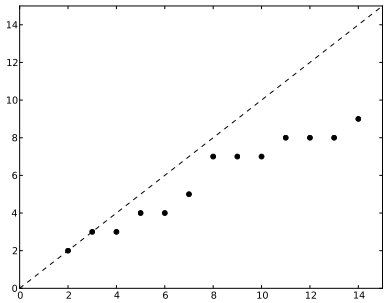
(a) R14



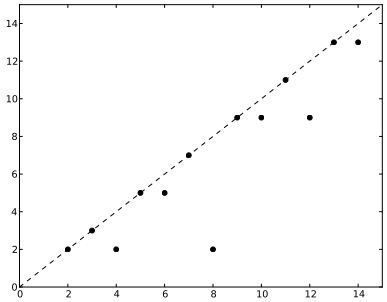
(b) R32



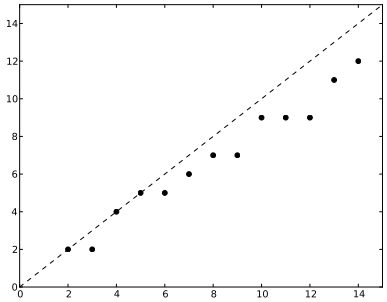
(c) R54



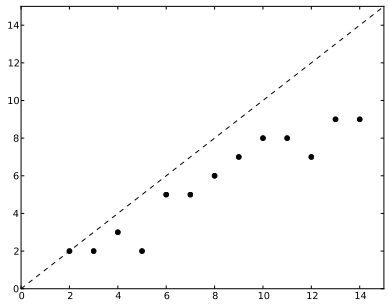
(d) R73



(e) R90



(f) R106



(g) R110

Figure 5.4: A representative sample of rules. Horizontal axis displays the size of the automaton, vertical axis displays the required size of subsets to recover 90% of the informational content. The dotted line corresponds to $\langle \Sigma \rangle = N$.

towards a constant value of $\langle\Sigma\rangle = 8$ (note however the large value of $\langle\Sigma\rangle$). The behaviour of rules 41 and 110 is difficult to infer from the size range we considered.

5.3 Information processing in ECA

The first section of this chapter could be said to correspond in a sense to a global approach since all the analysis there was based on the joint distribution p_N. In this section we would like to examine instead another approach we developed in [83], which is local since it focusses exclusively on the statistical dependences between neighbouring cells.

The intuition behind this approach is that long-range interactions or codependences are built up from elementary interactions occurring at the level of neighbouring cells. Cells made codependent by low-level interactions in turn interact with more and more distant parts of the system, each time exchanging information through the same elementary mechanism; so doing cells remote from each other come to be correlated through this kind of dominoes game. Our purpose here is to use information theory to try to capture quantitatively how local interactions posibly lead to complex emergent behaviours. In this viewpoint a cell's new state reflects its past interactions in the sense that it stores mutual information about the past states of upstream neighbouring cells. In the next iteration a downstream neighbouring unit interacts with this state, implicitly transferring this information and integrating it together with the information provided by others in order to build its own new state, and so on. Then each interaction among dynamical units is interpreted as a Shannon communication channel and we aim to trace the onward transmission and integration of information through the network of communication channels.

We shall retain the probabilistic framework set up in the previous section and deal with one- or two-timestep probabilities. In particular, denoting by V_i the neighbourhood of cell i, we shall express as

$$p(x_i(t+1), v_i(t)) \tag{5.8}$$

the probability that at time t the neighbourhood of i is in a state v_i and that at the next iteration the impacted cell takes value x_i. More relevant here however will be the conditional probability

$$p(x_i(t+1)|v_i(t)) = w_{v_i \to x_i} \tag{5.9}$$

With this conditional probabilty the joint probability (5.8) can be rewritten as

$$p(x_i(t+1), v_i(t)) = w_{v_i \to x_i} p(v_i(t)). \tag{5.10}$$

It is important to keep in mind that the dynamics of an ECA being deterministic, the conditional probabilities $w_{v_i \to x_i}$ can only take value 0 or 1. The probabilistic component thus relies on the initial distribution $p(v_i(t))$ only.

Example: As an example we can consider Rule 110, an elementary cellular automaton which is well known for displaying complex behaviour. The lookup table of this dynamics is given by

	0	1
000	1	0
001	0	1
010	0	1
011	0	1
100	1	0
101	0	1
110	0	1
111	1	0

(5.11)

so that we have $w_{000\to 0} = 1$, $w_{000\to 1} = 0$, and so forth.

5.3.1 Information processing features

The peculiarity of elementary cellular automata stems from the fact that the future of a cell within one time step is determined only by the present state of the right and left neighbours of the cell as well as by the present state of the cell itself. We can there therefore distinguish three types of information that contribute to the information flow between x_i and its neighbouhood:

1. *Memory*

   This is the mutual information between $x_i(t)$ and $x_i(t+1)$, or the knowledge provided by x_i about its state at the next iteration, namely

   $$I_{mem} = I(x_i(t), x_i(t+1). \tag{5.12}$$

2. *Transfer*

   This is the mutual information between $x_{i\pm1}(t)$ and $x_i(t+1)$, or the knowledge provided about x_i by either its left- or right neighbour. We can thus define both left- and right transfer information as

   $$I^L_{trans} = I(x_{i-1}(t), x_i(t+1), \tag{5.13}$$

   $$I^R_{trans} = I(x_{i+1}(t), x_i(t+1). \tag{5.14}$$

3. *Integration*

   This is the mutual information between the neighbourhood of x_i taken as a whole and x_i, namely

$$\begin{aligned} I_{int} &= I(v_i(t), x_i(t+1)) \\ &= I(\{x_{i-1}(t), x_i(t), x_{i+1}(t)\}, x_i(t+1). \end{aligned} \tag{5.15}$$

   Unlike the other two, integration information cannot therefore be reduced to one-to-one information, but represents, on the opposite, a global feature of the neighbourhood.

5.3.2 Information processing features in the randomized state

In this paragraph we want to show that in the very particular case where the inital state is randomized, *i.e.* when each cell is assigned an initial value of 0 or 1 with the same probability 1/2, then memory-, transfer- and integration information can be evaluated easily from the lookup table.

Integration

From the definition of integration information (5.15) and (2.33) we can rewrite sequentially

$$\begin{aligned} I_{int} &= I(v_i(t), x_i(t+1)) \\ &= \sum_{v_i(t), x_i(t+1)} p(v_i(t), x_i(t+1)) \ln \frac{p(v_i(t), x_i(t+1))}{p(v_i(t)) \cdot p(x_i(t+1))} \\ &= \sum_{v_i(t), x_i(t+1)} w_{v_i \to x_i} p(v_i(t)) \ln \frac{w_{v_i \to x_i} p(v_i(t))}{p(v_i(t)) \cdot p(x_i(t+1))} \\ &= \sum_{v_i(t), x_i(t+1)} w_{v_i \to x_i} p(v_i(t)) \ln \frac{w_{v_i \to x_i}}{p(x_i(t+1))} \\ &= \sum_{v_i(t), x_i(t+1)} w_{v_i \to x_i} p(v_i(t)) \ln \frac{w_{v_i \to x_i}}{\sum_{v_i(t)} w_{v_i \to x_i} p(v_i(t))}. \end{aligned} \tag{5.16}$$

We let intervene at this stage the randomization hypothesis which implies that

$$p(v_i(t)) = 2^{-n}, \tag{5.17}$$

where n denotes the size of a neighbourhood (in our case $n = 3$). Thus I_{int} can be rewritten further as

$$
\begin{aligned}
I_{int} &= \sum_{v_i(t),x_i(t+1)} w_{v_i\to x_i} p(v_i(t)) \ln \frac{w_{v_i\to x_i}}{\sum_{v_i(t)} w_{v_i\to x_i} p(v_i(t))} \\
&= 2^{-n} \sum_{v_i(t),x_i(t+1)} w_{v_i\to x_i} \ln \frac{w_{v_i\to x_i}}{2^{-n}\sum_{v_i(t)} w_{v_i\to x_i}} \\
&= -2^{-n} \sum_{v_i(t),x_i(t+1)} w_{v_i\to x_i} \ln \left(2^{-n} \sum_{v_i(t)} w_{v_i\to x_i} \right),
\end{aligned} \tag{5.18}
$$

where the last equality stems from the fact that w is either 0 or 1, implying that $w \ln w = 0$.

We now define a parameter λ through the formula

$$
\sum_{v_i} w_{v_i\to x_i} = \begin{cases} 2^n\lambda & \text{if } x_i = 1 \\ 2^n(1-\lambda) & \text{if } x_i = 0 \end{cases} \tag{5.19}
$$

In plain words, λ denotes the fraction of neighbourhood states that bring the central cell in state $x_i = 1$. For instance for rule 110 it can be deduced by simple inspection of the lookup table (5.11) that $\lambda_{110} = 5/8$. We can finally express I_{int} directly in terms of λ as

$$
\begin{aligned}
I_{int} &= -2^{-n} \sum_{v_i(t),x_i(t+1)} w_{v_i\to x_i} \ln \left(\lambda x_i + (1-\lambda)(1-x_i)\right) \\
&= -\sum_{x_i(t+1)} \left(\lambda x_i + (1-\lambda)(1-x_i)\right) \ln \left(\lambda x_i + (1-\lambda)(1-x_i)\right) \\
&= -(1-\lambda)\ln(1-\lambda) - \lambda \ln \lambda.
\end{aligned} \tag{5.20}
$$

Memory

The computation of memory information proceeds along similar lines. We get easily

$$
\begin{aligned}
I_{mem} &= I(x_i(t), x_i(t+1)) \\
&= \sum_{x_i(t),x_i(t+1)} p(x_i(t), x_i(t+1)) \ln \frac{p(x_i(t), x_i(t+1))}{p(x_i(t)) \cdot p(x_i(t+1))} \\
&= 2^{-n} \sum_{v_i(t),x_i(t+1)} w_{v_i\to x_i} \ln \frac{\sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to x_i}}{2^{-1}\sum_{v_i(t)} w_{v_i\to x_i}}.
\end{aligned} \tag{5.21}
$$

However the analysis has to be refined here. In analogy with (5.19), we define two parameters λ_0 and λ_1 that are the counterpart of λ restricted to the case that $x_i(t)$ takes value 0 or 1 respectively. Namely,

$$\sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to x_i} = \begin{cases} 2^{n-1}\lambda_0 & \text{if } x_i(t)=0 \text{ and } x_i(t+1)=1 \\ 2^{n-1}\lambda_1 & \text{if } x_i(t)=1 \text{ and } x_i(t+1)=1 \\ 2^{n-1}(1-\lambda_0) & \text{if } x_i(t)=0 \text{ and } x_i(t+1)=0 \\ 2^{n-1}(1-\lambda_1) & \text{if } x_i(t)=1 \text{ and } x_i(t+1)=0 \end{cases} \tag{5.22}$$

In the case of rule 110 we find for instance $\lambda_0 = 1/2$, $\lambda_1 = 3/4$. Of course we have a relation between λ_0, λ_1 and λ since $\lambda_0 + \lambda_1 = 2\lambda$.

Armed with this definition, I_{mem} becomes

$$\begin{aligned}
I_{mem} &= 2^{-n} \sum_{x_i(t+1)} \sum_{x_i(t)} \sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to x_i} \ln \frac{\sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to x_i}}{2^{-1}\sum_{x_i(t)}\sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to x_i}} \\
&= 2^{-n} \sum_{x_i(t)} \sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to 1} \ln \frac{\sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to 1}}{2^{-1}\sum_{x_i(t)}\sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to 1}} \\
&\quad + 2^{-n} \sum_{x_i(t)} \sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to 0} \ln \frac{\sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to 0}}{2^{-1}\sum_{x_i(t)}\sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to 0}} \\
&= 2^{-n} \sum_{x_i(t)} \sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to 1} \ln \frac{\sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to 1}}{2^{-1}(2^{n-1}\lambda_0 + 2^{n-1}\lambda_1)} \\
&\quad + 2^{-n} \sum_{x_i(t)} \sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to 0} \ln \frac{\sum_{x_{i-1}(t),x_{i+1}(t)} w_{v_i\to 0}}{2^{-1}(2^{n-1}(1-\lambda_0) + 2^{n-1}(1-\lambda_1))} \\
&= 2^{-1}\lambda_1 \ln \frac{\lambda_1}{2^{-1}(\lambda_0+\lambda_1)} + 2^{-1}\lambda_0 \ln \frac{\lambda_0}{2^{-1}(\lambda_0+\lambda_1)} \\
&\quad + 2^{-1}(1-\lambda_1) \ln \frac{(1-\lambda_1)}{2^{-1}((1-\lambda_0)+(1-\lambda_1))} \\
&\quad + 2^{-1}(1-\lambda_0) \ln \frac{(1-\lambda_0)}{2^{-1}((1-\lambda_0)+(1-\lambda_1))} \\
&= \frac{\lambda_1}{2} \ln \frac{\lambda_1}{\lambda} + \frac{\lambda_0}{2} \ln \frac{\lambda_0}{\lambda} + \frac{1-\lambda_1}{2} \ln \frac{1-\lambda_1}{1-\lambda} + \frac{1-\lambda_0}{2} \ln \frac{1-\lambda_0}{1-\lambda}.
\end{aligned} \tag{5.23}$$

Transfer

The case of transfer information proceeds exactly as for memory information, so that we shall skip detailed calculations. The only difference is that instead of defining parameters for fixed $x_i(t)$, we do it for fixed $x_{i+1}(t)$ (right transfer) or $x_{i-1}(t)$ (left transfer). More precisely, we define

$$\sum_{x_i(t),x_{i-1}(t)} w_{v_i\to x_i} = \begin{cases} 2^{n-1}\lambda_0^R & \text{if } x_{i+1}(t)=0 \text{ and } x_i(t+1)=1 \\ 2^{n-1}\lambda_1^R & \text{if } x_{i+1}(t)=1 \text{ and } x_i(t+1)=1 \\ 2^{n-1}(1-\lambda_0^R) & \text{if } x_{i+1}(t)=0 \text{ and } x_i(t+1)=0 \\ 2^{n-1}(1-\lambda_1^R) & \text{if } x_{i+1}(t)=1 \text{ and } x_i(t+1)=0 \end{cases} \tag{5.24}$$

$$\sum_{x_i(t),x_{i+1}(t)} w_{v_i\to x_i} = \begin{cases} 2^{n-1}\lambda_0^L & \text{if } x_{i-1}(t)=0 \text{ and } x_i(t+1)=1 \\ 2^{n-1}\lambda_1^L & \text{if } x_{i-1}(t)=1 \text{ and } x_i(t+1)=1 \\ 2^{n-1}(1-\lambda_0^L) & \text{if } x_{i-1}(t)=0 \text{ and } x_i(t+1)=0 \\ 2^{n-1}(1-\lambda_1^L) & \text{if } x_{i-1}(t)=1 \text{ and } x_i(t+1)=0 \end{cases} \tag{5.25}$$

Back to our example of rule 110, we find $\lambda_0^R = 1/2$, $\lambda_1^R = 3/4$, $\lambda_0^L = 3/4$ and $\lambda_1^L = 1/2$. As was the case for memory parameters, we still have the relation $\lambda_0^R + \lambda_1^R = \lambda_0^L + \lambda_1^L = 2\lambda$.

