THE UNIVERSITY OF WARWICK

University of Warwick institutional repository: http://go.warwick.ac.uk/wrap

A Thesis Submitted for the Degree of PhD at the University of Warwick

http://go.warwick.ac.uk/wrap/55996

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

THE UNIVERSITY OF WARWICK

AUTHOR: **Marina Diakonova** DEGREE: **Ph.D.**

TITLE: **Persistent Mutual Information**

DATE OF DEPOSIT: ....

I agree that this thesis shall be available in accordance with the regulations governing the University of Warwick theses.

I agree that the summary of this thesis may be submitted for publication.

I **agree** that the thesis may be photocopied (single copies for study purposes only).

Theses with no restriction on photocopying will also be made available to the British Library for microfilming. The British Library may supply copies to individuals or libraries. subject to a statement from them that the copy is supplied for non-publishing purposes. All copies supplied by the British Library will carry the following statement:

> "Attention is drawn to the fact that the copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's written consent."

AUTHOR'S SIGNATURE: ....

USER'S DECLARATION

1. I undertake not to quote or make use of any information from this thesis without making acknowledgement to the author.

2. I further undertake to allow no-one else to use this thesis while it is in my care.

DATE SIGNATURE ADDRESS

....

....

....

....

....

www.warwick.ac.uk

Persistent Mutual Information

by

Marina Diakonova

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Complexity Science and Physics

September 2012

THE UNIVERSITY OF WARWICK

Contents

Acknowledgments

I would like to thank my supervisors, Robin Ball and Robert MacKay, for their guidance, help and their great patience, all of which made this possible.

I also wish to thank Alex Stewart, without whom this would not have been started in the first place, Dave Howden, who spent so much of his own PhD helping others, and Mikołaj Sierżęga, who went through it with me in the end and generally kept me sane. A special thank you to my mother, grandmother and aunt, as well as Anna Bakanina and Vera Steinwald, who were not to be deterred even by the long spells of noncommunication. And finally to Tim Evans for that wonderful first course in mathematics twelve years ago, and all the conversations, encouragement and advice ever since.

Declarations

This thesis is original work, some of which has been published in Ball et al. [2010].

Abstract

We study Persistent Mutual Information (PMI), the information about the past that persists into the future as a function of the length of an intervening time interval. Particularly relevant is the limit of an infinite intervening interval, which we call Permanently Persistent MI. In the logistic and tent maps PPMI is found to be the logarithm of the global periodicity for both the cases of periodic attractor and multi-band chaos. This leads us to suggest that PPMI can be a good candidate for a measure of strong emergence, by which we mean behaviour that can be forecast only by examining a specific realisation.

We develop the phenomenology to interpret PMI in systems where it increases indefinitely with resolution. Among those are area-preserving maps. The scaling factor Γ for how PMI grows with resolution can be written in terms of the combination of information dimensions of the underlying spaces. We identify Γ with the extent of causality recoverable at a certain resolution, and compute it numerically for the standard map, where it is found to reflect a variety of map features, such as the number of degrees of freedom, the scaling related to existence of different types of trajectories, or even the apparent peak which we conjecture to be a direct consequence of the stickiness phenomenon. We show that in general only a certain degree of mixing between regular and chaotic orbits can result in the observed values of Γ. Using the same techniques we also develop a method to compute PMI through local sampling of the joint distribution of past and future.

Preliminary results indicate that PMI of the Double Pendulum shows some similar features, and that in area-preserving dynamical systems there might be regimes where the joint distribution is multifractal.

Chapter 1

Introduction

1.1 Complexity Science

The scientific method relies on the fact that reality is distinctly tractable (read predictable) on a number of levels. Here we do not mean Comte's layered separation of the subjects of human thought, though the history of emergence as a concept can certainly be traced along those lines. Rather by levels we mean categories of material substances defined by the particular manner of their interactions (Anderson [1972] or Marvin [1912] for a view that also includes the Logical).

Objects on a level of higher order are typically taken to be aggregates of objects of lower orders. The key questions here are about the extent and nature of this horisontal connectedness. They raise philosophical issues of the ontological and causal nature of level elements. Conversely these considerations could yield answers as to how to define a level in the first place.

Emergence is a phenomenon by which the difference between levels becomes in some ways fundamental, at least as far as the eye can see. This is expressed in the qualitatively different nature of element interactions, which in turn means that higher order behaviour cannot be predicted or explained using knowledge of lower-level processes.

Such conclusions are relevant in the scientific sense insofar as the limitations they place on the process of discovery. At the heart of Complexity Science are attempts to quantify the extent of unpredictability arising out of the differing nature of relations between conglomerates. Subjects of such studies that encompass distinct types of interactions or entities and that potentially display an extent of unexplainability are labelled Complex Systems.

Weaver [1948] made a point of differentiating between *complex* and *complicated* behaviour. The problem with defining a *complex* system exactly is linked to not knowing when and if a system would display emergent behaviour, which of course lies at the heart of the issue. This semantic interrelation between the two contexts is dangerous in the sense that defining one should not merely shift the weight on the other, as Bedau is criticised for by Thorén and Gerlee [2010].

Research presented here concerns a quantity that could potentially measure the extent of unpredictability and hence the level of emergence. We are not so much concerned with finding an appropriate semantic balance since we do not introduce any new philosophical definitions. For our purposes it is emergence, rather than complexity, that becomes the prism through which to view Complexity Science. This provides a framework in which to view the discipline. We therefore first review the history of the emergence concept and the reason behind the recent revival of scientific interest and only then talk about systems and languages in which notions from the theory of complexity are discussed, and in which our work will be based.

1.1.1 Emergence

One of the perceptions connected to emergence is of a new behaviour that was not obviously displayed by the components. There are so many ways in which objects can be combined - that detecting for example a pattern, which is of course a way of phrasing new relatedness - leads to the supposition of some predeliberation. The system must have already contained the notion of the pattern, of how things should be arranged at this higher level. The process of realising this, of something emerging, was perceived as being akin to magic - closed, inexplicable (Goldstein [1999]). The questions of "how" were replaced with speculations on "why". Philosophical considerations of emergence have always been at least partially theological[1].

Its roots go back to the beginnings of natural philosophy itself. There is a level on which this is not surprising, since it is postulates about the nature of reality that lie at the origin of science. Emergence as a thread running through the history of human thought is a sequence of ideas linking the appearance of order, Life, and Mind, to the mechanisms behind the universe as they appeared in contemporary understanding.

Ancient concepts linked to modern emergence are those involving a direction or potentiation. Aristotle is often misquoted to have said *the whole is greater than the sum of its parts* - but that is misleading. The context of this line from *Metaphysics* is an offered solution to Zeno's paradox, with the suggestion that the whole comes *before* the parts, whose being springs *from* the whole. Aristotle argued that all development is the processes of actualisation, the unfolding of some universal potential that is already contained as a seed in all things. Later on Plotinus had a similar notion related to an impersonal potential.

By the 19th century the world, and in particular life, was increasingly seen as being ultimately explainable. The old order was swept away, and according to Comte knowledge entered the third, positivist stage. As reductionism was taking hold, sciences were branching out and becoming more specialised. In this setting a new concept of an essentially

[1] In best of soviet traditions here we refer the reader to Engels. The argument of the transition of the quantitative into the qualitative, so preemptive of the ontological view of emergence, continues to resonate even today (see McGarr [1994] for a possibly politically-biased review).

immanent emergence was introduced by G.H.Lewes.

In *Problems of Life and Mind* Lewes bridged reductionism and Kant's transcendentalism by referring to one's perception of oneself as essentially non-dualist in nature. The force that combines elements of the Body to make up the Mind need not be external; and yet we do not need drop the apparent mystery altogether. Lewes juxtaposes two types of aggregates, the *Resultant* and the *Emergent.* Resultants arise out of simple aggregations; Emergents are outcomes of processes that *resist description.*

This was the origin of the term "emergence" and the basis for emergentism as a philosophical discipline. Further developments involved concepts differing based on whether any ontological or causal weight was attached to the aggregates, possible direction of causality, etc. These next major contributions came from an early 20th century group of mostly British scientists and philosophers; the context, similar to Lewes, was evolution.

These emergentists occupied a stance halfway between vitalists and reductionists, who were then referred to as *mechanists.* Vitalists like Bergson posited an *elan vital*, an external driving force as a major organisational principle. One of the first texts that offered an alternative position was *The Mind and its Place in Nature* by C.D.Broad. Broad recognises these organisational tendencies of organisms but rejects the necessity of bringing in a *deus ex machina.* Living beings are not machines; the aggregates of various orders that make them up display behaviour fundamentally different to that of the constituents. This was a statement of features and relatedness, and did not require a break with monoism. Interestingly his views single out the Mind as possessing an organisational centre, an ontological mental substance that gives rise to various mental processes. This is not dualistic in that this other kind of substance is not taken to preexist. Neither is it reductionist since by 'emergent' Broad means behaviours that are in principle not deducible but only *recognisable.*

This proto-emergent trend was picked up by C.L.Morgan. By today's more-scientific standards Morgan's philosophy is firmly in the camp of the 'strong' emergence. Clayton [2006] criticises his lack of parsimony in attributing the strongest possible, ontological connotations to higher-level objects, while insisting that the actual novel features can be expressed as statements of relatedness. Morgan makes several conjectures that could be viewed with the same reservations, such as allowing for downward causality, or considering evolution as a sequence of discrete jumps[2]. Nevertheless his claims “there is increasing complexity in integral systems as new kinds of relatedness are successively supervenient”, or “there is an ascending scale of what we may speak of as richness in reality” read like the motivation typically accompanying research that places itself firmly under the umbrella of Complexity Science.

By mid-twentieth century the hype had gone down. Optimising strategies for the firing of machine guns led to the realisation of the importance of feedback loops, and building the model of the Mind became but a matter of time: “seeing Man through the lens of logic, information and communication theory as transparent, with no hidden depths”, Goujon [2006]. Yet at the string of Macy conferences that followed the cyberneticists became increasingly confounded by psychologists presenting evidence from tighter, better controlled experiments in which human behaviour substantially differed from that of a robot. To quote Ludwig von Bertalanffy,“We may consider individuals as robots, and even transform them more and more into robots of consumption, of politics and of the industrial-military complex. But we pay for this dearly by moving nearer to *Brave New World* and *1984*; by neuroses, hippies, drug addiction, riots, wars and other symptoms of a sick society”.

This was said in, not surprisingly, 1968, at the Alpbach symposium organised to vent the frustration felt by the scientific community at the mechanistic approach that was increasingly perceived as failing. The answer, systems theory, was emergentist in that it called for “a change in basic categories of knowledge” (Arthur Koestler and John R. Smythies (editors) [1968]), noting that organised structures can be viewed as ‘wholes’ that show a different, new range of behaviour. The emphasis here was on the relations between the constituent parts that was seen to be independent of their ‘position’ in the ontological layered structure. This “isomorphism” is exactly what was picked up by the later proposals of *universality* in theories such as self-organised criticality. Yet another ‘emergence rule’ that is being proposed by A. Barabasi was foreseen in the lecture - that of similar behaviour of graph variables.