With (5.24), (5.25) we find at once that

$$\begin{aligned} I_{trans}^R = &\frac{\lambda_1^R}{2}\ln\lambda_1^R + \frac{\lambda_0^R}{2}\ln\lambda_0^R + \frac{(1-\lambda_1^R)}{2}\ln(1-\lambda_1^R) \\ &+ \frac{(1-\lambda_0^R)}{2}\ln(1-\lambda_0^R) - \lambda\ln\lambda - (1-\lambda)\ln(1-\lambda) \end{aligned} \tag{5.26}$$

$$\begin{aligned} I_{trans}^L = &\frac{\lambda_1^L}{2}\ln\lambda_1^L + \frac{\lambda_0^L}{2}\ln\lambda_0^L + \frac{(1-\lambda_1^L)}{2}\ln(1-\lambda_1^L) \\ &+ \frac{(1-\lambda_0^L)}{2}\ln(1-\lambda_0^L) - \lambda\ln\lambda - (1-\lambda)\ln(1-\lambda). \end{aligned} \tag{5.27}$$

5.3.3 Clustering of ECA according to their information features

Having proved that information features could be expressed in simple terms of the lookup table defining the dynamics is also interesting because of the link it ties with a classification scheme introduced by Langton [62]; indeed the λ we introduced above is precisely the parameter considered there. One might therefore expect that the generalized Langton parameters and information features introduced here could help in providing an accurate prediction of complexity classes.

To this end we computed the information features I_{int}, I_{mem}, I_{trans}^R and I_{trans}^L of the 88 inequivalent ECA and applied a clustering algorithm on the coordinates in the information-features space. The clustering scheme used takes as input the 88 vectors of information features and computes the Euclidian distance between these vectors. At each step of the procedure the two vectors or clusters separated by the smallest distance are clustered together. The distance between two clusters is defined as the distance between the most distant elements in each cluster.

The results are displayed in figure 5.5. ECA in class I are pictured in black, class II in red, class III in green and rules in class IV in blue. Though the global mismatch between the clusters emerging from the information features evaluated in the randomized state and the Wolfram classes is obvious, it should be pointed that there are certain cases where the clustering succeeds in predicting similarity of behaviour. This is the case in particular for rules 60, 90, 105 and 150 that are confounded in terms of their information features, and indeed belong all to class III. Unfortunately, other chaotic

rules are assigned to remote clusters. The converse situation also occurs where rules that belong to different classes are indistinguishable in terms of their features (see for instance rules 18 and 33 or 106 and 154).

Figure 5.6 presents the same clustering when informational features are evaluated not only in the randomized state but also in the stationary state. I_{int}, I_{mem}, I^R_{trans} and I^L_{trans} can no longer be calculated analytically but need to be found numerically. Enlarging the feature space has the effect that all uniform rules but one are clustered together. The remaining one, rule 168, has a transient regime which is essentially dominated by a translation of the initial state. This translational regime can be found as well in rules 2 and 130, which are classified in the same sub-spray as rule 168. Another good thing is that the spray bearing these uniform rules is distinct from the large branch bearing all chaotic or complex ECA; each of these sprays however does also contain rules pertaining to class II. It should also be noted that all rules in class III or IV are concentrated on two branches. These branches are unfortunately separated by another one bearing only periodic ECA. The overlap between Wolfram classes and clusters is therefore improved when considering information features in the stationary state, but the agreement is still far from perfect.

5.4 Discussion

The results of section 5.2 suggest that there is no obvious correspondence between complexity expressed in probabilistic terms using the maximum entropy decomposition scheme of [88], and complexity defined in terms of the evolution of some configurational pattern (except in some simple cases described above). In particular, we encountered several instances where expressing complexity in terms of maximum entropy probability distributions turned out to be non-intuitive. Though non-intuitivity is not necessarily unconvenient, this should at least prevent us from overinterpretations. Salient points to keep in mind are:

1. The notion of "order of interaction" as it appears here has little to do with the functional way it is usually presented in statistical physics.

2. A probabilistic approach allows dealing simultaneously with all possible initial conditions. The determination and discarding of particular initial conditions is of course possible, but it seems unlikely that it can be done *a priori*, and, therefore, seems meaningless (letting aside the fact that it would require a very considerable amount of work). Such a probabilistic characterization thus necessarily mixes up initial conditions giving rise to very different patterns, whence the discrepancy with definitions of complexity that rely on a specific initial condition.

3. Defining complexity as a scaling property brings into focus not so much the difference between simplicity and complexity but the difference between possible and impossible inference, or in other terms between reducibility and irreducibility. It also makes reducibility independent of any particular realization of a dynamics and dependent of the dynamics itself.

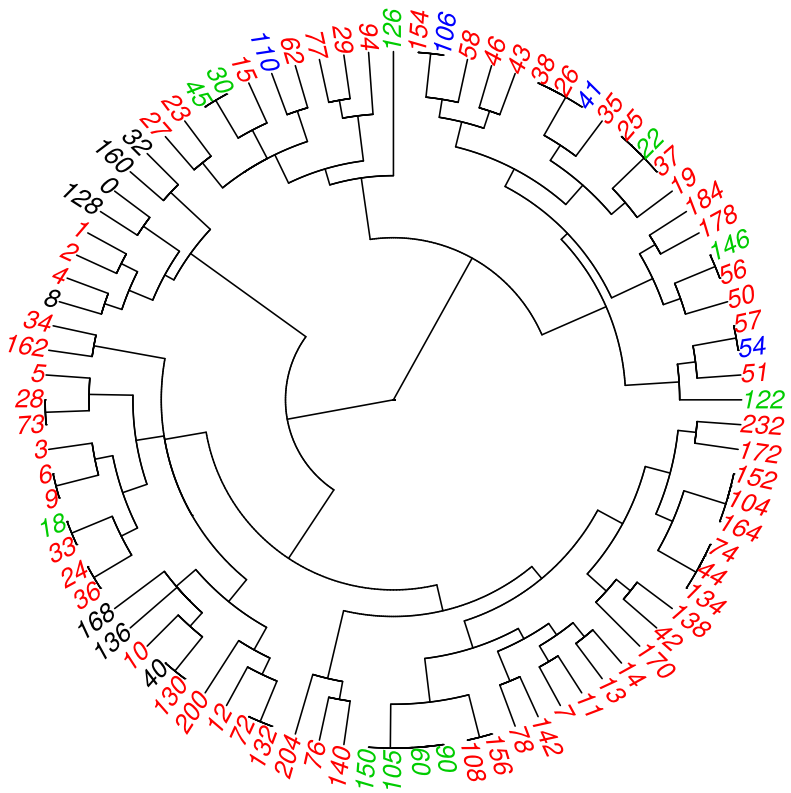


Figure 5.5: Inequivalent ECA clustered according to their informational features I_{int}, I_{mem}, I^R_{trans} and I^L_{trans} evaluated in the decorrelated state. CA in class I are shown in black, class II in red, class III in green and class IV in blue.

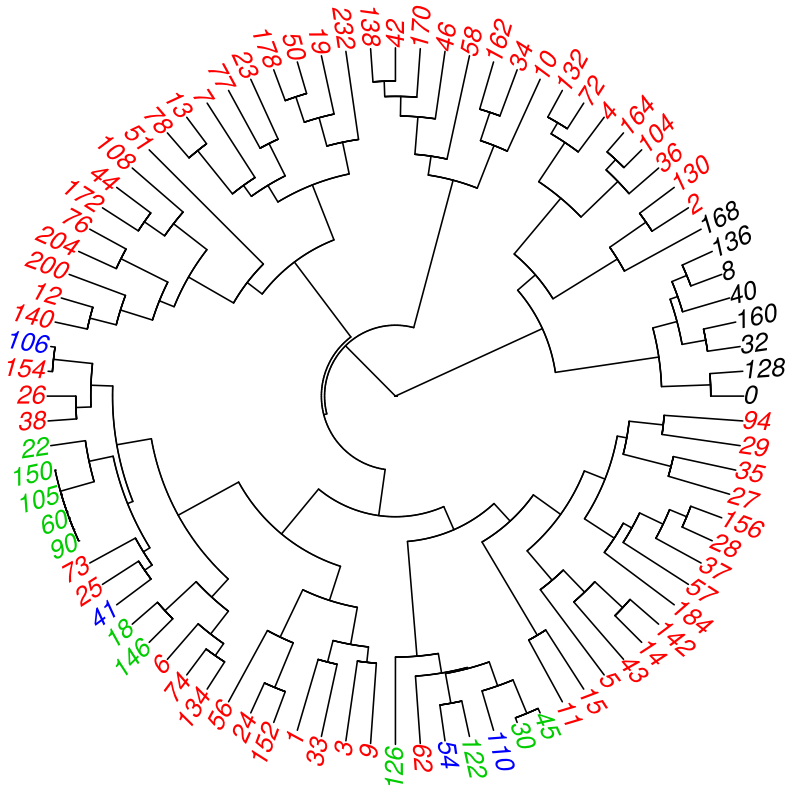


Figure 5.6: The same figure as 5.5, now with a feature space enlarged to account not only for features evaluated in the decorrelated space, but also in the stationary state.

Comparatively, the approach based on information flows occurring at the level of a neighbourhood seems more promising than the global approach. Dealing with local quantities also has the merit of being closer to the intuition. However we have seen that considering the information features at equilibrium brought a considerable improvemement of the clustering compared to the information features evaluated in the decorrelated state. The probability characterizing the entry state is therefore the result of the whole past history of the system, so that in a sense the information features can no longer be considered as strictly local. Overall, both approaches seem to capture a particular aspect of the dynamics considered, but none provides a completely satisfying picture.

Chapter 6

Revisiting the *Stosszahlansatz*

In chapter 3 we introduced maximum entropy models based on constrained marginals used in the previous chapter. We proved there that such distributions took a pseudo-factorized form such as (3.9). In particular in the simplest case where only *univariate* marginals are constrained this generalized notion of factorization happens to reduce to proper factorization. Since a maximum entropy hypothesis on marginals of order 1 amounts to a factorization hypothesis, our point in this chapter is to reverse the equivalence and envisage a factorization hypothesis as a maximum entropy hypothesis.

The relevance of this approach to statistical mechanics stems from the fact that from the very beginnings and at the very heart of kinetic theory of gases lies a technical assumption (known as *Stosszahlansatz* or *assumption of molecular chaos*) about the state of particles entering a binary collision, namely that the momenta of such particles are statistically independent of each other [99]. Of course this assumption did not prevent kinetic theory from achieving remarkable successes in diversified areas of statistical mechanics, fluid dynamics and others, and even from from a purely conceptual standpoint the controversies raised by the time irreversibility resulting from the *Stosszahlansatz* have been settled to a large extent, so that the molecular chaos can no longer be considered a central issue of fundamental physics.

But still: even though exact results have been obtained regarding its range of validity [95, 72], from a purist's perspective the ansatz is little more than an *ad hoc* assumption, and there are actually good reasons why it cannot hold in general. Unfortunately the way it should be complemented or generalized is anything but obvious, so it might seem that the ansatz is here to stay. Nonetheless our point in this chapter is to suggest that such a generalization can be achieved at the cost of a conceptual shift as to the actual meaning of the *Stosszahlansatz* itself. At first glance the scope of this assumption seems unambiguous, for the factorization hypothesis should be no more than the mathematical translation of a physical (though statistical) property of particles constituting the system. But if we consider that the factorization hypothesis on the 2-particle distribution is actually a maximum entropy estimate of this distribution, then the factorization becomes a heuristic hypothesis (almost) devoid of physical content. This move may appear at first as a rhetoric one, since from a mathematical standpoint the move is harmless. On the other hand it makes quite a difference when considered from a conceptual per-

spective, since while the factorization hypothesis is hard to amend on physical grounds the maximum entropy estimate lends itself nicely to generalization.

Our contribution in this chapter is, first, to show how the maximum entropy ansatz on the 3-particle distribution makes possible to close the BBGKY hierarchy of kinetic equations and derive a closed equation (6.16) describing the evolution of the 2-particle distribution [27]. Once the kinetic equation is set up, it becomes possible to follow the usual steps leading to the equilibrium distribution (6.32) and macroscopic balance equations. There will be however on the road a subtlety related to the definition of collisional invariants appropriate to the bilocal events under consideration, and it will happen that, besides conservation of (bilocal) mass, momentum and kinetic energy, it is necessary to consider a fourth invariant (6.20) that eventually accounts for the momentum correlation of particles.

6.1 Liouville equation and the BBGKY hierarchy

Let us consider N particles of mass m, the coordinates of which in phase space are their positions $\mathbf{x}_i$ and momenta $\mathbf{p}_i$. It will be convenient to define a condensed notation $\xi_i = (\mathbf{x}_i, \mathbf{p}_i)$. Let $f_N(\xi_1, ..., \xi_N, t)$ denote the joint distribution function characterizing the system. f_N obeys Liouville equation

$$\frac{df_N}{dt} = \frac{\partial f_N}{\partial t} + \sum_{i=1}^{N} \frac{\mathbf{p}_i}{m} \frac{\partial f_N}{\partial \mathbf{x}_i} + \sum_{i=1}^{N} \mathbf{F}_i \frac{\partial f_N}{\partial \mathbf{p}_i} = 0, \tag{6.1}$$

where $\mathbf{F}_i$ denotes the force exerted on particle i. We shall restrict ourselves to the case without external force and where particles interact pairwise through some radial potential $V(|\mathbf{x}_i - \mathbf{x}_j|) = V_{ij}$, so that $\mathbf{F}_i = -\sum_{j \neq i} \frac{\partial V_{ij}}{\partial \mathbf{x}_i}$. Reminding that f_N itself is normalized to $N!$, we can introduce the reduced s-particle distribution $f_s(\xi_1, ..., \xi_s, t) = \frac{N!}{(N-s)!} \int d\xi_{s+1} ... d\xi_N f_N(\xi_1, ..., \xi_N, t)$.