Alongside cybernetics it was information theory that was being challenged. Information theory was formalised by Shannon in 1948. Its birth can once more be attributed to wartime need, though this time the aim is that of reliable signal transmission. One of the measures

[2] His system of reality levels, called here ‘logical strata’, curiously places the mathematical at the foundation and the Mind at the top, while still maintaining pyramidal structure.

was that of the spread of the probabilities of possible outcomes. Shannon constructed a function that fit the specifications and on suggestion from (von Neumann) labelled it entropy (section 1.2.2). Comparing it to Boltzmann's entropy, we see that the information-theoretic entropy is a composite concept[3]. Thus a function effectively expressing the average information in a message became operationally equivalent to a purely thermodynamical measure of disorder. An easy to spot juxtaposition lies in the objective nature of one, and the very subjective nature of another. It is exactly this disassociation of information theory from meaning that started the questions about the suitability of using it to describe the more 'human' aspects. "Every culture creates a world by selecting from the background noise of events, certain signals which it treats as messages by giving them meaning" (cited in Goujon [2006]).

The growing trends thus stressed the more holistic approach. There were a number of fields in the second half of the twentieth century that fall broadly under the auspices of complexity science, and that brought about once more philosophical speculations about the nature of complexity and emergence; so much so that, to quote J. Goldstein,"Emergence functions not so much as an explanation but rather as a descriptive term pointing to patterns, structures, or properties that are exhibited on the macro-level.[...] An appeal to emergence is thus a way to describe the need to go to the macro level and its unique dynamics, laws, and properties in order to explain more adequately what is going on. The construct of emergence is therefore only a foundation on which to build an explanation, not its terminus". Thus complexity and emergence mean different things depending on one's background - and can range from the existence of phase transitions in many-body systems to the functioning of organisms. We illustrate this plurality of settings by an image from "Arts and Science Factory", see figure 1.1.

Current Understanding As complexity science gained footing, so too did the philosophical speculations return. The semantic distinction that has been applied most in the recent years is that between *strong* and *weak* emergence. The term *weak* was coined by Mark Bedau in an effort to find an appropriate operational definition to a concept already in use. In Bedau [1997] the description is that of behaviour resulting in a macrostate that is derivable only by simulations from the dynamical and the external (and initial) condition[4].

[3] This entropy of a stochastic process is fundamentally different to the entropy introduced by Kolmogorov and Sinai as a function of measurable dynamical systems.

[4] The phenomena covered by this description appear to be one the topics in the Santa Fe school.

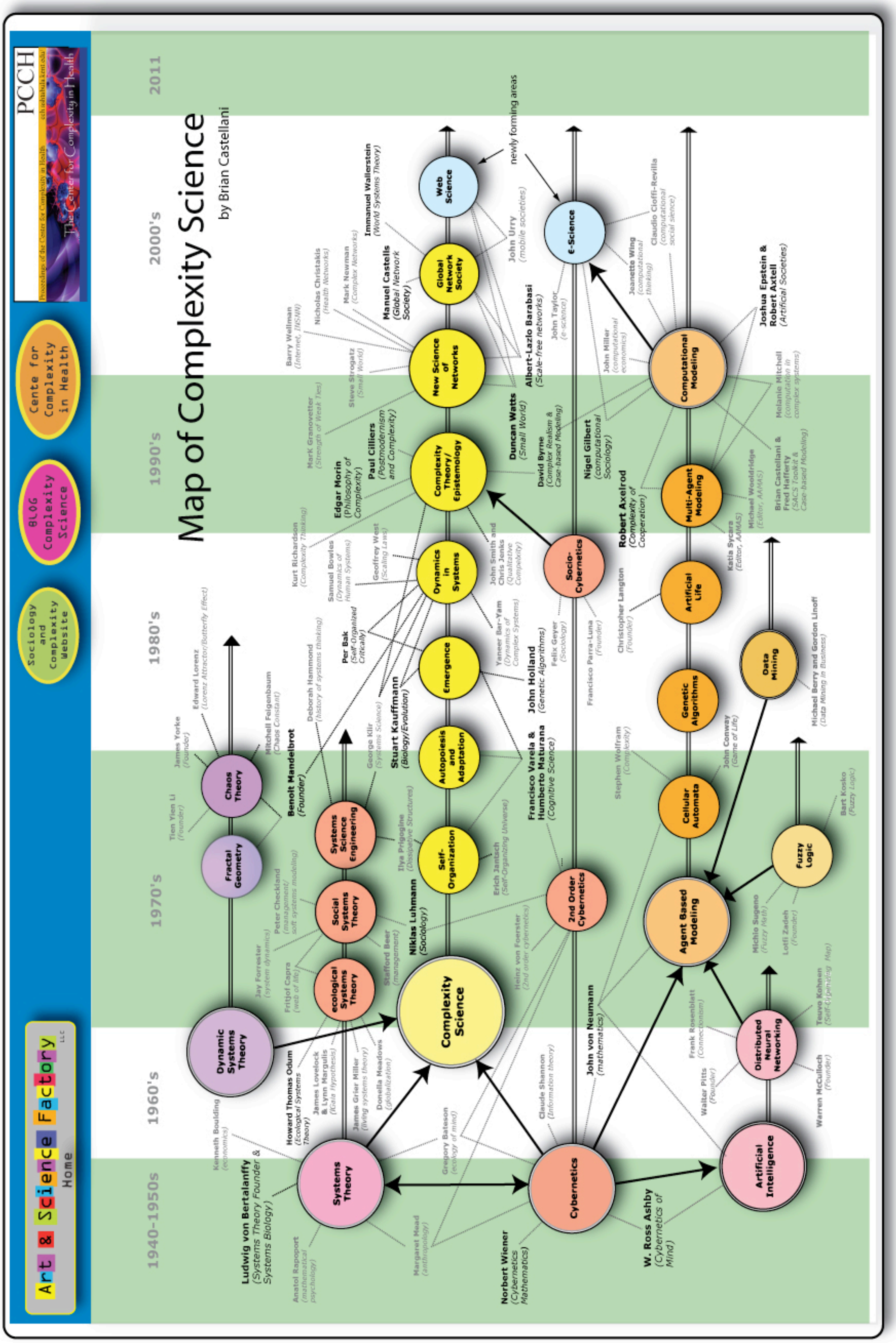

Art & Science Factory LLC
Home
Sociology and Complexity Website
BLOG Complexity Science
Cente for Complexity in Health
PCCH
The Center for Complexity in Health
Map of Complexity Science
by Brian Castellani
1940-1950s
1960's
1970's
1980's
1990's
2000's
2011
Systems Theory
Cybernetics
Artificial Intelligence
Dynamic Systems Theory
Complexity Science
ecological Systems Theory
Social Systems Theory
Systems Science Engineering
Fractal Geometry
Chaos Theory
Self-Organization
Autopoiesis and Adaptation
Emergence
Dynamics in Systems
Complexity Theory / Epistemology
New Science of Networks
Global Network Society
Web Science
2nd Order Cybernetics
Socio-Cybernetics
E-Science
Agent Based Modeling
Cellular Automata
Genetic Algorithms
Artificial Life
Multi-Agent Modeling
Computational Modeling
Distributed Neural Networking
Fuzzy Logic
Data Mining
newly forming areas
Ludwig von Bertalanffy (Systems Theory Founder & Systems Biology)
Anatol Rapoport (mathematical psychology)
Kenneth Boulding (economics)
Margaret Mead (anthropology)
Gregory Bateson (ecology of mind)
Norbert Wiener (Cybernetics Mathematics)
W. Ross Ashby (Cybernetics of Mind)
Claude Shannon (Information theory)
John von Neumann (mathematics)
Walter Pitts (Founder)
Warren McCulloch (Founder)
Frank Rosenblatt (Connectionism)
Howard Thomas Odum (Ecological Systems Theory)
James Lovelock & Lynn Margulis (Gaia Hypothesis)
James Grier Miller (living systems theory)
Donella Meadows (globalization)
Fritjof Capra (web of life)
Jay Forrester (system dynamics)
Peter Checkland (management / soft systems modeling)
Stafford Beer (management)
Niklas Luhmann (Sociology)
Heinz von Foerster (2nd order cybernetics)
Teuvo Kohonen (Self-Organizing Map)
Michio Sugeno (Fuzzy Math)
Lotfi Zadeh (Founder)
Tien Yien Li (Founder)
James Yorke (Founder)
Benoit Mandelbrot (Founder)
Mitchell Feigenbaum (Chaos Constant)
Edward Lorenz (Lorenz Attractor/Butterfly Effect)
Ilya Prigogine (Dissipative Structures)
Erich Jantsch (Self-Organizing Universe)
Francisco Varela & Humberto Maturana (Cognitive Science)
George Klir (Systems Science)
Stuart Kauffmann (Biology/Evolution)
Stephen Wolfram (Complexity)
Bart Kosko (Fuzzy Logic)
John Conway (Game of Life)
Deborah Hammond (history of systems thinking)
Per Bak (Self-Organized Criticality)
John Holland (Genetic Algorithms)
Yaneer Bar-Yam (Dynamics of Complex Systems)
Felix Geyer (Sociology)
Francisco Parra-Luna (Founder)
Christopher Langton (Founder)
Michael Berry and Gordon Linoff (Data Mining in Business)
Kurt Richardson (Complexity Thinking)
Samuel Bowles (Dynamics of Human Systems)
Geoffrey West (Scaling Laws)
John Smith and Chris Jenks (Qualitative Complexity)
Robert Axelrod (Complexity of Cooperation)
Katia Sycara (Editor, AAMAS)
Mark Granovetter (Strength of Weak Ties)
Edgar Morin (Philosophy of Complexity)
Paul Cilliers (Postmodernism and Complexity)
Duncan Watts (Small World)
David Byrne (Complex Realism & Case-based Modeling)
Nigel Gilbert (computational Sociology)
Michael Wooldridge (Editor, AAMAS)
Brian Castellani & Fred Hafferty (SACS Toolkit & Case-based Modeling)
Melanie Mitchell (computation in complex systems)
Barry Wellman (Internet, INSNA)
Steve Strogatz (Small World)
Nicholas Christakis (Health Networks)
Mark Newman (Complex Networks)
Manuel Castells (Global Network Society)
Albert-Lazlo Barabasi (Scale-free networks)
John Miller (computational economics)
John Taylor (e-science)
Joshua Epstein & Robert Axtell (Artificial Societies)
Immanuel Wallerstein (World Systems Theory)
John Urry (mobile societies)
Jeanette Wing (computational thinking)
Claudio Cioffi-Revilla (computational social science)

Figure 1.1: Schematic illustration of the history of ideas usually associated with Complexity Science ("Complexity Map" as published online by the "Arts and Science Factory").