Liouville equation is a direct consequence of Newtonian dynamics, and as such is reversible. In particular it should be reminded [80, 66] that – in contradistinction with the entropy of the 1-particle distribution – the Shannon entropy of the N-particle density, namely

$$H((f_N) = -\int d\xi_1 ... d\xi_N f_N(\xi_1, ..., \xi_N) \ln f_N(\xi_1, ..., \xi_N), \tag{6.2}$$

is conserved by (6.1). The emergence of irreversibility in the 1-particle description has been more than extensively discussed[1], but it is clear nowadays that it results unavoidably from integrating out degrees of freedom that are irrelevant to the description [69].

[1]The non-technically-oriented reader can find an interesting *and* elementary account of irreversibility in Statistical Mechanics in [45].

By integrating Liouville equation $(N-s)$ times, one obtains [61] a dynamical equation for f_s given by the so-called *BBGKY hierarchy* (from the non-chronological list of its co-discoverers' names : Bogoliubov, Born, Green, Kirkwood, Yvon) :

$$\frac{\partial f_s}{\partial t} + \sum_{i=1}^{s} \frac{\mathbf{p}_i}{m} \frac{\partial f_s}{\partial \mathbf{x}_i} - \sum_{i=1}^{s} \sum_{j \neq i}^{s} \frac{\partial V_{ij}}{\partial \mathbf{x}_i} \frac{\partial f_s}{\partial \mathbf{p}_i} - \int d\xi_{s+1} \sum_{i=1}^{s} \frac{\partial V_{i,s+1}}{\partial \mathbf{x}_i} \frac{\partial f_{s+1}}{\partial \mathbf{p}_i} = 0. \tag{6.3}$$

This expression forms a hierarchy since the dynamics for f_s is expressed in terms of the higher-order distribution f_{s+1}. Of course each equation can be deduced from its higher-order precursor by integration, at the cost of an information loss. In what follows we shall denote the s-th equation of the hierarchy as BBGKY-s. BBGKY1 and BBGKY2, in which we are primarily interested here, are

$$\frac{\partial f_1}{\partial t} + \frac{\mathbf{p}_1}{m} \frac{\partial f_1}{\partial \mathbf{x}_1} = \int d\xi_2 \frac{\partial V_{12}}{\partial \mathbf{x}_1} \frac{\partial f_2}{\partial \mathbf{p}_1} \tag{6.4}$$

and

$$\frac{\partial f_2}{\partial t} + \frac{\mathbf{p}_1}{m} \frac{\partial f_2}{\partial \mathbf{x}_1} + \frac{\mathbf{p}_2}{m} \frac{\partial f_2}{\partial \mathbf{x}_2} - \frac{\partial V_{12}}{\partial \mathbf{x}_1} \left(\frac{\partial}{\partial \mathbf{p}_1} - \frac{\partial}{\partial \mathbf{p}_2} \right) f_2 = \int d\xi_3 \left(\frac{\partial V_{13}}{\partial \mathbf{x}_1} \frac{\partial f_3}{\partial \mathbf{p}_1} + \frac{\partial V_{23}}{\partial \mathbf{x}_2} \frac{\partial f_3}{\partial \mathbf{p}_2} \right), \tag{6.5}$$

where it should be understood that $f_1 = f_1(\mathbf{p}_1, \mathbf{x}_1, t)$, $f_2 = f_2(\mathbf{p}_1, \mathbf{x}_1, \mathbf{p}_2, \mathbf{x}_2, t)$ and $f_3 = f_3(\mathbf{p}_1, \mathbf{x}_1, \mathbf{p}_2, \mathbf{x}_2, \mathbf{p}_3, \mathbf{x}_3, t)$. Our purpose is to investigate the second of these equations. As stressed above, we shall not try to express BBGKY1 and BBGKY2 as a set of coupled equations relating f_1 and f_2, since such an approach would not "fit" nicely in the spirit of the BBGKY approach. Instead, we shall manage to truncate BBGKY2 in order to obtain a single, self-standing equation for f_2.

6.2 From BBGKY2 to the kinetic equation

We now proceed to write down the kinetic equation for f_2. All through, we shall retain the usual assumptions of kinetic theory [61, 66], leading us to neglect triple collisions : the streaming term for the two-particle distribution characterizing particles '1' and '2' will thus be altered by 1) binary collisions between '1' and another particle, '2' being spectator, and 2) binary collisions between '2' and another particle, '1' being spectator. Sticking tightly to the assumptions made in the 1-particle theory is important in order to guarantee that any new prediction arising in the present 2-particle description can be ascribed to the statistical description considered and not to the introduction of new physical assumptions.

The binary interaction is defined as occurring when two particles meet in a ball B of radius R. Defining ternary interactions is more subtle since, inasmuch as the interaction

potential is the same whatever the order of the interaction, it seems artificial to introduce a specific cutoff. We shall therefore define the range of triple collisions as the lenticular overlap of balls $B_R^{(1)}$ and $B_R^{(2)}$ characterizing the domain of interaction with '1' and '2' respectively. Neglecting triple collisions thus amounts to assuming that $|\mathbf{x}_1 - \mathbf{x}_2| > 2R$.

We first compute the contribution of collisions of '1' with '3', '2' being left aside. Let us recall that the collision term is given by

$$\left(\frac{\partial f_2}{\partial t}\right)_{coll} = \int d\xi_3 \left(\frac{\partial V_{13}}{\partial \mathbf{x}_1}\frac{\partial f_3}{\partial \mathbf{p}_1} + \frac{\partial V_{23}}{\partial \mathbf{x}_2}\frac{\partial f_3}{\partial \mathbf{p}_2}\right). \tag{6.6}$$

In the usual derivation of the Boltzmann equation from the BBGKY hierarchy, the right-hand side of BBGKY1 is transformed using BBGKY2. Similarly, we can transform $(\partial_t f_2)_{coll}$ using BBGKY3, namely

$$\begin{aligned}&\frac{\partial f_3}{\partial t} + \frac{\mathbf{p}_1}{m}\frac{\partial f_3}{\partial \mathbf{x}_1} + \frac{\mathbf{p}_2}{m}\frac{\partial f_3}{\partial \mathbf{x}_2} + \frac{\mathbf{p}_3}{m}\frac{\partial f_3}{\partial \mathbf{x}_3} - \frac{\partial V_{12}}{\partial \mathbf{x}_1}\left(\frac{\partial}{\partial \mathbf{p}_1} - \frac{\partial}{\partial \mathbf{p}_2}\right) f_3 \\ &\quad - \frac{\partial V_{13}}{\partial \mathbf{x}_1}\left(\frac{\partial}{\partial \mathbf{p}_1} - \frac{\partial}{\partial \mathbf{p}_3}\right) f_3 - \frac{\partial V_{23}}{\partial \mathbf{x}_2}\left(\frac{\partial}{\partial \mathbf{p}_2} - \frac{\partial}{\partial \mathbf{p}_3}\right) f_3 = \left(\frac{\partial f_3}{\partial t}\right)_{coll}\end{aligned} \tag{6.7}$$

(we do not make explicit the collision term $(\partial_t f_3)_{coll}$ since we shall cancel it soon anyway). Under usual dimensional assumptions, we can write $\partial_t f_3 \approx 0$ and $(\partial_t f_3)_{coll} \approx 0$ so that, substituting in the collision term, $(\partial_t f_2)_{coll}$ is rewritten as

$$\begin{aligned}&\left(\frac{\partial f_2}{\partial t}\right)_{coll} = \\ &\int d\xi_3 \left(\frac{\mathbf{p}_1}{m}\frac{\partial f_3}{\partial \mathbf{x}_1} + \frac{\mathbf{p}_2}{m}\frac{\partial f_3}{\partial \mathbf{x}_2} + \frac{\mathbf{p}_3}{m}\frac{\partial f_3}{\partial \mathbf{x}_3} - \frac{\partial V_{12}}{\partial \mathbf{x}_1}\left(\frac{\partial}{\partial \mathbf{p}_1} - \frac{\partial}{\partial \mathbf{p}_2}\right) f_3 + \left(\frac{\partial V_{13}}{\partial \mathbf{x}_1} + \frac{\partial V_{23}}{\partial \mathbf{x}_2}\right)\frac{\partial f_3}{\partial \mathbf{p}_3}\right) \\ &= \int d\xi_3 \left(\frac{\mathbf{p}_1}{m}\frac{\partial f_3}{\partial \mathbf{x}_1} + \frac{\mathbf{p}_3}{m}\frac{\partial f_3}{\partial \mathbf{x}_3}\right)\end{aligned} \tag{6.8}$$

(the last term vanishes due to the boundary condition $f_3(|\mathbf{p}| \to \infty) = 0$, the penultimate since '1' and '2' are supposed far apart from each other and the first because f_3 depends but weakly on $\mathbf{x}_2$). More precisely,

$$\left(\frac{\partial f_2}{\partial t}\right)_{coll} = \int_{\mathbf{x}_3 \in B_R^{(1)}} d\mathbf{x}_3 d\mathbf{p}_3 \left(\frac{\mathbf{p}_1}{m}\frac{\partial f_3}{\partial \mathbf{x}_1} + \frac{\mathbf{p}_3}{m}\frac{\partial f_3}{\partial \mathbf{x}_3}\right). \tag{6.9}$$

The following is standard [61]. We introduce the relative coordinate $\mathbf{r}_{13} = \mathbf{x}_3 - \mathbf{x}_1$ and use Gauss' theorem in order to rewrite $(\partial_t f_2)_{coll}$ as a surface integral, so that

$$\left(\frac{\partial f_2}{\partial t}\right)_{coll} = \int_{\mathbf{r}_{13}\in B_R} d\mathbf{r}_{13}d\mathbf{p}_3 \frac{\mathbf{p}_3-\mathbf{p}_1}{m}\frac{\partial}{\partial \mathbf{r}_{13}} f_3(\mathbf{x}_1,\mathbf{p}_1,\mathbf{x}_2,\mathbf{p}_2,\mathbf{x}_3,\mathbf{p}_3,t)$$
$$= \int_{S_R} d\mathbf{p}_3 d\Sigma \cdot \frac{\mathbf{p}_3-\mathbf{p}_1}{m} f_3(\mathbf{x}_1,\mathbf{p}_1,\mathbf{x}_2,\mathbf{p}_2,\mathbf{x}_3,\mathbf{p}_3,t)$$
$$= \int_{S_R^-\cup S_R^+} d\mathbf{p}_3 d\Sigma \cdot \frac{\mathbf{p}_3-\mathbf{p}_1}{m} f_3(\mathbf{x}_1,\mathbf{p}_1,\mathbf{x}_2,\mathbf{p}_2,\mathbf{x}_3,\mathbf{p}_3,t), \tag{6.10}$$

where $d\Sigma$ denotes the surface element of the sphere S_R such that $|\mathbf{r}_{13}| = R$. The southern hemisphere is interpreted as the contribution of oncoming collisions since $(\mathbf{p}_3-\mathbf{p}_1)\cdot d\Sigma < 0$, while the northern one is the contribution of ending collisions since $(\mathbf{p}_3-\mathbf{p}_1)\cdot d\Sigma > 0$.

Orienting the polar axis along the vector $\mathbf{p}_3-\mathbf{p}_1$ allows rewriting the dot product as $d\Sigma\cdot(\mathbf{p}_3-\mathbf{p}_1) = |\mathbf{p}_3-\mathbf{p}_1|R^2\sin\theta\cos\theta d\theta d\phi$. This can be re-expressed in terms of the surface element of the azimuthal plane such that $\theta = \pi/2$. Letting r denote the radial component on the plane, we obviously have $r = R\sin\theta$, whence $dr = \pm R\cos\theta d\theta$ (depending on θ being lesser or larger than $\pi/2$) and $d\Sigma\cdot(\mathbf{p}_3-\mathbf{p}_1) = \pm|\mathbf{p}_3-\mathbf{p}_1|d\omega$. The collision term can thus be rewritten as (approximating $\mathbf{x}_3 \approx \mathbf{x}_1$)

$$\left(\frac{\partial f_2}{\partial t}\right)_{coll} = \int_{after} d\mathbf{p}_3 d\omega \frac{|\mathbf{p}_3-\mathbf{p}_1|}{m} f_3(\mathbf{x}_1,\mathbf{p}_1,\mathbf{x}_2,\mathbf{p}_2,\mathbf{x}_1,\mathbf{p}_3,t)$$
$$- \int_{before} d\mathbf{p}_3 d\omega \frac{|\mathbf{p}_3-\mathbf{p}_1|}{m} f_3(\mathbf{x}_1,\mathbf{p}_1,\mathbf{x}_2,\mathbf{p}_2,\mathbf{x}_1,\mathbf{p}_3,t). \tag{6.11}$$

6.3 The *Stosszahlansatz* for BBGKY2

The procedure leading from the BBGKY1 equation to a consistent kinetic equation for f_1 is standard : the *Stosszahlansatz* asserts that before colliding two particles are uncorrelated, *i.e.* f_2 factorizes as $f_2(\xi_1,\xi_2) = f_1(\xi_1)f_1(\xi_2)$. This allows us to express the collision integral in terms of f_1, so that BBGKY1 becomes a closed equation for f_1. Since this factorization hypothesis may be supported from a physical standpoint, it is tempting to use this ansatz in the collision term for BBGKY2 as well. But this raises an issue: if BBGKY2 can be cast into an equation relating a streaming term expressed in terms of f_2 to a collision term expressed in terms of f_1, then this equation is obviously not consistent by itself and has to be supplemented, so as to obtain a system of coupled equations.

Our point is that this issue vanishes if the *Stosszahlansatz* is reconsidered as a heuristic ansatz instead of a physically-grounded assumption. We propose to reformulate it as follows: since the exact codependence of particles entering the collision range is unknown, one has to make a reasonable guess on it, and the maximum entropy distribution

steps out at this point. The maximum entropy guess for f_2, compatible with the univariate distribution appearing in the streaming term, is the factorized one, but on the contrary the guess for f_3, compatible with the f_2 appearing in the left-hand side, will be quite different from a factorized distribution.