In Paul Davies and Niels H. Gregersen (editors) [2010] the editors observe that our conceptions of reality readily model themselves on the latest technological advances. Bedau's definition appears to fit the same trend - recent scientific progress relied heavily on the newfound ability to simulate behaviour. Consequently weak emergence views reality through this particular prism.

These are metaphysically noncommittal, scientifically comfortable stances. One does not need to reject the monoistic structure to admit unpredictability: the simple fact that equations are not analytically solvable means that there is a limit to how much can be forecast. There is thus a distinction between predictability in principle and in practice. A lot of the theoretically deducible phenomena can thus be called emergent. The prime examples here are deterministic cellular automata (Games of Life), behaviour of networks, or various aspects of evolution. Thus this description does not single out outcomes based on whether they are in any way interesting or surprising; but rather by indicating systems that we cannot (yet?) solve, it seems to have an operational-based support: most emergent macro phenomena are discovered only with the use of simulation. However, we do not know that in some years' time there won't be a new mathematics capable of giving the analytic result. Thus Grelling (as mentioned in Hempel and Oppenheim [1948]) points out that this view of weak emergence is more of a provisional construct.

In this respect it is half way to the more safe approach of doing complexity science without taking a metaphysical stance. From Thorén and Gerlee [2010]: "Contemporary accounts typically strive for weaker formulations trying to salvage some part of the concept whilst giving others up". Chalmers [2006] gives a slightly different definition. Here weak emergence concerns truths that are unexpected (in contrast Chalmers' strong emergence is about truths that are not deducible). Thus too deterministic cellular automata are weakly emergent - even if one would need to resort to calculations the general behaviour could still be deduced. Weak emergence becomes more of a statement of our understanding of the propagation of causality; giving our epistemological position relative to that of Laplace's demon.

Chalmers is also careful to mention that in general weak emergence should say something about the level of difficulty with which the inference takes place, as well as the difference between the complexity of the combination rules and the overall behaviour. The optimal definition of weak emergence thus seems to be a highly subjective operational concept achieved by including all the aspects desired intuitively. A phenomenon is weakly emergent if *complex, interesting high-level function is produced as a result of combining simple low-level mechanism in simple ways.* (Ibid.)

Strong emergence, on the other hand, tends to place itself in direct opposition to reductionism[5]. Accepting this hypothesis means allowing for the existence of laws other than the ones inferable from the scientific methodology, which in turn essentially involves a new kind of science. Here once again there are different schools based on what assumptions or consequences the authors are comfortable with ascribing to this concept. Thus for example Davies [2004] attributes to emergents novel causal powers, and admits downward causation, typically a problematic concept for scientists, one that is most required to be taken on faith. Kim [2006], on the other hand, suggests that philosophical coherence makes it not as simple as just picking attribute - and that admitting some may lead to undermining the whole concept, which is what happens with the circularity of downward causation[6].

Strong emergence is a philosophical conjecture, which for example for Kim [1999] should contain both irreducibility and supervenience. Starting from that approach the main question becomes whether strongly emergent phenomena exist, and if so, what they are. Chalmers supports the view that consciousness is exactly that. Depending on one's theological leanings God could also be 'analysed' in this way (Peacocke [2010], Gregersen [2010]). Though of course since the answers depend on the definition the results are possibly incomparable.

We will be attempting to quantitatively describe the extent to which initial information persists across in time. We too will use the distinction between the strong and weak notion in the loosest possible sense, focusing on epistemology rather than ontology even in the 'strong' case. That part of the thesis that refers back to it does so not because it claims to have found a phenomenon that we claim to be strongly emergent, but rather to notice that a certain statistical function can be used to differentiate between the two concepts *given they are defined in a certain way.* The data used is from chaotic dynamical systems, but our function sees chaos *as such* as a completely uninteresting (giving nothing in terms of forecastability) background noise, looking instead for global structures. The crucial conceptual link between low-dimensional dynamical systems and high-level complex

[5] Everyday usage had a diluting effect on the notion of 'strong'. If 'very strong' (Clayton [2006]) is already in literature, the next step is naturally some form of scale. Bauchau [2006] tentatively proposes one that places chaos somewhere low down, the top being defined by the class of universal computation.

[6] Chalmers also talks about downward causation as a phenomenon in its own right, not necessarily connected to strong emergence. This distinction allows one to view quantum wavefunction collapse as the former, whilst not necessarily supporting the strongly emergent view.

systems can be drawn in a number of ways, defining the 'higher' level at an arbitrary, subjective degree of complexity. One such is to consider the *trajectory* as a 'complex' object, which can be characterised by some aggregate variables - e.g. the Lyapunov exponent - but comes about as a result of, simply, applying the map. Alternatively the dynamical system itself, with the related quantities characterising the geometry, say, of the underlying strange attractors, can be thought of as an 'aggregate', whose succint properties can best be understood not by looking at the equation, but indeed by the aforementioned variables. In the next section we will see that according to *our* definition of emergence, a chaotic attractor with no interesting structure would not be considered as giving rise to emergent behaviour. This will be the case for the fully-developed chaos at the $r = 4$ regime of the logistic map. By contrast the intermediate r values, and in general area-preserving maps, would yield a richer set of results.

1.2 The Probabilistic Framework

We now review a common language in which various correlation, complexity and emergence measures are typically expressed.

The usual aim of physical sciences is to establish a link between observations and reality via an idealisation (a model). The distinction is that reality results in our observations that, in turn, lead to statements about the idealisation. Logic builds a reverse link and allows predictions from the model to be tested against new observations. Consider an archetypal process of tossing a fair coin. Without making a statement about reality we can successfully model the process by random variables. The key word here is 'successfully', which means that there do exist functions of results that are predictable by the model.
Development of probability theory can be traced in the correspondence of Pascal and Fermat, established after Pascal's friend Chevalier De Méré brought to his attention the issues facing gamblers at dice; especially the Autumn 1654 series. Along with establishing the basic rules of the calculus of probabilities, Pascal introduces probability as a value between 0 and 1 that is in some way "attached" to an event (rather than being dependent on the mind of the observer, as M. Miton (see Renyi [1972]) would have it). It expresses the extent of certainty that the event will happen, which Pascal identifies with the actual likelihood of an event coming to pass. The term "probability" is chosen especially so that its numerical value corresponds to our intuitive conceptual use of it[7].

Pascal also suggests that measuring the probability is equivalent to observing relative frequencies of occurrences in long trials. Probability is thus a fixed value around which the relative frequency oscillates in a random fashion. This leads to an effective two-level randomness - uncertainty in how sure one is in an event happening.

This put a start to both the mathematical and the scientific discipline. Probability can be approximated by observations, and subsequent manipulations using the calculus of probabilities allow for prediction, at least statistically. Pascal stresses that partial knowledge about the likelihood of an event occurring or not still constitutes some kind of knowledge about the event, even though the event might not actually come to pass.

[7] Nowadays Pascal would have even less reason to worry that the meaning of "probable" - as a theological conjecture the Vatican is yet to pronounce on - would be the first to spring to mind.

That these statements can be made scientifically rigorous[8], and can be put on a firm mathematical basis, has been postulated only relatively recently. It was Doob and Kolmogorov that proved that the rules of chance constitute a mathematical framework - see Getoor [2009] for a review.

We state the formal probability framework. Let $(\Omega, \mathbb{F}, \mathbb{P})$ be a probability space, and $(E, \mathbb{E})$ a measurable space. We interpret Ω as the space of all possible realisations of the given process. The σ-algebra $\mathbb{F}$ on Ω is then the respective *event* space, and $\mathbb{P}$ is the probability measure. We take E to be a subset of $\mathbb{R}^n$ for some integer n, and associate it with a *measurable* state space of the system.

A motivation in separating Ω from E, the space of possibilities from the potential results of measurements, can be traced to the wish to be more exact about the meaning of measurement. Consider performing any experiment, by which we mean some interaction with a system. It is more usual to measure some feature of the system. In this case it is more obvious that the result of the measurement would be a function of the *actual* state, $X : \Omega \to E$. Measuring the temperature of gas in a box falls in this category[9].

Our observations thus fall in E. Let $e \in E$. Since we identify what we observe with a function of the state of the system,

$$e = X(\omega), \tag{1.1}$$

where $\omega \in \Omega$ is the state of the system. We call function X a random variable, or a variate, or chance variable.

1.2.1 The Concept of Probability

Suppose we take the frequentist approach of associating the likelihood of seeing an outcome with the relative frequency with which this outcome has already been observed in systems of this kind. In this approach relative frequency serves the purpose of creating a measure on E. A random variable was setup as a link between observations in E and some "true" states in Ω. So the probability of seeing $e \in E$ can be thought of as resulting from some probability of the system being in those states that lead to observing e. Hence the common definition of probability: given a random variable X, the probability of observing it take a

[8] ignoring the 'truth' contained in them for a moment - see Diaconis et al. [2007]

[9] We make the optimistic assumption that there is a correspondence between reality and state of the system.

value $A \subseteq E$,

$$P(A) = P(X \in A) = \mathbb{P}\{\omega \in \mathbb{R} : X(\omega) \in A\}. \tag{1.2}$$

In information theory/computation mechanics literature the sets Ω and E are often identified with each other, and the random variables that question the state become the identity functions (although most of the time Ω is not being considered at all).

1.2.2 Entropy and Entropic Concepts

Entropy was introduced as an experimentally determinable quantity expressing the way a system absorbs heat at a given temperature. It was associated with the lack of organisation or order. The second law of thermodynamics posited that in a closed system entropy increases. Boltzmann attempted to justify the second law by replacing the imperative with, simply, vast differences on the scale of improbable. In his framework thermodynamic entropy measured the number of possible configurations of constituent parts that made up some distinct observable state.

Let $X : \Omega \to E$ be a random variable, and P defined by 1.2. The Shannon information of discrete-valued random variable X, introduced in Shannon [1948] [10] is

$$H(X) = -\sum_{x \in E} P(x) \log P(x). \tag{1.3}$$

In a countably infinite support space entropy is defined only if the series converges.

We will also use the differential Shannon entropy defined when $p(x), x \in E$ is probability distribution, and given by

$$H[p] = -\int_{x \in E} dx\ p(x) \log p(x) \tag{1.4}$$

but we will mention the difference between the two later in the text, in a particular context. Whatever information and uncertainty are, conceptually uncertainty is often understood to be the absence of information, and vice versa. Consider a random variable. Before observation there is some uncertainty as to the outcome. Observation corresponds to obtaining an amount $-\log P(x)$ of information. Thus entropy is defined as the average information of a message. Note that even information content in a message doesn't depend on the specific

[10] The probability P is understood to be given; the implication is that the variable is associated with only one probability. This interpretation is one where the variable is an outcome of a process, and so some 'natural', perhaps frequentist, probability can be assigned to it.

message itself, but rather on its probability, a property conferred on it by the system (or by the observer's knowledge of the system). Thus entropy is a function of the measure P and not of the support space.

Relative and Conditional Entropies Given two random variables X and Y, $X, Y : \Omega \to E$ we define the joint entropy

$$H(X,Y) = -\sum_{x,y \in E} P(x,y) \log P(x,y), \tag{1.5}$$

where $P(x,y)$ is the joint probability. The conditional entropy is then

$$H(X|Y) = H(X,Y) - H(Y). \tag{1.6}$$

Conditional entropy measures the amount of uncertainty in the outcome of one variable (here X) given that the outcome of another (Y) is known. Here we always use P to express the notion of probability. The way we defined it earlier rests on the assumption that each random variable comes with a probability we tacitly understand to be its own. Thus $P(x)$ is actually equal to the measure $\mathbb{P}^X\{X^{-1}(x)\}$, and $P(y)$ is $\mathbb{P}^Y\{Y^{-1}(y)\}$, where $\mathbb{P}^X$ and $\mathbb{P}^Y$ are for example given by the relative frequencies of the variables and are not necessarily the same. Thus P stands for a loose sense of 'probability of a random variable'.

The form 1.6 is the functional form of a 'distance' in the space of measures: if Let P, P' be measures on the space of measurable outcomes, then the relative entropy, or the Kullback-Leibler (KL) divergence between P and P', is defined to be

$$KL(P||P') = \sum_{x \in E} P(x) \log \frac{P(x)}{P'(x)}. \tag{1.7}$$

Here we separate P from P' because we view them in their capacities as probability measures.

The logarithm is defined to be equal to zero whenever $P'(x) = 0$ or $P(x) = 0$. KL is not symmetric, and is not technically a metric. Also 1.6 is not symmetric - the information about one outcome given another is not necessarily the same as the reverse.

Mutual Information The mutual information (MI) between X and Y is

$$I(X,Y) = H(X) + H(Y) - H(X,Y). \tag{1.8}$$

As entropy is extensive, the sum of entropies of independent variables should be the same as the entropy of the system made up of these variables. If the joint entropy is less than the sum of marginals it is understood that reduction in uncertainty is at the expense of some interdependence. Mutual information measures the deficit, and thus the degree of interdependence between two variables. It is zero if the two variables are independent (since the joint measure becomes the product of the marginals), is also completely symmetric and always positive.

MI can also be written as

$$I(X,Y) = H(Y) - H(Y|X). \tag{1.9}$$

This form expresses MI as the difference between uncertainty in one outcome (here Y) and the uncertainty in that outcome given that we know the result of another outcome (X). It is thus the information about one variable stored in the other, and is, too, symmetric[11].

Writing MI in terms of probabilities,

$$I(X,Y) = \sum_{x,y \in E} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}, \tag{1.10}$$

we see that mutual information between two variables is actually the relative entropy between the joint distribution and the product of the marginals. If the two variables are independent the joint *becomes* equal to the product of the marginals, and so the divergence between two elements that are actually the same point is zero (here the support space is actually $E \,\mathrm{x}\, E$).

1.2.3 Stochastic Processes: adding time

The framework into which this brings us is that of stochastic processes, i.e. systems where predictability of evolution can be treated using probabilistic tools. A stochastic process is

[11] The information-theoretic framework lends itself to verbal abstractions of the intensity limited only by the author's imagination. Thus in Prokopenko et al. [2009] mutual information is described as

mutual information = receiver's diversity - equivocation of receiver about the source.

defined as a sequence of random variables:

$$\{X_t, t \in T\}. \tag{1.11}$$

Some care must be taken when introducing time. The mathematical framework for discrete processes, otherwise known as sequences, ($T = \mathbb{Z}$) was established by Kolmogorov, and Doob did the same for $T = \mathbb{R}$ which presented more difficulties.

There are several ways of expressing the random variables. Behind the ideas are essentially three spaces: outcomes Ω, states E and time set T. $X(\omega)$ is the random variable independent of time. Including it produces $X(\omega, t)$, or $X_t(\omega)$, the latter notation being more common in the discrete time case.

The strength of this framework is that it allows to formulate dependencies between variables, which in this case are states at different times. It is a language of choice for models where evolution is probabilistic.

The mathematical object encoding any apparent causal structure between states at times in some set T is the *joint* probability of events indexed by elements of T.

Suppose that we have a discrete clock (which we take to be represented by $\mathbb{Z}$) that ticks from $-\infty$ to ∞, and that at every given time $i \in \mathbb{Z}$ a system yields a value from some alphabet $\mathbb{A}$. Thus a specific bi-infinite run of the system gives us a sequence (an element of space $\mathbb{A}^{\mathbb{Z}}$). We want to consider a random variable connected to a fixed time i, or more generally to a block of times from a to b. We can construct a probability space $(\Omega, \mathbb{F}, \mathbb{P})$, where $\Omega = E = \mathbb{A}^{\mathbb{Z}}$, $\mathbb{F}$ is a σ-algebra of cylinder sets, and $\mathbb{P}$ is a probability measure of Ω.

These random variables can be thought of as blocks, or subsequences. The above construct allows us to talk about probability over blocks of arbitrary length. Let $S_a^b = (S_a, S_{a+1}, ..S_b), b, a \in \mathbb{Z}, b \geq a$ be a block of length $b - a + 1$ s.t. $S_a := S_a^a$; and let $\overrightarrow{S_a}$ to be the semi-infinite block starting at a, $\overrightarrow{S_a} = (S_a, S_{a+1}, S_{a+2}..)$, and $\overleftarrow{S_a} = (..S_{a-2}, S_{a-1})$ to be one ending at and not inclusive of a. We define a stationary process as one whose marginals depend only on the length of the subsequence. No major global changes occur in such processes, changes that influence the relative frequency of subprocesses. Stationarity is defined as system with

$$P\left(S_a^{a+N} = A\right) = P\left(S_b^{b+N} = A\right), \tag{1.12}$$

$\forall a, b, N \in \mathbb{Z}^+$ and $A \in \mathbb{A}^{N+1}$. As such we will talk about probabilities of block with length

N, which we call $S^N, N \in \mathbb{Z}^+$.

Entropy Rate and related quantities

Consider the the uncertainty inherent in the system. A way of quantifying the amount (rather than perhaps the role) of chance is to view the data as an outcome of a stochastic process detailed above, and enquire after the entropy per symbol, where by symbol we mean an element of the alphabet $\mathbb{A}$. This quantity is also called the entropy rate. We follow the methodology established in Shannon [1948] and define Shannon entropy per block of length N, H_N, as

$$H_N = H\left[S^N\right] := -\sum_{A \in \mathbb{A}^N} P(S^N = A) \log P(S^N = A). \tag{1.13}$$

The block entropy is always nonnegative, $H_N \geq 0$, and grows monotonically with N, $H_{N'} \geq H_N, \forall N' > N,\ N, N' \in \mathbb{Z}^+$. Shannon defines two quantities, the entropy per symbol in a block of N random variables (starting at zero),

$$G_N := -\frac{1}{N} H[S_0^{N-1}], \tag{1.14}$$

and the average entropy of a new symbol given some past,

$$F_N := -H[S_1 \ |S_{-N+1}^0], \tag{1.15}$$

This is a function of random variables related to each other by the relative time of occurrence, so that the index of the block beginning is by itself arbitrary and is here shown as zero by default (see Cover and Thomas [2006]).

For stationary processes the limits for both G_N and F_N as $N \to \infty$ exist *and* coincide (Shannon [1948]). Hence the definition of the entropy rate h of a stochastic process S (considering that G_N is of course just the normalised block entropy):

$$h = \lim_{N \to \infty} -\frac{1}{N} H_N. \tag{1.16}$$

To illustrate features h picks up on consider:

- No causal link between the variates, and the process not necessarily stationary. S_i

become independent and hence

$$h = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N} H[S_i].$$

Here existence of h is assured unless H is a function of i, which is of course the blueprint of non-stationarity.

- S_i are independent and identically distributed (i.i.d.), then

  $$h = H[S_0],$$

  where the index is again arbitrary. The average entropy per symbol is *the* entropy of a symbol, since all symbols have the same uncertainty. This is not usually true, as h is a property of the system as a whole, a function of the information source rather than of the outcome at some single point in time. That the two are the same here shows that the information source does not store time dependencies.

- If, additionally, each i.i.d. S_i has a uniform measure of a support space of cardinality M, $H[S_i] = \log M$, and hence

  $$h = \log M.$$

  Thus for a coin toss modelled as a stochastic process with i.i.d. outcomes the alphabet would consist of two entries, giving the entropy rate of $\log 2$.

Any skewness in the measure towards a particular outcome of any variate would decrease the entropy rate of the process. Any dependency between variables would reduce the uncertainty per symbol and hence decrease the entropy rate even further. h measures both effects. As we have seen above, it is maximal for i.i.d. variates with uniform measure.

1.2.4 Symbolic Dynamics: linking Deterministic and Stochastic Frameworks

Consider a map $F : X \to X$ and a partition P on the state space $X = \bigsqcup_{i\in C} X_i$, $P_M : X \to \{1, 2, .., M\}$, where $\bigsqcup$ stands for the disjoint union.
$P_M(x \in X)$ gives the index of a cell that contains the point. A corresponding map, which

for convenience we here label with the same letter, turns each orbit O

$$O = \big(x, F(x), F^2(x), .. \big) \tag{1.17}$$

into a symbolic orbit sequence:

$$P_M : \mathbb{O} \to \Sigma_F \tag{1.18}$$
$$(x, F(x), ..) \mapsto (P_M(x), P_M(F(x))..)\,, \tag{1.19}$$

where $\mathbb{O}$ is the set of all orbits. Thus Σ_F is the set of all possible, or *admissible*, symbolic orbit sequences associated with the partition P_M of X, and map F. Note that orbits are defined as being bi-infinite: $s = (s_t)_{t=-\infty}^{\infty}$. Orbit sequences are thus sequences of integers labeling the position of the orbit in the coarse-grained version of the state space.

The symbolic dynamical system is defined as (Σ_F, σ), where the subshift σ is equivalent to the evolution operator, mapping each symbol to the next one (and is as such a function of the entire sequence itself, rather than the symbols):

$$\sigma : \Sigma_F \to \Sigma_F \tag{1.20}$$
$$\sigma\left(P_M(x), P_M(F(x))..\right) \mapsto \sigma\left(P_M(F(x)), P_M(F(F(x)))..\right). \tag{1.21}$$

This shows the process by which one can contextualise the study of dynamical systems in stochastic processes. In the next section we review the two archetypal dynamical systems.

1.3 Toy Models

1.3.1 The Logistic Map

The initial motivation was a model describing population growth. It is clear that in order to allow for some form of stability the system would have to be nonlinear. Interestingly enough, applying the same arguments behind parameters and form of dependencies to a continuous version produces a rather straightforward and unsurprising result, one that certainly does not admit chaos: one-dimensional iterative maps can exhibit a much broader range of behaviour then the corresponding one-dimensional ODE. Yet the map is only one of the possible ways to discretise the logistic equation, some of which produce quite different results. Behavioural richness of this particular version, the logistic map, was first noted in May [1976].