We showed in chapter 2 that, given bivariate marginals, the maximum entropy estimate for $f_3(\zeta_1, \zeta_2, \zeta_3)$ was given by

$$f_3^{ME}(\zeta_1, \zeta_2, \zeta_3) = \frac{1}{Z} \exp\left(\lambda_1(\zeta_1, \zeta_2) + \lambda_2(\zeta_1, \zeta_3) + \lambda_3(\zeta_2, \zeta_3)\right) \tag{6.12}$$

for some functions λ_1, λ_2 and λ_3. While this result is of limited practical range generally speaking, particle distribution functions of statistical mechanics have the crucial peculiarity that they are *symmetric* under exchange of the particles. This implies that these marginals are the same for *each* pair, and accordingly all three λ's are actually *the same*. Absorbing the normalization, one is therefore allowed to write that

$$f_3^{ME}(\zeta_1, \zeta_2, \zeta_3) = G(\zeta_1, \zeta_2)G(\zeta_1, \zeta_3)G(\zeta_2, \zeta_3) \tag{6.13}$$

for a function G that is nevertheless still unknown, except for the fact that it has to satisfy the marginal constraint

$$G(\zeta_1, \zeta_2) \int d\zeta_3 G(\zeta_1, \zeta_3)G(\zeta_2, \zeta_3) = f_2(\zeta_1, \zeta_2). \tag{6.14}$$

It results from the previous consideration that before the collision the maximum entropy estimation of the three-particle distribution function is

$$f_3(\mathbf{x}_1, \mathbf{p}_1, \mathbf{x}_2, \mathbf{p}_2, \mathbf{x}_1, \mathbf{p}_3, t) = G(\mathbf{x}_1, \mathbf{p}_1; \mathbf{x}_2, \mathbf{p}_2; t)G(\mathbf{x}_1, \mathbf{p}_1; \mathbf{x}_1, \mathbf{p}_3; t)G(\mathbf{x}_2, \mathbf{p}_2; \mathbf{x}_1, \mathbf{p}_3; t). \tag{6.15}$$

The ansatz may be extended after the collision using the fact that, by Liouville's equation, $f_3(\mathbf{x}_1, \mathbf{p}_1, \mathbf{x}_2, \mathbf{p}_2, \mathbf{x}_1, \mathbf{p}_3, t) = f_3(\mathbf{x}_1^{-\tau}, \mathbf{p}_1, \mathbf{x}_2^{-\tau}, \mathbf{p}_2', \mathbf{x}_1^{-\tau}, \mathbf{p}_3', t - \tau)$, where τ is the retardation such that at $t - \tau$ the particles are entering the collision range with momenta $\mathbf{p}_1'$, $\mathbf{p}_3'$. Since $\mathbf{x}_i^{-\tau} \approx \mathbf{x}_i$ and $t \approx t - \tau$, and since $\mathbf{p}_1'$, $\mathbf{p}_3'$ are pre-collisional momenta, the ansatz may be introduced in the first integral as well with arguments $\mathbf{p}_1'$, $\mathbf{p}_3'$. We are therefore eventually led to the following Boltzmann-like form for the second-order BBGKY equation :

$$\begin{aligned}
&\frac{\partial f_2}{\partial t} + \frac{\mathbf{p}_1}{m}\frac{\partial f_2}{\partial \mathbf{x}_1} + \frac{\mathbf{p}_2}{m}\frac{\partial f_2}{\partial \mathbf{x}_2} \\
&= \int d\mathbf{p}_3 d\omega \frac{|\mathbf{p}_3 - \mathbf{p}_1|}{m} (G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1',\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_1}_{\mathbf{p}_1',\mathbf{p}_3'} G^{\mathbf{x}_2,\mathbf{x}_1}_{\mathbf{p}_2,\mathbf{p}_3'} - G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_1}_{\mathbf{p}_1,\mathbf{p}_3} G^{\mathbf{x}_2,\mathbf{x}_1}_{\mathbf{p}_2,\mathbf{p}_3}) \\
&\quad + \int d\mathbf{p}_4 d\omega \frac{|\mathbf{p}_4 - \mathbf{p}_2|}{m} (G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_2'} G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_4'} G^{\mathbf{x}_2,\mathbf{x}_2}_{\mathbf{p}_2',\mathbf{p}_4'} - G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_4} G^{\mathbf{x}_2,\mathbf{x}_2}_{\mathbf{p}_2,\mathbf{p}_4}),
\end{aligned} \tag{6.16}$$

where for the sake of readibility we have used the shortcut $G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_2} = G(\mathbf{x}_1,\mathbf{p}_1;\mathbf{x}_2,\mathbf{p}_2)$. The second term accounts for the contribution of collisions undergone by particle '2'.

Equation (6.16) is coherent for f_2 since G can - in principle - be solved in terms of f_2. This implicit form of the collision term bears a close ressemblance with the one appearing in the standard Boltzmann equation. This ressemblance might however turn deceptive since G is likely to be a complicated functional of f_2, but it happens that in spite of this mathematical complication we can push the analysis further following Boltzmann's footsteps.

6.4 Collisional invariants

Our immediate purpose is to deduce the equilibrium distribution resulting from (6.16). To this end we shall define invariants χ that are quantities conserved in a bilocal collision of pairs ('1','2') and ('3','4') (occurring in $\mathbf{x}$ or $\mathbf{y}$ for the (1,3) and (2,4) collision respectively). A collisional invariant has to be defined in this case such that [2]

$$\chi(\mathbf{p}'_1,\mathbf{p}'_2)+\chi(\mathbf{p}'_3,\mathbf{p}'_4)=\chi(\mathbf{p}_1,\mathbf{p}_2)+\chi(\mathbf{p}_3,\mathbf{p}_4). \tag{6.17}$$

Beside mass invariance, obvious invariants are

$$(\mathbf{p}'_1+\mathbf{p}'_2)+(\mathbf{p}'_3+\mathbf{p}'_4)=(\mathbf{p}_1+\mathbf{p}_2)+(\mathbf{p}_3+\mathbf{p}_4) \tag{6.18}$$

and

$$(\mathbf{p}'^2_1+\mathbf{p}'^2_2)+(\mathbf{p}'^2_3+\mathbf{p}'^2_4)=(\mathbf{p}^2_1+\mathbf{p}^2_2)+(\mathbf{p}^2_3+\mathbf{p}^2_4). \tag{6.19}$$

We assert that in our case these invariants should be complemented by

$$(\mathbf{p}'_1\cdot\mathbf{p}'_2)+(\mathbf{p}'_3\cdot\mathbf{p}'_4)=(\mathbf{p}_1\cdot\mathbf{p}_2)+(\mathbf{p}_3\cdot\mathbf{p}_4). \tag{6.20}$$

The rationale for introducing this extra invariant is that although collisions of external particles with both '1' and '2' have to be taken into consideration, these events do not, with overwhelming probability, occur simultaneously, and either '1' or '2' does not undergo collision and is thus left unaltered. In such "spurious bilocal collisions", (6.17) should therefore actually read

$$\chi(\mathbf{p}'_1,\mathbf{p}'_2)+\chi(\mathbf{p}'_3,\mathbf{p}'_2)=\chi(\mathbf{p}_1,\mathbf{p}_2)+\chi(\mathbf{p}_3,\mathbf{p}_2) \tag{6.21}$$

[2] We should also indicate spatial arguments, but in this context they are not relevant and can be dropped out.

or

$$\chi(\mathbf{p}_1', \mathbf{p}_2') + \chi(\mathbf{p}_1', \mathbf{p}_4') = \chi(\mathbf{p}_1, \mathbf{p}_2) + \chi(\mathbf{p}_1, \mathbf{p}_4) \tag{6.22}$$

for the first and second term on r.h.s. of (6.16) respectively. This interpretation amounts to introducing an extra particle so as to treat these two simple collisions as pair-pair collisions, which is harmless as far as the collision term is concerned since this fictitious particle is free. Then (6.20) becomes

$$\mathbf{p}_2 \cdot (\mathbf{p}_1' + \mathbf{p}_3') = \mathbf{p}_2 \cdot (\mathbf{p}_1 + \mathbf{p}_3)$$

(or

$$\mathbf{p}_1 \cdot (\mathbf{p}_2' + \mathbf{p}_4') = \mathbf{p}_1 \cdot (\mathbf{p}_2 + \mathbf{p}_4)$$

respectively) which is obviously true. As we shall see below, this "exotic" collisional invariant is necessary to enforce momentum correlation between particles at equilibrium.

6.5 Stationary state

To proceed towards the 2-particle distribution at equilibrium, let us first note that $G(\mathbf{x}_1, \mathbf{p}_1'; \mathbf{x}_1, \mathbf{p}_3') = G(\mathbf{x}_1, \mathbf{p}_1; \mathbf{x}_1, \mathbf{p}_3)$ and $G(\mathbf{x}_2, \mathbf{p}_2'; \mathbf{x}_2, \mathbf{p}_4') = G(\mathbf{x}_2, \mathbf{p}_2; \mathbf{x}_2, \mathbf{p}_4)$. This is certainly true of f_2 itself by Liouville's theorem, so we can reasonably suppose this passes to G; this can also be verified directly from (6.23) since for a binary collision we have that $\mathbf{p}_1' \cdot \mathbf{p}_3' = \mathbf{p}_1 \cdot \mathbf{p}_3$ (this relation follows from expressing post-collisional velocities in terms of the apsidal vector characterizing the event [48]).

Then the condition for the collision integrals to vanish is that

$$\begin{cases} G^{eq}(\mathbf{x}_1, \mathbf{p}_1'; \mathbf{x}_2, \mathbf{p}_2) G^{eq}(\mathbf{x}_2, \mathbf{p}_2; \mathbf{x}_1, \mathbf{p}_3') = G^{eq}(\mathbf{x}_1, \mathbf{p}_1; \mathbf{x}_2, \mathbf{p}_2) G^{eq}(\mathbf{x}_2, \mathbf{p}_2; \mathbf{x}_1, \mathbf{p}_3) \\ G^{eq}(\mathbf{x}_1, \mathbf{p}_1; \mathbf{x}_2, \mathbf{p}_2') G^{eq}(\mathbf{x}_1, \mathbf{p}_1; \mathbf{x}_2, \mathbf{p}_4') = G^{eq}(\mathbf{x}_1, \mathbf{p}_1; \mathbf{x}_2, \mathbf{p}_2) G^{eq}(\mathbf{x}_1, \mathbf{p}_1; \mathbf{x}_2, \mathbf{p}_4) \end{cases}$$

Taking the logarithm of both sides, we recognize in $\ln G^{eq}$ a collisional invariant as defined by (6.21), (6.22), so that this quantity is necessarily a linear combination of the invariants introduced above, *i.e.*

$$G^{eq}(\mathbf{p}_1; \mathbf{p}_2) = e^{A + \mathbf{B} \cdot (\mathbf{p}_1 + \mathbf{p}_2) + C(\mathbf{p}_1^2 + \mathbf{p}_2^2) + D \mathbf{p}_1 \cdot \mathbf{p}_2}. \tag{6.23}$$

This form is actually not the most general since the coefficients could possibly depend on the space variables. While letting A, $\mathbf{B}$ and C depend on $\mathbf{x}, \mathbf{y}$ might seem a jesuitic subtlety, this dependence would be relevant for the coefficient in front of $\mathbf{p}_1 \cdot \mathbf{p}_2$. This

however introduces additional complications when connecting coefficients to observable quantities. We thus focus on the case of constant coefficients for the moment and postpone the rest of the discussion to a later section.

From this expression for G^{eq} we can now deduce the expression for the 3-particle distribution f_3^{eq} that makes the collision term vanish. We get (substituting $\mathbf{p}_3 \to \mathbf{k}$ in order to single out the integration variable)

$$\begin{aligned} f_3^{eq}(\mathbf{x}_1, \mathbf{p}_1; \mathbf{x}_2, \mathbf{p}_2; \mathbf{x}_1, \mathbf{k}) &= G^{eq}(\mathbf{p}_1; \mathbf{p}_2) G^{eq}(\mathbf{p}_1; \mathbf{k}) G^{eq}(\mathbf{p}_2; \mathbf{k}) \\ &= e^{A+\mathbf{B}(\mathbf{p}_1+\mathbf{p}_2)+C(\mathbf{p}_1^2+\mathbf{p}_2^2)+D\mathbf{p}_1\mathbf{p}_2} \\ &\quad \times e^{A+\mathbf{B}(\mathbf{p}_1+\mathbf{k})+C(\mathbf{p}_1^2+\mathbf{k}^2)+D\mathbf{p}_1\mathbf{k}} \\ &\quad \times e^{A+\mathbf{B}(\mathbf{p}_2+\mathbf{k})+C(\mathbf{p}_2^2+\mathbf{k}^2)+D\mathbf{p}_2\mathbf{k}} \\ &= e^{3A+2\mathbf{B}(\mathbf{p}_1+\mathbf{p}_2+\mathbf{k})+2C(\mathbf{p}_1^2+\mathbf{p}_2^2+\mathbf{k}^2)+D(\mathbf{p}_1\mathbf{p}_2+\mathbf{p}_2\mathbf{k}+\mathbf{p}_1\mathbf{k})}. \end{aligned} \tag{6.24}$$

It remains to integrate on $\mathbf{k}$ in order to obtain f_2^{eq}:

$$\begin{aligned} f_2^{eq}(\mathbf{x}_1, \mathbf{p}_1; \mathbf{x}_2, \mathbf{p}_2) &= e^{3A+\mathbf{B}(\mathbf{p}_1+\mathbf{p}_2)+2C(\mathbf{p}_1^2+\mathbf{p}_2^2)+D\mathbf{p}_1\mathbf{p}_2} \int d\mathbf{k} e^{(2\mathbf{B}+D\mathbf{p}_1+D\mathbf{p}_2)\mathbf{k}+2C\mathbf{k}^2} \\ &= \left(-\frac{\pi}{2C}\right)^{3/2} e^{3A+2\mathbf{B}(\mathbf{p}_1+\mathbf{p}_2)+2C(\mathbf{p}_1^2+\mathbf{p}_2^2)+D\mathbf{p}_1\mathbf{p}_2-\frac{(2\mathbf{B}+D\mathbf{p}_1+D\mathbf{p}_2)^2}{8C}} \\ &= \left(-\frac{\pi}{2C}\right)^{3/2} e^{3A-\frac{\mathbf{B}^2}{2C}} e^{\left(2-\frac{D}{2C}\right)\mathbf{B}(\mathbf{p}_1+\mathbf{p}_2)+\left(2C-\frac{D^2}{8C}\right)(\mathbf{p}_1^2+\mathbf{p}_2^2)+\left(D-\frac{D^2}{4C}\right)\mathbf{p}_1\mathbf{p}_2}. \end{aligned} \tag{6.25}$$