The logistic map f is a one-dimensional dissipative system displaying the period-doubling route to chaos. For $0 \leq r \leq 4$, $f : [0, 1] \to [0, 1]$, and for $r > 4$ the trajectories are no longer confined. If $x_{n+1} = f(x_n)$,

$$x_{n+1} = rx_n(1 - x_n). \tag{1.22}$$

For small r the motion is periodic. With increased r the periodicity successively doubles until what is known as the period-doubling accumulation point at r_c< 4. The underlying pitchfork bifurcation produces unstable periodic points, making the attractor at r_c be nowhere dense. It can be shown that then the attractor is a Cantor set, with a variety of computable fractal dimensions (see for instance Grassberger and Procaccia [1983a], Grassberger and Procaccia [1983b]). At $4 > r > r_c$ motion is confined to chaotic bands. These then merge in a symmetric way until at $r = 4$ the attractor fills $[0, 1]$ and motion is mixing, in the terminology of Collet and Eckmann.

Figure 1.2 shows the bifurcation diagram. On this scale it would not matter if it was produced by following single trajectories, or taking a number of certain initial conditions and recording the iterates at a specified time. The only persistent feature of the map is the clock. Chaotic motion conforms to this by making every T^{th} iterate be located in the same band (if T is the number of bands), but leaves the location of the point within the band to be varied with a certain positive Lyapunov exponent $\lambda(r)$. Lorenz called this motion 'noisy periodicity'.

Figure 1.3 shows the variation of the Lyapunov exponent (of which there is only one, since the system is one-dimensional) across r. The gaps where $\lambda(r) = 0$ correspond

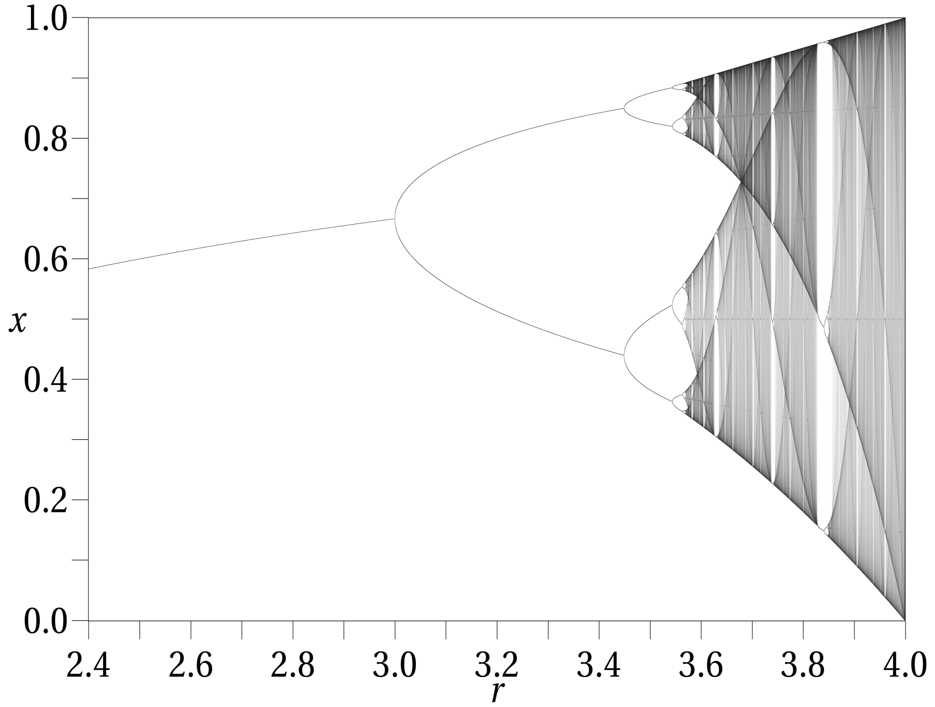


Figure 1.2: The standard bifurcation diagram of the logistic map. For lower values of r the trend continues, the attractor x having a periodicty one (source: wikipedia).

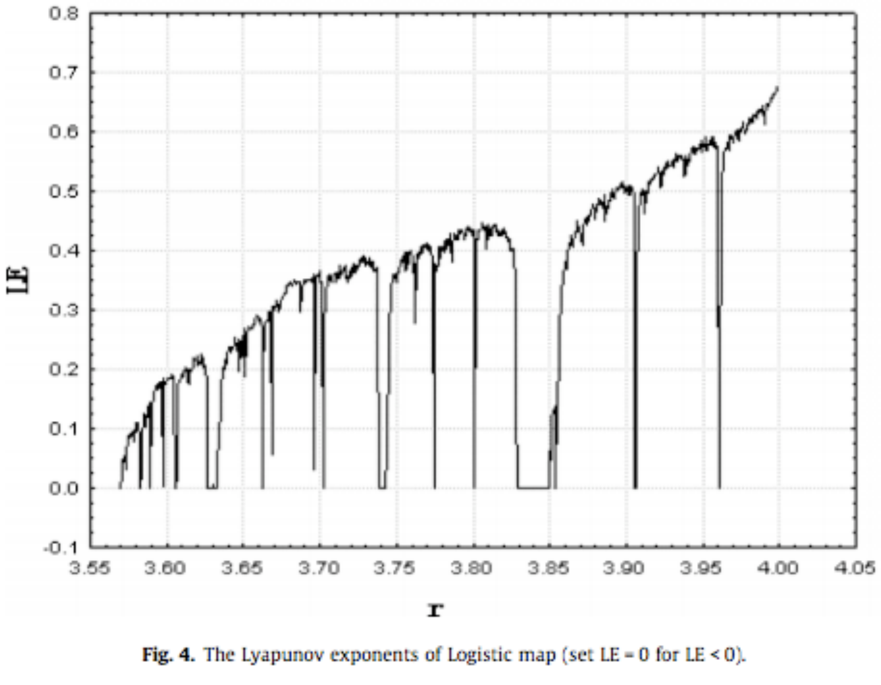


Fig. 4. The Lyapunov exponents of Logistic map (set LE = 0 for LE < 0).

Figure 1.3: The Lyapunov exponent of the logistic map (taken from Luo et al. [2009]).

to windows of regular motion. On the bifurcation diagram these are regions with distinct lines. The biggest window is around $r = 1 + \sqrt{8}$, where the attractor is a period-3 limit cycle. These bursts of periodicity happen at all scales of $r_c < r < 4$. Moreover, they do not necessarily then lead to chaos in the same way. Period-tripling, and other combinations and mergers can be detected if only the r resolution is large enough.

The logistic map lies in the broad class of one-dimensional unimodal maps which all share the qualitative features of the bifurcation pattern (to be more precise, through kneading theory they can be shown to be topologically equivalent). These maps are projections of higher-dimensional systems to lower planes, and as such are not invertible (for example through having several of higher-dimensional orbits happening to have an equal coordinate). One of the reasons behind their generality is that often the dynamics of these original systems happens only on a small subset of the state space, and as such motion can effectively be described by simpler lower-dimensional maps. The general theory of 1D maps is limited: it is for instance not possible to *find* all ranges of (to use our example) r corresponding to motion of a particular type. Something similar is possible in reverse (Singer [1978]): satisfaction of a certain condition on the Schwarzian derivative (a function of the derivates of various orders) can demonstrate a limit on the number of stable periodic orbits. The opposite means the attractor is either infinite (a cantor set), or motion is mixing with all the traits of chaos. In this respect the logistic maps belongs to the class of maps with an everywhere-negative Schwarzian derivate, labelled S-maps.

The existing general result concerns the types of motion possible, and is in fact the reason the logistic map displays both mixing, periodic and 'ergodic' (infinite attractor) behaviour. It is that subsets of r that result in these three motion types are all of positive Lesbegue measures. Another interesting result is the Sarkovskii sequence, which says that if an observed period is present in the given sequence, then the system also has motion with arbitrarily long periods. The lowest periodicity in the sequence is three, which is exactly the value mentioned above for the logistic map. This result also proves that an infinite range of other periodicities can indeed be detected. In fact since for low values of r the period doubles, it implies that the periodic windows (which do not have to have period equal to 2^n) can be infinite in number. That is indeed the case. In fact the sequence does not limit the number of windows with the same period.

There are three main routes to chaos present in the logistic map. It is in a universal class of systems defined by flip bifurcations. Periodicity doubles every r_i, leaving behind unstable fixed points, so that if

$$\delta_i = \frac{r_i - r_{i+1}}{r_{i+1} - r_{r+2}} \tag{1.23}$$

then the Feigenbaum constant $\delta_\infty = 4.6692$ defines a certain class of maps. Another way chaos sets in is through intermittency. This is a direct effect of the tangent bifurcations that are the underlying reasons behind the attractor suddenly turning periodic. This process leaves trajectories for some finite time stuck near specific points. This is the effect that makes us see the pattern of folded shadows in the bifurcation diagram: these specific regions are exactly ones which, after a small increase in r, become the stable periodic limit cycles. Inside these periodic windows after periodicity increases (in a manner that is not necessarily doubling the period) noisy periodicity occurs again, until an ‘explosion’ happens. This - or the ‘interior crisis’ coined by Grebogi - is the sudden jump in the size of the attractor.

At $r = 4$ under a change of variables the motion is equivalent to the Bernoulli shift map (bit shift map) (and also to the behaviour of the Tent map at $\mu = 2$, see later section), given by

$$x_{n+1} = 2x_n \mod [1] \tag{1.24}$$

If we represent x in binary form then points are sequences composed of two symbols. Iterations can then be viewed as shifting the sequence (which is to the right of the decimal point) one step to the left. One of the ways in which this shift in framework is useful is in how it helps to understand the effects of chaos, represented in the logistic map by mixing motion. Chaos is often characterised by sensitive dependence on initial condition. In practice this means that *finite* information about an initial condition will soon be lost. Any finite information is represented by a finite binary string. Hence after the number of iterations becomes greater than the length of the initial string no information about the original string would be left. More exactly, if two trajectories differ by some finitely-specified amount, there is a time after which this difference would be nullified[12].

This is one the reasons we use chaotic dynamical systems in our study of how information gets preserved across time. We do not view chaos as *the* emergent phenomenon; we are only partially interested in its phenomenology. From the perspective of this work chaotic motion

[12] Initial conditions that are rational numbers would thus be repeated ever finite number of steps, since their binary expansion contains repeated regions that will get moved forward. Irrational number are dense; hence chaotic motion at $r = 4$ is simply more ‘likely’.

merely serves as a mechanism that after a finite time makes computing the true final state impossible. Note, however, that that does not mean that we cannot say anything about where trajectories are likely to end up. The invariant measure is a beta function and is not flat. That means that independent of the initial condition there are guesses about the position at some arbitrarily far future, guesses which are more likely to be correct than not (for same subset size). Accordingly, in our investigations we focus not on prediction but on 'forecastability' (the difference is clarified in the section on PMI).