The coefficients have now to be determined so as to match observational constraints on average momentum, average kinetic energy ϵ and momentum correlation φ. Since the average momentum is proportional to $\mathbf{B}$, we can restrict ourselves for simplicity to the case of a gas without global translational motion and put $\mathbf{B} = 0$ so that

$$f_2^{eq}(\mathbf{p}_1; \mathbf{p}_2) = \left(-\frac{\pi}{2C}\right)^{3/2} e^{3A} e^{\left(2C-\frac{D^2}{8C}\right)(\mathbf{p}_1^2+\mathbf{p}_2^2)+\left(D-\frac{D^2}{4C}\right)\mathbf{p}_1\mathbf{p}_2}. \tag{6.26}$$

Regarding average energy we should have (keeping in mind that f_2 is normalized, by convention, to $N(N-1) \approx N^2$)

$$\epsilon = \frac{\int d\mathbf{p}_1 d\mathbf{p}_2 \left(\frac{\mathbf{p}_i^2}{2m}\right) f_2^{eq}}{\int d\mathbf{p}_1 d\mathbf{p}_2 f_2^{eq}} = \frac{V^2}{2mN^2} \langle \mathbf{p}_i^2 \rangle. \tag{6.27}$$

Regarding the correlation coefficient of momenta, it follows from our assumption of a gaz without global translational motion that

$$\varphi = \frac{\langle \mathbf{p}_1 \cdot \mathbf{p}_2 \rangle - \langle \mathbf{p}_1 \rangle \langle \mathbf{p}_2 \rangle}{\sqrt{\langle \mathbf{p}_1^2 \rangle - \langle \mathbf{p}_1 \rangle^2} \sqrt{\langle \mathbf{p}_2^2 \rangle - \langle \mathbf{p}_2 \rangle^2}} = \frac{\langle \mathbf{p}_1 \cdot \mathbf{p}_2 \rangle}{\langle \mathbf{p}^2 \rangle}. \tag{6.28}$$

where

$$\langle O \rangle = \int d\mathbf{p}_1 d\mathbf{p}_2 O f_2^{eq}. \tag{6.29}$$

Performing integrations on $\mathbf{p}_1, \mathbf{p}_2$ therefore allows relating φ to C and D as

$$\varphi = \frac{D^2 - 4CD}{16C^2 - D^2} \tag{6.30}$$

so that $D = -4C\varphi/(1+\varphi)$. Using the expression for ϵ, the coefficient C can then be related to φ as

$$C = \frac{3(\varphi+1)}{8m\epsilon(2\varphi+1)(\varphi-1)}. \tag{6.31}$$

Normalizing we finally arrive at the conclusion that

$$f_2^{eq}(\mathbf{p}_1; \mathbf{p}_2) = \left(\frac{N}{V}\right)^2 \left(\frac{3m}{4\pi\epsilon}\right)^3 (1-\varphi^2)^{-3/2} \exp\left(-\frac{3}{4m\epsilon(1-\varphi^2)}\left(\mathbf{p}_1^2 + \mathbf{p}_2^2\right) + \frac{3\varphi}{2m\epsilon(1-\varphi^2)}\mathbf{p}_1 \cdot \mathbf{p}_2\right). \tag{6.32}$$

It therefore appears that equilibrium distributions, such as (6.32), can -at least in principle- be found for which particles stay correlated with each other over time. This correlation is nonetheless hidden as long as 1-particle distributions only are examined, since integrating on $\mathbf{p}_2$ we find

$$f_1^{eq}(\mathbf{p}_1) = \frac{N}{V}\left(\frac{3m}{4\pi\epsilon}\right)^{3/2} e^{-\frac{3}{4m\epsilon}\mathbf{p}_1^2} \tag{6.33}$$

which is none but Maxwell's usual distribution of velocities.

6.6 Momentum correlation depending on inter-particle distance

We alluded above to the fact that the coefficients in front of collisional invariants in (6.23) can depend on space coordinates. Though this possibility comes as an irrelevant complication for what regards A, $\mathbf{B}$ and C, we want to retain the possibility to consider

a momentum correlation depending on the respective positions of particles so that D becomes $D(\mathbf{x}_1, \mathbf{x}_2)$.

The distribution f_3 can be obtained proceeding as previously, except that it is necessary to symmetrize the argument in the exponential with respect to events occurring in $\mathbf{x}_1$ and $\mathbf{x}_2$. Accordingly (6.24) becomes

$$f_3^{eq}(\mathbf{x}_1, \mathbf{p}_1; \mathbf{x}_2, \mathbf{p}_2; \mathbf{k}) = e^{3A+2\mathbf{B}(\mathbf{p}_1+\mathbf{p}_2+\mathbf{k})+2C(\mathbf{p}_1^2+\mathbf{p}_2^2+\mathbf{k}^2)} \cdot e^{D(\mathbf{x}_1,\mathbf{x}_2)(\mathbf{p}_1\mathbf{p}_2+\frac{1}{2}\mathbf{p}_1\mathbf{k}+\frac{1}{2}\mathbf{p}_2\mathbf{k})+\frac{1}{2}D(\mathbf{x}_1,\mathbf{x}_1)\mathbf{p}_1\mathbf{k}+\frac{1}{2}D(\mathbf{x}_2,\mathbf{x}_2)\mathbf{p}_2\mathbf{k}}. \quad (6.34)$$

We can actually convoke an argument of isotropy to support the assumption that the correlation is likely to depend only on the distance r between particles but not on their positions themselves, that is $D(\mathbf{x}_1, \mathbf{x}_2)$ becomes $D(|\mathbf{x}_1 - \mathbf{x}_2|) = D(r)$. Equation (6.34) can be simplified to yield

$$f_3^{eq}(\mathbf{x}_1, \mathbf{p}_1; \mathbf{x}_2, \mathbf{p}_2; \mathbf{k}) = e^{3A+2\mathbf{B}(\mathbf{p}_1+\mathbf{p}_2+\mathbf{k})+2C(\mathbf{p}_1^2+\mathbf{p}_2^2+\mathbf{k}^2)+D(r)(\mathbf{p}_1\mathbf{p}_2+\frac{1}{2}(\mathbf{p}_1+\mathbf{p}_2)\mathbf{k})+\frac{1}{2}D(0)(\mathbf{p}_1+\mathbf{p}_2)\mathbf{k}}. \quad (6.35)$$

Integrating on $\mathbf{k}$ as previously we find for f_2^{eq} that

$$f_2^{eq}(\mathbf{x}_1, \mathbf{p}_1; \mathbf{x}_2, \mathbf{p}_2) = \left(-\frac{\pi}{2C}\right)^{3/2} e^{3A-\frac{\mathbf{B}^2}{2C}} e^{\left(2-\frac{D(r)+D(0)}{4C}\right)\mathbf{B}(\mathbf{p}_1+\mathbf{p}_2)} \cdot e^{\left(2C-\frac{(D(r)+D(0))^2}{32C}\right)(\mathbf{p}_1^2+\mathbf{p}_2^2)+\left(D(r)-\frac{(D(r)+D(0))^2}{16C}\right)\mathbf{p}_1\mathbf{p}_2}. \quad (6.36)$$

Without global translational motion, we put $\mathbf{B} = 0$ so that (6.25) is modified into (with the arguments reduced to their minimal form – denoting the probability to find two particles at distance r apart, with momenta $\mathbf{p}_1$ and $\mathbf{p}_2$)

$$f_2^{eq}(\mathbf{p}_1; \mathbf{p}_2; r) = \left(-\frac{\pi}{2C}\right)^{3/2} e^{3A+\left(2C-\frac{(D(r)+D(0))^2}{32C}\right)(\mathbf{p}_1^2+\mathbf{p}_2^2)+\left(D(r)-\frac{(D(r)+D(0))^2}{16C}\right)\mathbf{p}_1\mathbf{p}_2}. \quad (6.37)$$

In establishing the connection of (6.37) with observable average kinetic energy, it is required to integrate (6.37) over r; it is therefore mandatory to introduce an *ansatz* or experimentally-educated guess on the form of $D(r)$, which sadly falls outside the scope of our theoretical investigations.

6.7 Balance equations

Allowing for momentum correlation between particles has a two-fold consequence as to the macroscopic description of the fluid. First, it makes necessary to take account of this

correlation in the usual balance equations for mass, momentum and energy; second these have to be complemented by a fourth balance equation corresponding to the collisional invariant (6.20).

The macroscopic conservation equations are derived from (6.16) by multiplying both sides by $\chi(\mathbf{p}_1, \mathbf{p}_2)$ and integrating on momenta, so as to get

$$\begin{aligned}
&\int d\mathbf{p}_1 d\mathbf{p}_2 \chi(\mathbf{p}_1, \mathbf{p}_2) \left(\frac{\partial}{\partial t} + \frac{\mathbf{p}_1}{m} \frac{\partial}{\partial \mathbf{x}} + \frac{\mathbf{p}_2}{m} \frac{\partial}{\partial \mathbf{y}} \right) f_2(\mathbf{x}, \mathbf{p}_1; \mathbf{y}, \mathbf{p}_2; t) \\
&= \int d\mathbf{p}_1 d\mathbf{p}_2 d\mathbf{p}_3 d\omega \chi(\mathbf{p}_1, \mathbf{p}_2) \frac{|\mathbf{p}_3 - \mathbf{p}_1|}{m} (G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}'_1,\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_1}_{\mathbf{p}'_1,\mathbf{p}'_3} G^{\mathbf{x}_2,\mathbf{x}_1}_{\mathbf{p}_2,\mathbf{p}'_3} - G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_1}_{\mathbf{p}_1,\mathbf{p}_3} G^{\mathbf{x}_2,\mathbf{x}_1}_{\mathbf{p}_2,\mathbf{p}_3}) \\
&\quad + \int d\mathbf{p}_1 d\mathbf{p}_2 d\mathbf{p}_4 d\omega \chi(\mathbf{p}_1, \mathbf{p}_2) \frac{|\mathbf{p}_4 - \mathbf{p}_2|}{m} (G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}'_2} G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}'_4} G^{\mathbf{x}_2,\mathbf{x}_2}_{\mathbf{p}'_2,\mathbf{p}'_4} - G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_4} G^{\mathbf{x}_2,\mathbf{x}_2}_{\mathbf{p}_2,\mathbf{p}_4}).
\end{aligned} \tag{6.38}$$

We now have to go through the usual sequence of relabelling and permutations [61]. Substituting $\mathbf{p}_1 \leftrightarrow \mathbf{p}_3$ in the first integral of the collision term and $\mathbf{p}_2 \leftrightarrow \mathbf{p}_4$ in the second (which is allowed both being integrated variables) allows us to rewrite the r.h.s. of (6.38) as

$$\begin{aligned}
&\int d\mathbf{p}_3 d\mathbf{p}_2 d\mathbf{p}_1 d\omega \chi(\mathbf{p}_3, \mathbf{p}_2) \frac{|\mathbf{p}_1 - \mathbf{p}_3|}{m} (G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}'_3,\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_1}_{\mathbf{p}'_3,\mathbf{p}'_1} G^{\mathbf{x}_2,\mathbf{x}_1}_{\mathbf{p}_2,\mathbf{p}'_1} - G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_3,\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_1}_{\mathbf{p}_3,\mathbf{p}_1} G^{\mathbf{x}_2,\mathbf{x}_1}_{\mathbf{p}_2,\mathbf{p}_1}) \\
&\quad + \int d\mathbf{p}_1 d\mathbf{p}_4 d\mathbf{p}_2 d\omega \chi(\mathbf{p}_1, \mathbf{p}_4) \frac{|\mathbf{p}_2 - \mathbf{p}_4|}{m} (G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}'_4} G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}'_2} G^{\mathbf{x}_2,\mathbf{x}_2}_{\mathbf{p}'_4,\mathbf{p}'_2} - G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_4} G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_2} G^{\mathbf{x}_2,\mathbf{x}_2}_{\mathbf{p}_4,\mathbf{p}_2}).
\end{aligned} \tag{6.39}$$

Adding (6.38) and (6.39) and dividing the result, the collision term becomes

$$\begin{aligned}
&\frac{1}{2} \int d\mathbf{p}_1 d\mathbf{p}_2 d\mathbf{p}_3 d\omega \left[\chi(\mathbf{p}_1, \mathbf{p}_2) + \chi(\mathbf{p}_3, \mathbf{p}_2) \right] \frac{|\mathbf{p}_3 - \mathbf{p}_1|}{m} \\
&\qquad \cdot (G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}'_1,\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_1}_{\mathbf{p}'_1,\mathbf{p}'_3} G^{\mathbf{x}_2,\mathbf{x}_1}_{\mathbf{p}_2,\mathbf{p}'_3} - G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_1}_{\mathbf{p}_1,\mathbf{p}_3} G^{\mathbf{x}_2,\mathbf{x}_1}_{\mathbf{p}_2,\mathbf{p}_3}) \\
&\quad + \frac{1}{2} \int d\mathbf{p}_1 d\mathbf{p}_2 d\mathbf{p}_4 d\omega \left[\chi(\mathbf{p}_1, \mathbf{p}_2) + \chi(\mathbf{p}_1, \mathbf{p}_4) \right] \frac{|\mathbf{p}_4 - \mathbf{p}_2|}{m} \\
&\qquad \cdot (G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}'_2} G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}'_4} G^{\mathbf{x}_2,\mathbf{x}_2}_{\mathbf{p}'_2,\mathbf{p}'_4} - G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_4} G^{\mathbf{x}_2,\mathbf{x}_2}_{\mathbf{p}_2,\mathbf{p}_4}).
\end{aligned} \tag{6.40}$$

Now relabelling $\mathbf{p}_1 \leftrightarrow \mathbf{p}'_1$, $\mathbf{p}_2 \leftrightarrow \mathbf{p}'_2$, $\mathbf{p}_3 \leftrightarrow \mathbf{p}'_3$ and $\mathbf{p}_4 \leftrightarrow \mathbf{p}'_4$, expression (6.40) becomes

$$\begin{aligned}
&\frac{1}{2} \int d\mathbf{p}'_1 d\mathbf{p}'_2 d\mathbf{p}'_3 d\omega \left[\chi(\mathbf{p}'_1, \mathbf{p}'_2) + \chi(\mathbf{p}'_3, \mathbf{p}'_2) \right] \frac{|\mathbf{p}'_3 - \mathbf{p}'_1|}{m} \\
&\qquad \cdot (G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}'_2} G^{\mathbf{x}_1,\mathbf{x}_1}_{\mathbf{p}_1,\mathbf{p}_3} G^{\mathbf{x}_2,\mathbf{x}_1}_{\mathbf{p}'_2,\mathbf{p}_3} - G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}'_1,\mathbf{p}'_2} G^{\mathbf{x}_1,\mathbf{x}_1}_{\mathbf{p}'_1,\mathbf{p}'_3} G^{\mathbf{x}_2,\mathbf{x}_1}_{\mathbf{p}'_2,\mathbf{p}'_3}) \\
&\quad + \frac{1}{2} \int d\mathbf{p}'_1 d\mathbf{p}'_2 d\mathbf{p}'_4 d\omega \left[\chi(\mathbf{p}'_1, \mathbf{p}'_2) + \chi(\mathbf{p}'_1, \mathbf{p}'_4) \right] \frac{|\mathbf{p}'_4 - \mathbf{p}'_2|}{m} \\
&\qquad \cdot (G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}'_1,\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}'_1,\mathbf{p}_4} G^{\mathbf{x}_2,\mathbf{x}_2}_{\mathbf{p}_2,\mathbf{p}_4} - G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}'_1,\mathbf{p}'_2} G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}'_1,\mathbf{p}'_4} G^{\mathbf{x}_2,\mathbf{x}_2}_{\mathbf{p}'_2,\mathbf{p}'_4}).
\end{aligned} \tag{6.41}$$

Using the fact that $d\mathbf{p}_1' d\mathbf{p}_2' d\mathbf{p}_3' = d\mathbf{p}_1 d\mathbf{p}_2 d\mathbf{p}_3$ and $d\mathbf{p}_1' d\mathbf{p}_2' d\mathbf{p}_4' = d\mathbf{p}_1 d\mathbf{p}_2 d\mathbf{p}_4$, and that $|\mathbf{p}_3' - \mathbf{p}_1'| = |\mathbf{p}_3 - \mathbf{p}_1|$ and $|\mathbf{p}_4' - \mathbf{p}_2'| = |\mathbf{p}_4 - \mathbf{p}_2|$, this last expression (6.41) can be added to (6.40) and the result, divided by two, yields for the collision term

$$\begin{aligned}
&\frac{1}{4}\int d\mathbf{p}_1 d\mathbf{p}_2 d\mathbf{p}_3 d\omega \left[\chi(\mathbf{p}_1,\mathbf{p}_2) + \chi(\mathbf{p}_3,\mathbf{p}_2) - \chi(\mathbf{p}_1',\mathbf{p}_2') - \chi(\mathbf{p}_3',\mathbf{p}_2')\right] \frac{|\mathbf{p}_3 - \mathbf{p}_1|}{m} \\
&\qquad \cdot \left(G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1',\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_1}_{\mathbf{p}_1',\mathbf{p}_3'} G^{\mathbf{x}_2,\mathbf{x}_1}_{\mathbf{p}_2,\mathbf{p}_3'} - G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_1}_{\mathbf{p}_1,\mathbf{p}_3} G^{\mathbf{x}_2,\mathbf{x}_1}_{\mathbf{p}_2,\mathbf{p}_3}\right) \\
&+ \frac{1}{4}\int d\mathbf{p}_1 d\mathbf{p}_2 d\mathbf{p}_4 d\omega \left[\chi(\mathbf{p}_1,\mathbf{p}_2) + \chi(\mathbf{p}_1,\mathbf{p}_4) - \chi(\mathbf{p}_1',\mathbf{p}_2') - \chi(\mathbf{p}_1',\mathbf{p}_4')\right] \frac{|\mathbf{p}_4 - \mathbf{p}_2|}{m} \\
&\qquad \cdot \left(G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_2'} G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_4'} G^{\mathbf{x}_2,\mathbf{x}_2}_{\mathbf{p}_2',\mathbf{p}_4'} - G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_2} G^{\mathbf{x}_1,\mathbf{x}_2}_{\mathbf{p}_1,\mathbf{p}_4} G^{\mathbf{x}_2,\mathbf{x}_2}_{\mathbf{p}_2,\mathbf{p}_4}\right). \qquad (6.42)
\end{aligned}$$

which vanishes by the definition of collisional invariants (6.21), (6.22). It result that (6.38) reduces to

$$\int d\mathbf{p}_1 d\mathbf{p}_2 \chi(\mathbf{p}_1,\mathbf{p}_2) \left(\frac{\partial}{\partial t} + \frac{\mathbf{p}_1}{m}\frac{\partial}{\partial \mathbf{x}} + \frac{\mathbf{p}_2}{m}\frac{\partial}{\partial \mathbf{y}}\right) f_2(\mathbf{x},\mathbf{p}_1;\mathbf{y},\mathbf{p}_2;t) = 0. \qquad (6.43)$$

This general balance equation can therefore be rewritten as

$$\begin{aligned}
0 = &\frac{\partial}{\partial t}\int d\mathbf{p}_1 d\mathbf{p}_2 \chi f_2 - \int d\mathbf{p}_1 d\mathbf{p}_2 f_2 \frac{\partial \chi}{\partial t} \\
&+ \frac{1}{m}\frac{\partial}{\partial \mathbf{x}}\int d\mathbf{p}_1 d\mathbf{p}_2 \chi \mathbf{p}_1 f_2 - \frac{1}{m}\int d\mathbf{p}_1 d\mathbf{p}_2 \mathbf{p}_1 f_2 \frac{\partial \chi}{\partial \mathbf{x}} \\
&+ \frac{1}{m}\frac{\partial}{\partial \mathbf{y}}\int d\mathbf{p}_1 d\mathbf{p}_2 \chi \mathbf{p}_2 f_2 - \frac{1}{m}\int d\mathbf{p}_1 d\mathbf{p}_2 \mathbf{p}_2 f_2 \frac{\partial \chi}{\partial \mathbf{y}}. \qquad (6.44)
\end{aligned}$$

For $\chi = 1$ the general expression (6.44) becomes

$$\frac{\partial}{\partial t}\int d\mathbf{p}_1 d\mathbf{p}_2 f_2 + \frac{1}{m}\frac{\partial}{\partial \mathbf{x}}\int d\mathbf{p}_1 d\mathbf{p}_2 \mathbf{p}_1 f_2 + \frac{1}{m}\frac{\partial}{\partial \mathbf{y}}\int d\mathbf{p}_1 d\mathbf{p}_2 \mathbf{p}_2 f_2 = 0, \qquad (6.45)$$

for $\chi = p_1^j + p_2^j$ it becomes

$$\begin{aligned}
&\frac{\partial}{\partial t}\int d\mathbf{p}_1 d\mathbf{p}_2 (p_1^j + p_2^j) f_2 + \frac{1}{m}\frac{\partial}{\partial \mathbf{x}}\int d\mathbf{p}_1 d\mathbf{p}_2 (p_1^j + p_2^j)\mathbf{p}_1 f_2 \\
&\qquad + \frac{1}{m}\frac{\partial}{\partial \mathbf{y}}\int d\mathbf{p}_1 d\mathbf{p}_2 (p_1^j + p_2^j)\mathbf{p}_2 f_2 = 0, \qquad (6.46)
\end{aligned}$$

for $\chi = \mathbf{p}_1^2 + \mathbf{p}_2^2$ we have

$$\frac{\partial}{\partial t}\int d\mathbf{p}_1 d\mathbf{p}_2(\mathbf{p}_1^2+\mathbf{p}_2^2)f_2+\frac{1}{m}\frac{\partial}{\partial \mathbf{x}}\int d\mathbf{p}_1 d\mathbf{p}_2(\mathbf{p}_1^2+\mathbf{p}_2^2)\mathbf{p}_1 f_2 + \frac{1}{m}\frac{\partial}{\partial \mathbf{y}}\int d\mathbf{p}_1 d\mathbf{p}_2(\mathbf{p}_1^2+\mathbf{p}_2^2)\mathbf{p}_2 f_2 = 0 \tag{6.47}$$

and at last $\chi = \mathbf{p}_1 \cdot \mathbf{p}_2$ yields

$$\frac{\partial}{\partial t}\int d\mathbf{p}_1 d\mathbf{p}_2(\mathbf{p}_1\cdot\mathbf{p}_2)f_2+\frac{1}{m}\frac{\partial}{\partial \mathbf{x}}\int d\mathbf{p}_1 d\mathbf{p}_2(\mathbf{p}_1\cdot\mathbf{p}_2)\mathbf{p}_1 f_2 + \frac{1}{m}\frac{\partial}{\partial \mathbf{y}}\int d\mathbf{p}_1 d\mathbf{p}_2(\mathbf{p}_1\cdot\mathbf{p}_2)\mathbf{p}_2 f_2 = 0. \tag{6.48}$$

When reexpressed in terms of macroscopic quantities, (6.45) takes the form of two independent copies of the usual local conservation equation for mass density (which is expected since the correlation between particles is about their momenta), while (6.46) and (6.47) represent two coupled copies of the corresponding local equations. Equation (6.48) alone contains the physics brought in by considering 2-particle distributions.

6.8 Comments

It appears from our analysis that applying the criterion of maximum entropy as a heuristic tool to infer the 3-particle distribution based on a requirement of compatibility with 2-particle marginals allows to set up a consistent equation for the dynamics of pairs of particles, with the consequences that equilibrium states potentially exhibit correlation and that a conservation equation exists besides conservation of mass, momentum and energy. Two remarks are in order here.

Since the criterion of maxium entropy relies on a subjective ingredient (*i.e.* the physicist's uncertainty about the exact 3-particle distribution), the reader might feel uncomfortable using this approach in this context. Moreover, the maximum entropy approach to statistical mechanics is often considered with circumspection since this subjective ingredient provides an alternative derivation of the supposedly exact and objective Maxwellian distribution. It should be reminded that we have seen in chapter 3 that Wallis' argument provides an objective rationale supporting the principle of maximum entropy, namely that the distribution having the largest entropy is also the most probable one in the absence of an *a priori*. Postulating the maximum entropy estimate of f_3 thus amounts to replacing the actual f_3 by the most probable distribution compatible with f_2. In this respect the maximum entropy approach is not subjective properly said. Moreover, our result (6.32) makes clear that Maxwell's distribution is but the result of our relative lack of interest in dealing with more than one particle at a

time, and provides only a first-order approximation. In this respect it is not particularly objective.

The equilibrium 2-particle distribution might be expected to follow the standard canonical distribution, in which case the correlating term in equation (6.32) would appear to come in contradiction with standard results of statistical mechanics, but this is not the case for two reasons. Would the canonical distribution be supposed to apply to the system as a whole, it should be reminded that the canonical ensemble is concerned with systems immersed in a heat bath; hence it would be a pointless assumption to regard the isolated N-particle system as canonically distributed. More importantly, it must be underlined that the conservation of the N-particle entropy (6.2) comes in contradiction with the microcanonical postulate, *i.e.* the assumption that systems are distributed equiprobably on the shell of constant energy.[3] Since this postulated equiprobable distribution f_N^* is the (only) one maximizing $H(f_N)$, then a kinetic description becomes essentially pointless given that $f_N(t) = f_N^*$. On the other side, would the canonical distribution be justified using the combinatorial argument of the most probable distribution, it should be reminded that the derivation relies on the assumption that individual particles can be distributed independently over the μ-space. This is no longer the case when dealing with pairs of particles, for assigning a pair ('a','b') to a point of the (bilocal) μ-space puts constraints on all pairs that involve either 'a' or 'b' (in other words pairs do not obey Boltzmann's statistics).

[3]Let us recall here that there exists other circumstances where a direct use of statistical ensembles raises issues. This is for instance the case for long-range interacting systems, where in particular the equivalence of microcanonical and canonical ensembles is not guaranteed [79, 21]

Perspectives and Epilogue

Notwithstanding the preliminary chapter devoted to the generalized Minority Game, whose aim was to investigate how information - taken in its most intuitive sense - could impact the global behaviour of a relatively simple non-equilibrium system, the primary task we assigned ourselves in this thesis was to assess the relevance of information theory in the study of complex systems, and in particular to highlight the versatility of maximum entropy criteria by examining three domains of application that *a priori* are not particularly tightly related to each other.

Regarding reconstruction of time series in chapter 4 and description of complexity in chapter 5, our focus was to take a critical look at existing theoretical tools based on the principle of maximum entropy in order to see to what extent these approaches would turn relevant when brought at work on actual problems. It was unavoidable that in testing the frontiers of a method we eventually went a bit too far and confronted situtions in which this method fails or, at least, produces disappointing outcomes.

Looking retrospectively at the material presented in chapter 4 at the light of the work undertaken since, our results there were actually rather promising in spite of the computational difficulties as well as our discutable assumption of reversibility. Since applying our approach on empirical financial series turned out to give surprisingly relevant results, the temptation is great of pushing the experiment further in this direction. This requires however to solve some computational issues, possibly with the help of the competing but complementary approach based on the transfer matrix (which raises comparable numerical issues but has the great advantage of overcoming the restriction to reversible series). In the case when the state space can be expressed as the space of configurations of some set of variables, a point that might also be investigated is the possibility to introduce marginal constraints on the transition matrix instead of constraints on autocorrelation, by imposing transition probabilities for subsets of variables.

The material presented in the first part of chapter 5 came early in our research on information theory and actually constituted our entry point in maximum entropy models. In our view it has the merit of highlighting, among others, that the decomposition of the total information cannot be interpreted as handily as it is sometimes suggested. Our choice of working with cellular automata was not necessarily optimal since confronting a notion of complexity that cannot receive an immediate interpretation with another (Wolfram's) that can be grasped at first glance is bound to result in a certain defiance towards the former (though possibly unjustified). In any way this approach to complexity would benefit greatly from theoretical advances regarding the computation of coefficients that enter the information decomposition, but this is made difficult

by the fact that the maximum entropy distributions based on constrained marginals cannot so far be computed analytically (it actually came to us as a surprise that multivariate functions taking a pseudo-factorized form such as (3.9) or even (6.13) seemed to have received very little attention from mathematicians). In this respect the local approach discussed in section 5.3 seems more convenient to work with both technically and conceptually, and seems more promising as to the results obtained so far.

We showed in chapter 6 that maximum entropy models based on marginals can find an utility nonetheless, for it seems that most steps of local kinetic theory can be replicated in the bilocal setting introduced there. In our view the crucial physical fact resulting from our analysis, more than in the existence of correlated equilibrium states, lies in the fourth collisional invariant and the associated conservation equation (6.48). The main difficulty when dealing with bilocal objects whose local counterparts are familiar resides in the concomitant loss in intuition. We cannot but hope that this has not plagued our calculations too much; it must be stressed however that until the ideas presented here are supported or infirmed experimentally or numerically they must be considered as no more than a theoretical construction. Likely the most convenient way to tackle this validation is through a discrete numerical model, that must however display some kind of hamiltonianity. On the experimental side, a full hydrodynamical treatment of the balance equations has still to be carried through in order to produce testable predictions.