As such the interesting features we find stem from other persistent features of the system, or from a variety of motion, not just chaotic; or else from the different ways in which chaotic motion can happen. The latter two are explored by a different system which we give in the section below. Unlike the logistic map it is not dissipative but rather admits coexistence of various types of trajectories, exhibiting a different route to chaos and is thus accompanied by a range of new phenomena.

1.3.2 The Standard Map

The standard map, also sometimes called the Chirikov standard map, was considered by Bryan Taylor, and introduced by Boris Chirikov in Chirikov [1979]. A two-dimensional area-preserving map with a single parameter, it is a Poincaré cross-section of a Hamiltonian system that demonstrates the now-classic route to the onset of chaos described by the KAM framework. As such it has been found useful in such a wide variety of situations (see Zaslavsky [2012]) that its common name has come to reflect its applicability. The classical interpretation of the associated Hamiltonian system is that of a kicked rotor. The quantum version of the Hamiltonian behind the map is used to test the Anderson Localisation.

The map is paradigmatical in its demonstration of Hamiltonian chaos (according to Cambell [1987], it plays the same role for Hamiltonian chaos the logistic map did for chaos in dissipative systems). What makes this map so tractable as a toy model is that there is only one parameter that essentially controls the system regime. The fact that the map is iterative also means computations can be performed relatively fast, with potential errors stemming only from numerical approximation and not the necessarily inexact solver algorithms.

The standard map is given by

$$p_{n+1} = p_n + K \sin\theta_n \tag{1.25}$$

$$\theta_{n+1} = \theta_n + p_{n+1}. \tag{1.26}$$

Without loss of generality we take K to be positive, and since here we will be considering the dynamics on a torus, both variables are confined to the fundamental domain $[0, 2\pi]$, and taken $\mod [2\pi]$. A negative K corresponds to a translation of angle to $[-\pi, \pi]$, and graphically it merely shifts the position of the main structure surrounding the stable fixed point. The map is reversible and has a number of symmetries.

The extent of chaos increases with K, so that at $K = 0$ all the orbits are either periodic or quasi-periodic, and at $K = 2\pi$ the system is ergodic, at least on the level of available resolutions (finding the measure of these islands of regular motion for large K is one the open problems - see Sinai [2010]). We will restrict our interest to $0 \leq K \leq 2\pi$.

The original Hamiltonian for the kicked rotor, with kicks of strength K, is

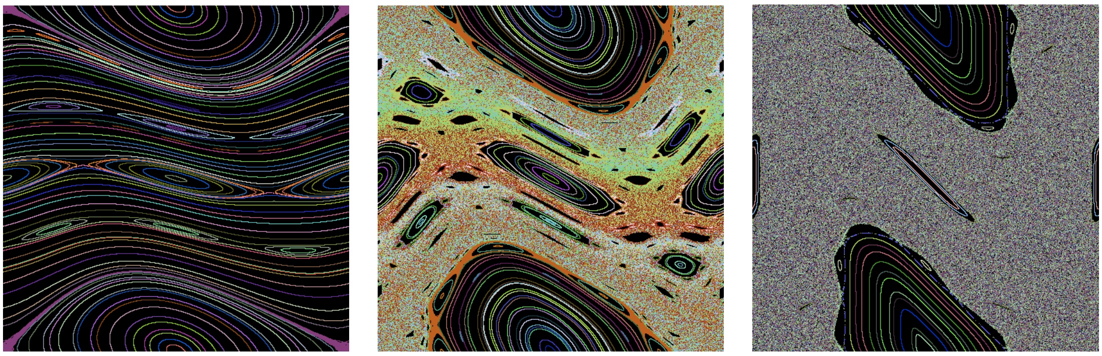
$$H(\theta, p) = \frac{1}{2}p^2 + K \cos\theta \sum_{n=-\infty}^{\infty} \delta\left(\frac{t}{T} - n\right), \tag{1.27}$$

where p is the canonical momentum, and δ represents instantaneous kicks at frequency $2\pi/T$. It is clear that while θ is continuous throughout, p gets changed by a finite amount. Therefore one can look at the Poincaré plane defined by the t just before successive kicks. These difference equations are equivalent to the standard map, and can be derived from Hamilton's equations associated with eq.(1.27)

In this respect the state space of the standard map can be interpreted as the phase space of the Hamiltonian, and momentum p and angle θ as polar coordinates of the trajectory as it goes through the Poincaré plane.

The range of map behaviour is demonstrated in figure 1.4 that traces the evolution of a number of trajectories for three different K. Broadly speaking, circles correspond to regular orbits and absence of structure indicates chaos. These graphs show one of the more striking (Zaslavsky [2012]) features of Hamiltonian chaos - the coexistence of regions of regular and chaotic motion.

This dependence of motion type on the initial condition is made possible by the lack of



(a) $K = 0.6$ (b) $K = 1.1$ (c) $K = 2$

Figure 1.4: Evolution of a number of trajectories using the standard map with increasing K (here the axes are (θ, p), $-\pi \leq \theta \leq \pi$). Orbits are tagged by colour. Notice that at $K = 2$ no single area (apart from maybe near the resonances) is dominated by a single trajectory. This is not the case at $K = 1.1$, where for large enough times trajectories are still seen to stick in subsets of the broad chaotic area. See figure 1.5 for more of this effect.

attractors. The volume (say the set of trajectories) does not contract to a small subset of the initial state space. Hamiltonian systems by definition conserve energy, or the phase space volume, which in terms of the standard map translates to area-preservation. Varying K therefore changes the general type and the specifics of motion given by an initial conditions. Thus the absence of kicks modelled by strength $K = 0$ renders the original Hamiltonian integrable. Just by looking at the equations shows that this is because momentum is now a conserved quantity (along with energy). If θ had not been confined the system would simply be describing free motion. As it stands the invariant manifolds are described by circles, each defined by a winding number $\omega(p_0) = p_0$:

$$p_{n+1} = p_0 \tag{1.28}$$

$$\theta_{n+1} = \theta_0 + p_0 n. \tag{1.29}$$

This regular motion, which involves trajectories confined to horizontal lines on the phase diagram, comes in two types. If ω is rational then after a finite number of iterations the trajectory begins to retrace its steps. Thus in periodic motion for some initial angle the horizontal lines fill in to a greater extent (with smaller gaps) depending on the specifics of ω. They do so without any gaps, densely covering the circle, if ω is irrational, in which case the motion is quasi-periodic. Hence $(0, 0)$ is a fixed point, every point on the $\omega = \pi$ is a period-2 fixed point, etc.

As K increases by a small amount some fixed points disappear, and the winding number is no longer equal to p_0. For example at $\omega = \pi$ the $\theta = 0$ and $\theta = \pi$ are now stable fixed points between which lie hyperbolic fixed points. Point stability can be tested by comparing the trace of the Jacobian to 2 in order to compute Greene's residue. The stable fixed points become surrounded by elliptic orbits, and the hyperbolic fixed points are associated to hyperbolic orbits and thin stochastic bands. All these have an associated periodicity so that ellipses around the period-two fixed points are populated by trajectories alternating between them at every time step. Thus the horizonal frequency of these elliptic islands can easily be predicted. These ellipses come in what can be described as 'islands', or resonances. Circles associated with periodic motion - rational ω - typically break down first, at $K = 0$. According to the Poincaré-Birkhoff theorem for every $\omega = m/n$ there will be at least two periodic orbits left, with period n (Meiss [2005]). This appears as n islands, the chain called a resonance. At least one of those will be on the $p = 0$ line, the 'dominant' symmetry line (ibid.). As K increases new elliptic orbits are created around each elliptic orbit based on the associated ω. Thus structures form on all scales, though this is still not proven. In terms of universality, MacKay [1983] used renormalisation group techniques to show that the island structure around the golden curve is the same for all smooth maps (twist maps). The arrangement of islands of periodic motion is non-trivial. A single chaotic orbit will encounter obstacles on all scales, which corresponds to there being a specific distribution of island sizes. The area occupied by a single chaotic orbit will be finite (Umberger and Farmer [1985]), turning the orbit into a 'fat fractal'. If it is computed by for example breaking up the state space and counting the visited squares then this number will have definite scaling regime with resolution. The reverse holds too and the regular motion also occupies a finite area (Cambell [1987]). Growing K is generally associated with deformation of the horizonal lines, or rotational circles (the circles *seen* as circles in the state space do not actually encircle a torus, and are called librational circles). As these encroach on each others' spaces the stable manifold of one crosses the unstable manifold of the other in a 'resonance overlap'. This produces a homoclinic intersection, and therefore an infinity of homoclinic intersections. Partially motivated by the study of motion in plasma, Chirikov [1960] computed the criteria for the overlap of the resonances. If the state space is viewed as a cylinder then the destruction of the final barrier allows the 'particle' to escape, i.e. momentum to increase indefinitely. This gives an estimate of some $K = K_c$.

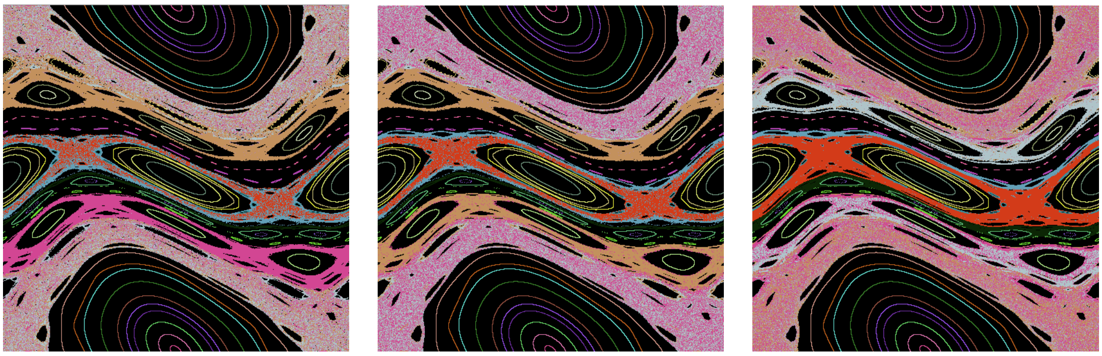
It is possible to determine existence of a rotational circle by looking at convergence of residues of the orbits remaining after the destruction of the m/n orbit MacKay [1992].

Thus with increase in K fewer and fewer circles are left. This relationship between the winding number associated with the remaining circle and the K value can be made exact. Works such as Black and Satija [1989] show the 'fractal' nature of this dependency. The last circles to be destroyed correspond to ones with $\omega = \gamma \pm m$, $m \in \mathbb{Z}$, where γ is the 'most' irrational (the criteria assigning the extent of 'irrationality' is related to the asymptotic tails in the fraction expansion) number, the golden mean. MacKay and Percival [1985] proved that no circles are left for $K > 63/64$. We use the notation K_g to denote the exact point of the breakdown of the golden circle. Although no analytic expression exists, numerically it is found to be $K \approx 0.97$. $K_c \geq K_g$, and the two values are usually associated.

All the above is usually phrased in terms of flows in the state space of the original Hamiltonian, so that invariant circles are cross-sections of the invariant tori, called the KAM (Kolmogorov-Arnold-Moser) tori. The KAM theorem is then exactly the statement about persistence and breakdown conditions of these KAM tori (and hence cantori). Also in this framework K can be viewed as perturbation away from integrability, in at least one meaning of the word.

Transport in the Standard map Stochastic motion occurs between the invariant rotational circles. A region that is bordered by them and containing nothing inside to limit the chaotic motion is called a 'zone of stability'. Mather [1991] showed the existence of orbits that get asymptotically close to the regions' borders. These regions may be difficult to pinpoint when $K \approx K_c$ since then the structures are self-similar and appear on all scales. According to the Aubry-Mather theory irrational winding numbers are associated with trajectories dense on either the circle or a Cantor set. Since the circles stop existing after some finite K, it follows that what remains must become a cantor set. These 'cantori' will thus contain holes which then admit movement to the other side, and chaotic trajectories can pass through.

Figure 1.5 shows the consequence of this method of freeing up the space. Since passing through the obstacles that are cantori is difficult, there are time scales (possibly location-dependent) at which trajectories are essentially stuck in specific regions. While there they mimic the rotational motion that characterised those regions before the circle breakup. MacKay et al. [1984] showed that the local flux of trajectories through a cantorus



(a) t_1 (b) $t_2 > t_1$ (c) $t_3 > t_2$

Figure 1.5: The same run of the standard map showing the evolution of trajectories at $K = 0.971635$ up to some t_i. As before a trajectory has its own colour, and the colour of a pixel is determined by the same orbit sequence across all pixels. Therefore if there are two areas that change in colour, but are at some time coloured differently with no mixing, it means that there is a time period in which at least one signed trajectories is *not* entering a particular subset.
Notice how occupation of the different areas of the graph fluctuates, the most uniformly colour areas being near the separatrices - and the distinct change in colour of the two bands that appear to be symmetric about the golden circle, which lies roughly in the middle.

can be written as

$$\Delta W \propto (K - K_c)^a \,, \tag{1.30}$$

$a \approx 3$. This is roughly in line with the prediction in Chirikov [1979] expressed in terms of *time* of transitions between regions. These results can be expressed in terms of the diffusion coefficient, calculated using the Fokker-Planck framework in which it makes sense to consider passing through a barrier as a probabilistic phenomenon.

A global picture with analysis integrating diffusion across the different trajectories suggests anomalously slow relaxation due to the cantori. Poincaré recurrences (Chirikov and Shepelyansky [1999]) and for instance the number of trapped particles in a region then decay algebraically in t. In Bensimon and Kadanoff [1984] there is an algebraic decay in the escape area with n at K_c.

1.4 Quantifying Complexity

The existing measures of complexity and emergence all vary depending on the mathematical framework one considers, the system in question, and of course on what one's intuitive notions about the extent of 'emergence' in this system are. Broadly speaking there are measures based on single realisations and multiple realisation, or ensembles. To the former category belongs the algorithmic Kolmogorov complexity (see below). We hold the view that randomness should not be equated with complexity or emergence, and so our proposed measure is a function of probabilities. As seen above, in that setting the order-disorder relation is usually phrased in terms of entropies, which is exactly our aim.

The measure in existing literature to which our function comes closest is Effective Measure Complexity (EMC, otherwise known as the excess entropy). This quantity is usually conceptually twinned with the entropy rate, in the sense that defining one can define the other, and certainly understanding one helps with having a clear picture of the other. Entropy rate was already introduced in eq. (1.16) in the context of sequences. Thus EMC and entropy rate are measures of systems with a discrete alphabet (and by extension discrete time). Excess entropy is usually applied to sequences obtained from symbolic dynamics or probabilistic cellular automata, whereas the EMC incarnation (the original) was studied in formal languages and grammar.

Symbolic dynamics then looks at the map as a potential means of randomisation. The same notions can be defined in terms of continuous state spaces to obtain metric entropy and its measure-free counterpart, topological entropy. These are standard quantifiers in dynamical systems theory.

In our work we use data from dynamical systems without any discretisation. There is still a notion of resolution, but that is now related to the depth of sampling, and is therefore phrased in terms of varying the measures of subsets rather than their linear size. Our initial aim is to test a function that could detect shared information between the past and future of a distribution over the attractor, with a variable time gap, and understand what features of the system would qualify it for being labelled, in this definition, as 'strongly emergent'.

In the following section we set the mathematical context of the various quantifiers of order and disorder mentioned above, and then review the toy models that we will use. These are the logistic map - a one-dimensional dissipative dynamical system, and the Standard

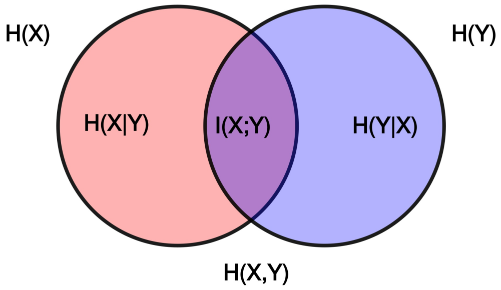


Figure 1.6: Venn diagram where the extent of overlap of two variables measures the degree of their interdependence on each other, in other words their Mutual Information (denoted here by $I(X;Y)$. $H(X,Y)$ is the joint entropy of X and Y. Source: wikipedia.) Some systems that we will study have the constant marginal entropies, which means that looking at the cross-section is equivalent to looking at the joint entropy of the system.

map, a two-dimensional area-preserving map. Both are maps with one-parameter, changing which affects the extent of chaos in the trajectories. Chaos is a standard setting in which to talk about 'weak' emergence. In general in dynamical systems attractors are sometime said to 'emerge' as parameters are varied. We view chaos as a mechanism that results in the loss of initial information. In the language of dynamical systems that is described (and defined) by the rate of exponential divergence of nearby trajectories, and the quantity that measures it is related to the metric entropy. Yet in the section below it will be seen how the *initial* motivation behind the various entropy-related concepts in dynamical systems was actually at least partially pure information-theoretical. Kolmogorov, who developed some of these concepts, also worked on information - Kolmogorov [1965]. In that area, apart from introducing the aforementioned algorithmic complexity measure taken up by G. Chaitin, he also stressed the importance and use of mutual information. His method was not probabilistic - it was simply to count the proportion of filled squares in the joint distribution, which implied the setting of a sequence along with uniform measure. Mutual information and its variants are one of the primary measures of choice for nonlinear correlations, at least partially because it can be understood rather intuitively in terms of the 'extensive'-entropy framework - see figure 1.6. Using it to quantify various complexity concepts is in line with the tacit understanding that the emergence can be viewed as some form of interdependency between the variables, that is not present in 'simple' systems.

These correlations can be searched for among more than one variable. For a example adding an conditional dependency of the variable pair would give the Conditional Mutual Information. This measure is sometimes used to infer network structure, as for example is done in the recent work by H. Jensen's group (Jensen and Razal [2012]), who were looking at electroencephalography data to investigate causal connectivity of the brain.

Conditional mutual information can itself be extended as a measure of stochastic interdependencies on all scales. Considering subsets of all sizes it can be shown in Studeny and Vejnarová [1998] that this is a valid way to decompose multi-information. Multi-information, or the 'total cohesion function' looks at the difference between the joint entropy and the sum of marginal entropies, except for, unlike in the case of mutual information, the marginals are now defined on the underlying power set. Approching this problem from the information geometry point of view, Erb and Ay [2004] motivates multi-information by showing it can be decomposed into a sum of mutual information between distributions that differ only by the extent to which their marginals agree. The authors prove that in the thermodynamic limit the multi-information for the 2D Ising model is maximised at the phase transition, whereas the 1D Ising model shows only a steady increase with β. However the same can be shown by simply considering mutual information between two neighbouring spins (Matsuda et al. [1996]). Yet the idea of measuring level-specific dependencies can be extended to give a vector-valued measure of complexity. The motivation, according to Kahle et al. [2009], is to "quantify complexity by measuring how far it is from being reducible to a theory of k-interactions." Consider the distance (here using the KL metric) between the original distribution and the set of distributions generated by a Hamiltonian with only k-particle terms. An element in the complexity vector is then the difference between the k^{th} and the $(k-1)^{th}$ such distance. It represents the optimal improvement in understanding the system by including interactions one order higher that were not present in the original k-order subsystem. This is very much an ongoing research, since computation is costly; moreover, results for the couple chaotic systems (Galla and Gühne [2012]) still require clear interpretation.

The lack of symmetry in 'distance' measures computed using the KL metric may be a conceptual impediment to clarity of definition. It is possible to move away from desiring the what is essentially a distance quantifier to have the clear interpretation of relative entropy, and consider proper metrics. For example in MacKay [2009] a range of metrics are compared through behaviour with respect to parameters, and a solution, Dobrushin metric, is proposed as a candidate. MacKay also proposes definitions of emergence in terms of space-time phases (Diakonova and MacKay [2011]). Here, emergence is the Dobrushin distance between a phase and the product measure of individual components, whereas since strong emergence is characterised by more than one phase, measuring it could be a matter of evaluating the diameter of the set of phases.

1.4.1 Some Probabilistic Order/Disorder Measures and Related Quantities from Dynamical Systems.

Probabilistic Measures v Algorithmic Complexity Chaitin [1975] gives a nice motivation for defining complexity in terms of pattern. Consider tossing a coin a (large) number of times. We would expect to see a sequence of heads and tails with no discernible order. If we achieve a perfect alternating sequence we would be surprised - it seems that outcomes could be predicted. Yet both of these outcomes have an equal chance of coming up. We therefore want to distinguish them not via some source, but simply by considering them as given, and looking for one with the most pattern, or predictability.

Algorithmic complexity, or Kolmogorov-Chaitin complexity, is the length of the shortest computer program that could produce the sequence. If there is absolutely no discernable patter, no way of compressing the sequence, then the shortest program would simply be given the values themselves, and its length would be the length of the sequence. This is how randomness is here defined - probability in the Pascal sense does not come into it at all.

Consider tossing a coin and obtaining a sequence of just heads. This will happen with the same probability as any other sequence. The point is it is easy to store this outcome in our head, just as it is easy to store an alternating sequence: they would all be distinguishable. But complicate the sequence by reversing a few outcomes and already the result would be hard to memorise, and hence hard to compare with the result obtained if we had reversed another subset of tosses. It is arguably easier to memorise predictable patterns, so when thinking about algorithmic complexity the notion of a reproducible algorithm could be substituted for ease of commiting a sequence to memory.