* * *

Since this thesis was concerned to a large extent with the question of assessing maximum entropy models, we would like to conclude it by mentioning a recent work by Obuchi & al. [77, 78] where this question is tackled from a very different perspective that sheds another light on the line of thought presented here.

These authors consider the problem of estimating a target distribution over a set of N variables taking binary values $\{\pm 1\}$, so that each distribution can be seen as a point in the 2^N-dimesional simplex. Then they choose *randomly* M observables among the set of polynomial functions of the variables and compute their average values over the target distribution. These M expectation values provide the constraints on the problem, and admissible distributions are defined as the ones satisfying these constraints within some fixed accuracy. Their next step is to introduce a biased measure over the admissible space, expressed in terms of a biasing parameter Γ such that the distribution with the largest entropy is selected when $\Gamma \to \infty$ while when $\Gamma = 0$ the measure is uniform.

Once the admissible solution space and its measure are introduced, the next step is to compute the distance R between the target distribution and the center-of-mass of the solution space varies with Γ. The main findings reported in [77] are that

- expectedly, R decreases with M
- surprisingly, R does not depend on the bias, so that in this setup the maximum entropy distribution is not actually better than any other admissible solution

- however when the tolerance on the accuracy of the solutions becomes too large R grows with Γ; this is expected since when the constraints get loose they become irrelevant, and the maximum entropy solution tends to go to the uniform distribution .

The obvious drawback of the above analysis lies in the fact that usual target distributions of interest are smooth and in the random selection of observables, implying that these observables are not chosen according to their *a priori* relevance. Though overcoming these drawbacks leads to considerable mathematical difficulties, preliminary results [78] suggest that when considering smooth distributions and adequately selected observables the maximum entropy admissible distribution steps out. Nevertheless, the quality of this solution seems to be highly sensitive to the informativeness of the observables.

Appendix: Decomposition of total information

We give here an implementation of the iterative scaling algorithm leading to the decomposition of total information in terms of order of interaction. The present implementation in Python aims less at performance than at self-completeness. As much as possible we avoid using built-in functions, except for the function returning the power set of an ensemble.

```
# Imports ************************************************************

from math import *
from numpy import *
from random import *
import itertools as it
import functools as fun
from pylab import plot, show, legend


# Create the list of agents ******************************************

NA = 5

ListA = []
for k in range(NA):
   ListA.append(k)

NS = 2**NA     # Number of possible states


# Create the topology of the network that agents are put on. ************

# Are implemented here a random (Erdös-Rényi) network with any two nodes
linked with probability p (here each node is linked to itself), and a
circular topology where each node is linked to its right and left
neighbours.
```

```
p = 0.5

M = zeros([NA, NA], float)

# Random graph

for l in range(NA):
   for c in range(NA):
      if l == c:
         M[l, c] = 1
      if l < c:
         if random() < p:
            M[l, c] = 1
      else:
         M[l, c] = M[c, l]


# Circular topology

for l in range(NA):
   for c in range(NA):
      if abs(c-l)%(NA-1) <= 1:
         M[l, c] = 1


# Generate the dynamics of the model  ******************************

# For the sake of variety we implement here instead of an ECA
an Ising-like dynamics and a voter model in which an agent takes
value 1 with probability 0.8*Q+0.1 where Q denotes the average
opinion of its neighbourhood.


# "Neighbourhood" returns the list of nearest neighbours of a node given
an adjacency matrix.

def Neighbourhood(agent, MatAdj):
   N = []
   for k in range(NA):
      if MatAdj[agent, k] == 1:
         N.append(k)
   return N
```

```
# VOTER MODEL

# "Will_be_one" returns the probability for an agent to take value 1
given the current state of the system.

def Will_be_one(agent, state, MatAdj):
   neigh = Neighbourhood(agent, MatAdj)
   b_state = list(map(int, format(state, "0" + str(NA) + "b")))
   sum_neigh = 0
   for k in range(len(neigh)):
      sum_neigh += 0.8 * (b_state[(NA-1) - neigh[k]]) + 0.1
   will_be_one = sum_neigh / len(neigh)
   return will_be_one


# "Stateswitch" returns the probability for the system to transition
from a state s1 to another s2.

def Stateswitch(s1, s2, MatAdj):
   b_s1 = list(map(int, format(s1, "0" + str(NA) + "b")))
   b_s2 = list(map(int, format(s2, "0" + str(NA) + "b")))
   prob = 1
   for k in range(NA):
      if b_s1[(NA-1) - k] == 0 and b_s2[(NA-1) - k] == 0:
         prob *= 1 - Will_be_one(k, s1, MatAdj)
      elif b_s1[(NA-1) - k] == 0 and b_s2[(NA-1) - k] == 1:
         prob *= Will_be_one(k, s1, MatAdj)
      elif b_s1[(NA-1) - k] == 1 and b_s2[(NA-1) - k] == 0:
         prob *= 1 - Will_be_one(k, s1, MatAdj)
      else:
         prob *= Will_be_one(k, s1, MatAdj)
   return prob


W = zeros([NS, NS], float)
for l in range(NS):
   for c in range(NS):
      W[l, c] = Stateswitch(l, c, M)


# ISING MODEL

def Will_be_one_Ising(agent, state, MatAdj):
```

```
   neigh = Neighbourhood(agent, MatAdj)
   b_state = list(map(int, format(state, "0" + str(NA) + "b")))
   en_neigh = 0
   for k in range(len(neigh)):
      en_neigh -= (2 * b_state[(NA-1)-neigh[k]] - 1)
                                          * (2 * b_state[(NA-1)-agent] - 1)
   en_neigh_flip = 0
   for k in range(len(neigh)):
      en_neigh_flip += (2 * b_state[(NA-1)-neigh[k]] - 1)
                                          * (2 * b_state[(NA-1)-agent] - 1)
      if b_state[(NA-1) - agent] == 0:
         if en_neigh <= en_neigh_flip :
            will_be_one = exp(0.02 * (en_neigh - en_neigh_flip))
         else:
            will_be_one = 1
      else:
         if en_neigh <= en_neigh_flip :
            will_be_one = 1 - exp(0.02 * (en_neigh - en_neigh_flip))
         else:
            will_be_one = 0
   return will_be_one


def Stateswitch_Ising(s1, s2, MatAdj):
   b_s1 = list(map(int, format(s1, "0" + str(NA) + "b")))
   b_s2 = list(map(int, format(s2, "0" + str(NA) + "b")))
   prob = 1
   for k in range(NA):
      if b_s1[(NA-1) - k] == 0 and b_s2[(NA-1) - k] == 0:
         prob *= 1 - Will_be_one_Ising(k, s1, MatAdj)
      elif b_s1[(NA-1) - k] == 0 and b_s2[(NA-1) - k] == 1:
         prob *= Will_be_one_Ising(k, s1, MatAdj)
      elif b_s1[(NA-1) - k] == 1 and b_s2[(NA-1) - k] == 0:
         prob *= 1 - Will_be_one_Ising(k, s1, MatAdj)
      else:
         prob *= Will_be_one_Ising(k, s1, MatAdj)
   return prob


W_Ising = zeros([NS, NS], float)
for l in range(NS):
   for c in range(NS):
      W_Ising[l, c] = Stateswitch_Ising(l, c, M)
```

```
# Now comes the handling of probabilities *************************

# "Compl" returns the complement of a set.

def Belong(k, List):
   compt = 0
   for n in range(len(List)):
      if k == List[n]:
         compt += 1
      else:
         compt += 0
   return(compt)

def Compl(ListMarg, ListAll):
   compl = []
   for k in range(len(ListAll)):
      if Belong(ListAll[k], ListMarg) == 1:
         compl = compl
      else:
         compl.append(k)
   return compl


# The built-in function "Subsets" returns all subsets containing
k elements taken among a set.

def Subsets(List, k):
   return set(it.combinations(List, k))


# "SubsetsCon" returns connected k-subsets by checking if a particular
subset generated by "Subsets" is actually connected.

def SubsetsCon(List, k, M):
   subsets_connected = []
   candidates = list(Subsets(List, k))
   candidates.sort()
   for n in range(len(candidates)):
      M_red = zeros([k, k], float)
      for l in range(k):
         for c in range(k):
            M_red[l, c] = M[(candidates[n])[l], (candidates[n])[c]]
```

```
      for m in range(k-1):
         M_red = dot(M_red, M_red)
      index = 1
      for l in range(k):
         for c in range(k):
            if M_red[l, c] == 0:
               index = 0
      if index == 1:
         subsets_connected.append(candidates[n])
   return subsets_connected


# "KLD" computes the Kullback divergence between two distributions P, Q.

def KLD(P, Q):
   d = 0
   for k in range(len(P)):
      if P[k] != 0:
         d += P[k] * log(P[k] / Q[k])
      else:
         d += 0
   return d


# Initialize the joint distribution

p_tot = array([range(NS), [0.0]*NS]).T
for k in range(NS):
   p_tot[k,1] = random()
p_tot[:, 1] = p_tot[:, 1] / sum(p_tot[:, 1])



# Zero-th order, uniform, estimation

p_null = array([range(NS), [0.0]*NS]).T
for k in range(NS):
   p_null[k,1] = 1 / NS


p_guess_temp = array([range(NS), [0.0]*NS]).T
```

```
# "p_marg" returns the marginal distribution of p over some subset
of the variables.

def p_marg(p, marg):
   NM = len(marg)
   NSM = 2**NM
   p_marg = array([range(NSM), [0.0]*NSM]).T
   for k in range(NSM):
      # Converts state k from decimal to binary.
      # Here the last bit codes the first agent.
      b_k = list(map(int, format(k, "0" + str(NM) + "b")))
      for n in range(NS):
         b_n = list(map(int, format(n, "0" + str(NA) + "b")))
         L = list(map(lambda i: b_n[(NA - 1) - marg[i]]
                                    == b_k[(NM - 1) - i], range(NM)))
         if fun.reduce(lambda x,y: x and y, L):
            p_marg[k, 1] += p[n, 1]
   return p_marg


# "Fact" returns the factorization of some distibution p.

def Fact(p):
   p_fact = array([range(NS), [0.0]*NS]).T
   for k in range(NS):
      b_k = list(map(int, format(k, "0" + str(NA) + "b")))
      p_fact[k, 1] = 1
      for l in range(NA):
         p_fact[k, 1] *= p_marg(p_tot, [l])[b_k[(NA-1) - l], 1]
   return p_fact



# "MultiInfo" returns the total information encapsulated in p.

def MultiInfo(p):
   MI = KLD(p[:,1], Fact(p)[:,1])
   return MI



# "Brown" is the function that finally decomposes the total information
by using  the building blocks defined above.
```

```
def Brown(p_tot, M):
   p_guess = Fact(p_tot)   # We start from the factorized distribution.
   distancestop_tot = zeros(NA+1, float)
   distancestop_tot[0] = KLD(p_tot[:,1], p_null[:,1])
   distancestop_tot[1] = KLD(p_tot[:,1], Fact(p_tot)[:,1])
   distancestop_tot[NA] = 0
   # Loop over the order of interaction. The loop could start from 1
      but we already know that the first order approximation is p_fact.
   for r in range(2, NA+1):
      subsets = list(SubsetsCon(ListA, r, M))
      subsets.sort()
      # We choose to make two adjustments for each subset.
      for k in range(2 * len(subsets)):
         compl = Compl(subsets[k%len(subsets)], ListA)
         for l in range(NS):
            s = list(map(int, format(l, "0" + str(NA) + "b")))
            for m in range(len(compl)):
               s.pop((NA - 1 - m) - (compl[m] - m))
            s_dec = 0
            for g in range(r):
               s_dec += s[g] * 2**(r - 1 - g)
            # The scaling algorithm properly said
            p_guess_temp[l, 1] = p_guess[l, 1]
                    * p_marg(p_tot, subsets[k%len(subsets)])[s_dec, 1]
                    / p_marg(p_guess, subsets[k%len(subsets)])[s_dec, 1]
         p_guess = p_guess_temp
         p_guess[:,1] = p_guess[:,1] / sum(p_guess[:,1])
      distancesap_tot[r] = KLD(p_tot[:,1], p_guess[:,1])
   Coefs = zeros(NA+1, float)
   # Since C_0 is not defined the corresponding slot is used
      for storing the total information
   Coefs[0] = MultiInfo(p_tot)
   for k in range(1, NA+1):
      Coefs[k] = distancesap_tot[k-1] - distancesap_tot[k]
   return Coefs


# Main ************************************************************

# We operate on p_tot through the transition matrix W obtained by
either dynamics above.

itermax = 20
results = zeros([itermax, NA+1], float)
```

```
results[0, :] = Brown(p_tot, M)

for t in range(1, itermax):
   print(t)
   p_tot[:, 1] = dot(p_tot[:, 1].T, W)
   results[t, :] = Brown(p_tot, M)


print(results)

for k in range(NA+1):
   plot(results[:, k])
show()


# END ************************************************************
```

Bibliography

[1] S. Amari and N. Hiroshi. *Methods of Information Geometry*. American Mathematical Society, Rhode Island, 2000.

[2] J. V. Andersen and D. Sornette. The $-game. *Eur. Phys. Jour. B*, 31:141–145, 2003.

[3] A. Arbona, C. Bona, C. Bona, and A. Plastino. A fisher-gradient complexity in systems with spatio-temporal dynamics. *Physica A*, 448:216–223, 2015.

[4] A. Arbona, C. Bona, B. Miñano, and A. Plastino. Statistical complexity measures as telltale of relevant scales in emergent dynamics of spatial systems. *Physica A*, 410:1–8, 2014.

[5] Aristotle. *Metaphysics*. Harvard University Press, Cambridge, Massachusetts, 1933.

[6] B. W. Arthur. Inductive reasoning and bounded rationality: the El Farol problem. *An. Econ. Rev.*, 84:406–411, 1994.

[7] N. Ay, E. Olbrich, N. Bertschinger, and J. Jost. A Geometric Approach to Complexity. *Chaos*, 21, 2011.

[8] B. E. Baaquie. *Quantum Finance: Path Integrals and Hamiltonians for Options and Interest Rates*. Cambridge University Press, Cambridge, 2004.

[9] P. Bak. *How Nature Works: The Science of Self-Organized Criticality*. Copernicus, New York, 1996.

[10] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: an explanation of $1/f$ noise. *Phys. Rev. Lett.*, 59:381–384, 1987.

[11] Y. Bar-Yam. A Mathematical Theory of Strong Emergence Using Multiscale Variety. *Complexity*, 9(6):15–24, 2004.

[12] Y. Bar-Yam. Multiscale Complexity/Entropy. *Advances in Complex Systems*, 7:47–63, 2004.

[13] C. Beck. Generalized information and entropy measures in physics. *Contemporary Physics*, 50(4):495–510, 2009.

[14] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. Walczak. Statistical Mechanics for Natural Flocks of Birds. *Proc. Natl. Acad. Sci. (USA)*, 109:4786–4791, 2012.

[15] J. Bouchaud and M. Potters. *Theory of Financial Risk and Derivative Pricing.* Cambridge University Press, Cambridge, 2009.

[16] P. Brandimarte. *Handbook in Monte Carlo Simulation.* Wiley & Sons, New York, 2014.

[17] P. J. Brockwell and R. A Davis. *Time Series: Theory and Methods.* Springer, Berlin, 1991.

[18] D. C. Brody and E. Graefe. Information Geometry of Complex Hamiltonians and Exceptional Points. *Entropy*, 15:3361–3378, 2013.

[19] D. T. Brown. A Note on Approximations to Discrete Probability Distributions. *Information and Control*, 2:386–392, 1959.

[20] T. Bury. *Collective behaviours in the stock market: A maximum entropy approach.* PhD thesis, Université libre de Bruxelles, 2014.

[21] A. Campa, T. Dauxois, D. Fanelli, and S. Ruffo. *Physics of Long-Range Interacting Systems.* Oxford University Press, Oxford, 2014.

[22] A. Cavagna, I. Giardina, F. Ginelli, T. Mora, D. Piovani, R. Tavarone, and A. Walczak. Dynamical Maximum Entropy Approach to Flocking. *Phys. Rev. E*, 89:042707, 2014.

[23] D. Chalet and Y. C. Zhang. On the minority game: Analytical and numerical studies. *Physica A*, 256:514–532, 1998.

[24] J.-R. Chazottes and G. Keller. Pressure and Equilibrium States in Ergodic Theory. In *Encyclopedia of Complexity and Systems Science.* Springer, 2009.

[25] G. Chliamovitch, A. Dupuis, and B. Chopard. Maximum Entropy Rate Reconstruction of Markov Dynamics. *Entropy*, 17:3738–3751, 2015.

[26] G. Chliamovitch, A. Dupuis, A. Golub, and B. Chopard. Improving Predictability of Time Series Using Maximum Entropy Methods. *EPL*, 110:10003, 2015.

[27] G. Chliamovitch, O. Malaspinas, and B. Chopard. A Truncation Scheme for the BBGKY2 Equation. *Entropy*, 17:7522–7529, 2015.

[28] B. Chopard and M. Droz. *Cellular Automata Modeling of Physical Systems.* Cambridge University Press, Cambridge, 1998.

[29] R. Cofré and B. Cessac. Exact computation of the maximum-entropy potential of spiking neural-network models. *Phys. Rev. E*, 89(052117), 2014.

[30] P. Collet and J.-P. Eckmann. *Concepts and Results in Chaotic Dynamics*. Springer, Berlin, 2006.

[31] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, 2006.

[32] D. R. Cox and H. D. Miller. *The Theory of Stochastic Processes*. Chapman & Hall, Boca Raton, 1965.

[33] A. Czaplicka, J. A. Holyst, and P. M. A. Sloot. Noise enhances information transfer in hierarchical networks. *Nature Scientific Reports*, 3(1223), 2013.

[34] A. Czaplicka, K. Suchecki, B. Miñano, M. Trias, and J. A. Holyst. Information slows down hierarchy growth. *Phys. Rev. E*, 89(062819), 2014.

[35] R. Descartes. *Discours de la méthode*. Ian Maire, Leiden, 1637.

[36] R. C. Dewar. Maximum Entropy Production as an Inference Algorithm that Translates Physical Assumptions into Macroscopic Predictions: Don't Shoot the Messenger. *Entropy*, 11:931–944, 2009.

[37] P. D. Dixit. Stationary properties of maximum-entropy random walks. *Phys. Rev. E*, 92(042149), 2015.

[38] P. D. Dixit and K. A. Dill. Inferring Microscopic Kinetic Rates from Stationary State Distributions. *J. Chem. Theory Comput.*, 10:3002–3005, 2014.

[39] P. D. Dixit, A. Jain, G. Stock, and K. A. Dill. Inferring Transition Rates of Networks from Populations in Continuous-Time Markov Processes. *J. Chem. Theory Comput.*, 11:5464–5472, 2015.

[40] R. P. Feynman and A. R. Hibbs. *Quantum Mechanics and Path Integrals*. McGraw-Hill, New York, 1965.

[41] P. Fiedor. Maximum Entropy Production Principle for Stock Returns. In *IEEE Symposium Series on Computational Intelligence*, pages 695–702, 2015.

[42] A. Fronczak, P. Fronczak, and J. A. Holyst. Thermodynamic forces, flows, and onsager coefficients in complex networks. *Phys. Rev. E*, 76(061106), 2007.

[43] S. Galam. *Sociophysics: A Physicist's Modeling of Psycho-political Phenomena*. Springer, New York, 2012.

[44] F. R. Gantmacher. *The Theory of Matrices*. Chelsea, New York, 1984.

[45] G. A. Gottwald and M. Oliver. Boltzmann's Dilemma: An Introduction to Statistical Mechanics via the Kac Ring. *SIAM Rev.*, 51:613–635, 2009.

[46] A. Greven, G. Keller, and G. Warnecke. *Entropy*. Princeton University Press, Princeton, 2003.

[47] O. Har-Shemesh, R. Quax, A. G. Hoekstra, and P. M. A. Sloot. Information geometric analysis of phase transitions in complex patterns: the case of the Gray-Scott reaction-diffusion model. *Journal of Statistical Mechanics: Theory and Experiment*, 2016(4), 2016.

[48] S. Harris. *An Introduction to the Theory of the Boltzmann Equation*. Holt, Rinehart, and Winston, New York, 1971.

[49] K. Huang. *Statistical Mechanics*. John Wiley & Sons, New York, 1963.

[50] D. Hume. *A Treatise on Human Nature*. John Noon, London, 1739.

[51] K. Ilinski. *Physics of Finance: Gauge Modelling in Non-equilibrium Pricing*. John Wiley & Sons, New York, 2001.

[52] E. T. Jaynes. Information Theory and Statistical Mechanics. *Phys. Rev.*, 106(4):620–630, 1957.

[53] E. T. Jaynes. Information Theory and Statistical Mechanics. II. *Phys. Rev.*, 108(2):171–190, 1957.

[54] E. T. Jaynes. The Minimum Entropy Production Principle. *Annual Review of Physical Chemistry*, 31:579–601, 1980.

[55] E. T. Jaynes. On the rationale of maximum entropy methods. *Proc. IEEE*, 70:939–952, 1982.

[56] E. T. Jaynes and G. L. Bretthorst. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, 2003.

[57] G. Kaniadakis. Statistical mechanics in the context of special relativity. *Phys. Rev. E*, 66:056125, 2002.

[58] G. Kaniadakis. Statistical mechanics in the context of special relativity. II. *Phys. Rev. E*, 72:036108, 2005.

[59] A. Y. Khinchin. *Mathematical Foundations of Information Theory*. Dover, Mineola, 1957.

[60] H. Kleinert. *Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets*. World Scientific, Singapore, 2009.

[61] H. J. Kreuzer. *Nonequilibrium Thermodynamics and its Statistical Foundations*. Oxford University Press, Oxford, 1981.

[62] C. G. Langton. Computation at the Edge of Chaos: Phase Transitions and Emergent Computation. *Physica D*, 42:12–37, 1990.

[63] F. C. Leone, L. S. Nelson, and R. B. Nottingham. The Folded Normal Distribution. *Technometrics*, 3(4):867–887, 1961.

[64] H. Levy, M. Levy, and S. Solomon. *Microscopic Simulation of Financial Markets*. Academic Press, New York, 2000.

[65] W. Li. Mutual Information Functions versus Correlation Functions. *J. Stat. Phys.*, 60(5/6):823–837, 1990.

[66] R. L. Liboff. *Kinetic Theory*. Springer, New York, 2003.

[67] K. Lindgren. Correlations and Random Information in Cellular Automata. *Complex Systems*, 1:529–543, 1987.

[68] K. Lindgren and M. G. Nordahl. Complexity Measures and Cellular Automata. *Complex Systems*, 2:409–440, 1988.

[69] M. C. Mackey. *Time's arrow*. Springer, New York, 1992.

[70] R. N. Mantegna and H. E. Stanley. *Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge University Press, Cambridge, 1999.

[71] O. Marre, S. El Boustani, Y. Fregnac, and A. Destexhe. Prediction of Spatiotemporal Patterns of Neural Activity from Pairwise Correlations. *Phys. Rev. Lett.*, 102:138101, 2009.

[72] S. Mischler and C. Mouhot. Kac's program in kinetic theory. *Invent. math.*, 193(1):1–147, 2013.

[73] T. Mora and W. Bialek. Are Biological Systems Poised at Criticality ? *J. Stat. Phys.*, 144:268–302, 2011.

[74] R. M. Neumann. entropic approach to Brownian motion. *Am. J. Phys.*, 48:354–357, 1980.

[75] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, New York, 2010.

[76] M. E. J. Newman and G. T. Barkema. *Monte Carlo Methods in Statistical Physics*. Oxford University Press, New York, 1999.

[77] T. Obuchi, S. Cocco, and R. Monasson. Learning Probabilities From Random Observables in High Dimensions: The Maximum Entropy Distribution and Others. *J. Stat. Phys.*, 161:598–632, 2015.

[78] T. Obuchi and R. Monasson. Learning probability distributions from smooth observables and the maximum entropy principle: some remarks. *J. Phys.: Conf. Ser.*, 638:012018, 2015.

[79] T. Padmanabhan. Statistical mechanics of gravitating systems. *Physics Reports*, 188:285–362, 1990.

[80] O. Penrose. *Foundations of Statistical Mechanics.* Oxford University Press, Oxford, 1970.

[81] R. D. Peters, M. Le Berre, and Y. Pomeau. Prediction of catastrophes: An experimental model. *Phys. Rev. E*, 86(026207), 2012.

[82] R. Quax, A. Apolloni, and P. M. A. Sloot. The diminishing role of hubs in dynamical processes on complex networks. *Journal of the Royal Society Interface*, 10(88), 2013.

[83] R. Quax, G. Chliamovitch, A. Dupuis, J.-L. Falcone, B. Chopard, A. G. Hoekstra, and P. M. A. Sloot. Information processing features can detect behavioral regimes of dynamical systems. In preparation.

[84] R. Quax, D. Kandhai, and P. M. A. Sloot. Information dissipation as an early-warning signal for the lehman brothers collapse in financial time series. *Sci. Rep.*, 3(1898), 2013.

[85] D. Ruelle. *Thermodynamic Formalism: The Mathematical Structure of Equilibrium Statistical Mechanics.* Addison-Wesley, New York, 1978.

[86] A. Rényi. On measures of information and entropy. *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, 1961.

[87] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek. Weak Pairwise Correlations Imply Strongly Correlated Network States in a Neural Population. *Nature*, 440:1007–1012, 2006.

[88] E. Schneidman, S. Still, M. J. Berry, and W. Bialek. Network Information and Connected Correlations. *Phys. Rev. Lett.*, 91(238701), 2003.

[89] T. Schreiber. Measuring Information Transfer. *Phys. Rev. Lett.*, 85:461–464, 2000.

[90] E. Schrödinger. *What Is Life ?* Cambridge University Press, Cambridge, 1944.

[91] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423, 1948.

[92] Y. G. Sinai. *Introduction to ergodic theory.* Princeton University Press, Princeton, 1976.

[93] H. E. Stanley. *Introduction to Phase Transitions and Critical Phenomena.* Oxford University Press, Oxford, 1971.

[94] G. J. Stephens and W. Bialek. Statistical Mechanics of Letters in Words. *Phys. Rev. E*, 81(066119), 2010.

[95] A.-S. Sznitman. Equations de type de Boltzmann, spatialement homogènes. *Z. Warscheinlichkeitstheorie verw. Gebiete*, 66:559–592, 1984.

[96] C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.*, 52(479-487), 1988.

[97] C. Tsallis, M. Gell-Mann, and Y. Sato. Asymptotically scale-invariant occupancy of phase space makes the entropy S_q extensive. *Proc. Nat. Acad. Sci.*, 102(42):15377–15382, 2005.

[98] P. Turchin, T. E. Currie, E. A. L. Turner, and S. Gavrilets. War, space, and the evolution of old world complex societies. *Proc. Nat. Acad. Sci.*, 110(41):16384–16389, 2013.

[99] Jos Uffink. Compendium of the foundations of classical statistical physics. In *Handbook for the Philsophy of Science*. North Holland, 2006.

[100] E. Van der Straeten. Maximum Entropy Estimation of Transition Probabilities of Reversible Markov Chains. *Entropy*, 11:867–887, 2009.

[101] N. G. Van Kampen. *Stochastic processes in physics and chemistry*. North-Holland, Amsterdam, 1992.

[102] J. C. Vasquez, A. Palacios, O. Marre, M. J. Berry, and B. Cessac. Gibbs distribution analysis of temporal correlation structure on multicell spike trains from retina ganglion cells. *J. Physiol. Paris*, 106(2):120–127, 2012.

[103] E. Verlinde. On the origin of gravity and the laws of Newton. *Journal of High Energy Physics*, 2011(4):1–27, 2011.

[104] J. Voit. *The Statistical Mechanics of Financial Markets*. Springer, Berlin, 2005.

[105] S. Watanabe. Information Theoretical Analysis of Multivariate Correlation. *IBM Journal*, 14(3):66–82, 1960.

[106] J. N. Webb. *Game Theory: Decisions, Interaction and Evolution*. Springer, London, 2007.

[107] A. D. Wissner-Gross and C. E. Freer. Relativistic statistical arbitrage. *Phys. Rev. E*, 82:056104, 2010.

[108] A. D. Wissner-Gross and C. E. Freer. Causal Entropic Forces. *Phys. Rev. Lett.*, 110:168702, 2013.

[109] S. Wolfram. Statistical Mechanics of Cellular Automata. *Reviews of Modern Physics*, 55(3):601–644, 1983.

[110] S. Wolfram. *A New Kind of Science*. Wolfram Media, Champaign, 2002.