It should be noted that difficulty in forecasting is not necessarily related to correctness, or possibility, of the forecast result. As Grassberger notes, a random string is impossible to predict correctly, but the best prediction is just guesswork. Correspondingly, Bennett [1988] introduces the concept of *logical depth*, which measures the effort taken to make a prediction (contrast it with Kolmogorov Complexity, which measures the amount of information associated with recreating system output). The drawbacks of using algorithmic complexity as a measure of our intuitive understanding of the concept was mentioned by

Atlan whilst introducing the concept of self-organisation (Atlan [1986]), and in Huberman and Hogg [1986]. Grassberger [1986] motivates the need for a statistical description of complexity based on the fact that many observables in physics are statistical by their nature, such as temperature and pressure. The ideal gas situation, which arguably is an excellent example of emergent phenomenon in the sense of possessing a small number of observables capable of accurately describing the general properties, would thus be untractable using the deterministic, algorithmic notions of complexity above. Grassberger puts forward the notion that our intrinsic processing is also done in terms of ensembles (this made it possible for the recent learning algorithm by Google to identify the concept of cat without knowing what to look for - see Le et al. [2012]). A solution to Chaitin's example is thus that although all three sequences would have the same probability of being produced, the fully random example would be indistinguishable from another fully random one (to use our language, because it would be harder to memorise), so when we mention the complexity of something we are actually defining an ensemble of observations (see below).

Statistical Complexity and ϵ-machines

Grassberger [1986] mentions that Kolmogorov's complexity is an intuitive quantifier of randomness, not complexity. A complex system is thus a system producing a pattern somewhere between perfectly ordered (trivial) and completely random (P.G. actually uses the standard map as an example) - see figure 1.4.1. In a later paper Crutchfield and Packard [1983] back this notion up, mentioning that while the ordered case is entirely predictable, the random one admits a compact physical description, and hence complexity lies somewhere in between on the spectrum of predictability. The two extremes are both *computationally* simple - and hence what one needs to consider is a measure based on the system's internal computation. Note that to treat these two examples as computationally simple we need to assume a distribution over ensembles, which means the second case is a matter of using a random number generator; otherwise the Kolmogorov complexity would be high for the second case.

Thus statistical complexity requires the ensemble to be reduced in some systematic way. One such measure uses symmetry based on predictive properties. Specifically, Grassberger [1986] sees the system (stationary, discrete) as a formal language. The rules defining combination of symbols from some alphabet is a "grammar", and the probabilities over

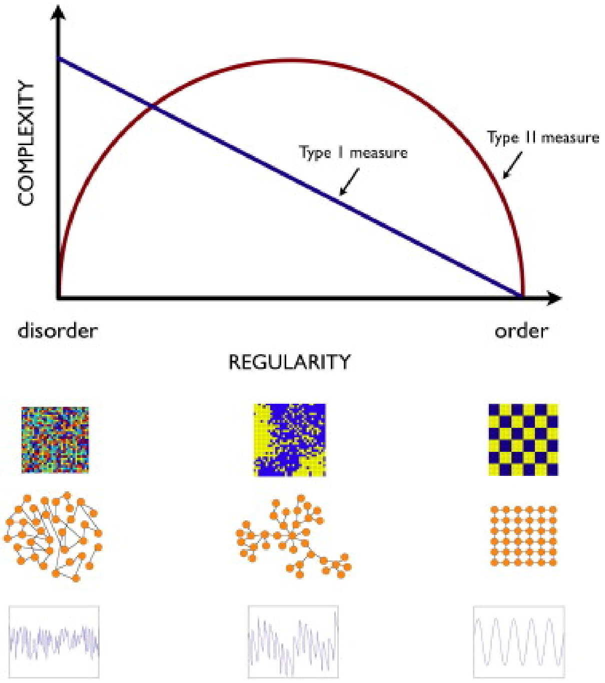


Figure 1.7: Intuitive understandings of various complexity measures, where type II is the one under investigation here, and an example of a type I measure is algorithmic complexity (taken from Parrott [2010]).

"words" (strings of arbitrary length) a "style". The resultant reduction is a deterministic finite automaton (DFA). The former adjective refers not to the process, but to the representation.

DFA is a graph where nodes are states defined by the same probability of future outcomes. From each node to another node are links that stand for the possible symbols. A path on this graph is thus a word, and the probability over the links defines the style. Minimal graph corresponding to a language is one with the smallest number N of nodes. Grassberger in (ibid.) called $\log(N)$ the algorithmic complexity AC, but notes that in Wolfram [1984] the same quantity is called the complexity of regular languages. The deterministic part refers to the property that a labelled link stemming from a node is associated with only one other node. Grassberger also refers to these structures as Unifilar Hidden Markov Chains (UHMC). Attaching a frequency measure $p(i)$ to nodes that are now labelled by i, statistical complexity is defined by

$$SC = H[p]. \tag{1.31}$$

Here we assume stationarity, and associate p to the stationary measure.

It also clear that $SC \leq AC$. Both SC and AC play roles similar to topological and metric entropies. The former counts the number of distinct sequences normalised by their size, whereas AC measures the number of nodes on a graph. Grassberger mentions that if the grammar is known then the minimal graph, and hence AC, can be found. However a minimal graph does not necessarily correspond to a graph with the smallest SC. Larger graphs can produce nodes that are visited less frequently. For the purposes of computation, i.e. complexity, the support space should be allowed to become large, so that the only variability is the measure - that way the universal computer would have a setup that allows for easier execution.

Crutchfield and Young [1989] reinvent statistical complexity in the context of dynamical systems by discretising the state space and building a language using symbolic dynamics. The DFA construct is labelled the ϵ-machine, and complexity of the original system is defined in terms of the processing capacity of the ϵ-machine, which is phrased in terms of the complexity of the graph; that, in turn, is expressed in terms of generalised Renyi entropies C_n of the asymptotic vertex probabilities $p(i)$. Using the same notation as above, $C_0 = \log(N)$ is clearly the AC of Grassberger, yet here it is named **probabilitistic algorithmic complexity**, and C_1 is just SC. However, all of this is heavily dependent on the original partition (the examples used a simple halfway cutoff point in the state space of the logistic map).

Modelling a system in terms of what are effectively causal states (evident even in Grassberger's description) represents accessing the structure of a system's intrinsic computation. It is in this context that questions about internal processing, the memory required by the system to statistically reproduce a state, the information storage and transfer, are answered (Feldman and Crutchfield [1998]). This is the basis of the broad designation of the work around the field of statistical complexity as Computational Mechanics.

A sequence of only 1s will have one node in the DFA (or probabilistic finite-state machine, a term used by the Santa Fe school), a period-two sequence two nodes. Statistical reproduction means that distributions over subsequences of any length are the same. Consider a sequence of alternating ones and zeros, and two graphs (fig.1.4.1). The first has one node and two circular links to itself, each representing different symbols, and each associated with probability of a half. The second machine has two nodes, with two links forming a loop, each corresponding to a symbol but this time with a probability of 1. Even though the first graph is minimal, only the second will have reproduced the probabilities over subsequences, a distribution giving zero on anything non-alternating. A random sequence would have only

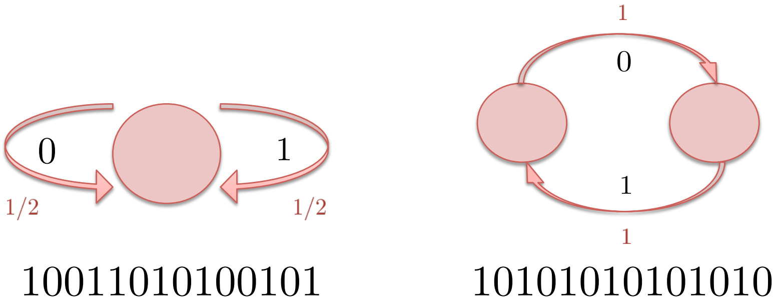


Figure 1.8: Illustrations of DFA and samples of their output. Note that nodes do *not* necessarily correspond to the different symbols. The nodes are causal states, the numbers in black are outputs, and the smaller numbers in light red are the probabilities.

one node but two links. Thus AC would be small for both the entirely random and the extremely memorable (not diversified) sequence, and would increase with the periodicity (or pattern), behaviour detailed in section 1.4.1.

Crutchfield and Young [1989] and their later works detail the method for constructing the ϵ-machine (for reviews see works by D. Feldman). The construction method utilises predictability of equivalence classes. A method was given by P.Grassberger in Zambella and Grassberger [1988], and later in various publications by C.Shalizi (e.g. Shalizi et al. [2004], Shalizi and Shalizi [2004]). The latter algorithms assume the UHMCs do not retain the transient states, though without metioning it explicitly.

Metric and Topological Entropies

Metric and Topological Entropies in Symbolic Dynamics The most straightforward way of understanding topological entropy is through the n-cylinder framework (Parry [1964]; see also Crutchfield and Packard [1983]). Consider a symbolic dynamical system (Σ_F, σ) induced by $F : X \to X$ on the partition P (now we use P in the sense of partition, not partitioning) as described above. We define an n-cylinder equivalence class on Σ_F by comparing the first n symbols. Label the elements of the resultant partition of Σ_F according to the first n elements of the sequences belonging to that equivalence class: let the n-cylinder s^n be a set of (permissable first n) symbols $(s_0^n, ..s_{n-1}^n)$, in other words, the set of all admissable n-element sequences.

This equivalence class also induces a partition on X. To each n-cylinder corresponds a set of initial conditions $x \in X$, each the (practical) start of an orbit whose first n symbols in a corresponding sequence are the same as of the n-cylinder - or, put more simply, the set of all initial conditions resulting in a particular n-symbol sequence for its first n symbols.

An n-cylinder partition on X is a way of grouping elements of the state space into sets that result in similar orbits. The extent of similarity is expressed by the n integer (all of this, of course, has an intrinsic dependence on P). Increasing n is a *refinement* of the partition in the string sense, i.e. the cardinality of this partition cannot decrease.

In fact the cardinality of this partition, in other words the number of different equivalence classes given a refinement n, $N(n)$, is essentially a measurement of the number of trajectories distinguishable using P and n. Given some P, the number of equivalence classes given the limiting case of infinite refinement corresponds to the ultimate number of trajectories distinguishable with partition P. If it were postulated that the number of such trajectories grows in a specific fashion, namely exponentially, then the rate of growth can be written as

$$h_\sigma(P, F) = \frac{\log N(n)}{n}. \tag{1.32}$$

This limit is proven to exist (Parry [1964]). Maximising $N(n)$ over the partition gives

$$h_\sigma(F) = \sup_P h_\sigma(P, F). \tag{1.33}$$

$h_\sigma(F)$ is a function of the number of maximal number of different trajectories that can be found by partitioning the state space.

If we use this dynamical system to send signals (i.e. as an information source), how many distinguishable messages will there be? We can associate an orbit to a signal. The reception has errors, so each point will be decoded with an error. The size of this error is related to the strength of the coarse-graining. A message, or orbit, is thus transmitted as a sequence of symbols from the finite alphabet (whose size depends on the error magnitude). Given this setup, we ask the question of how many messages will the receiver be able to distinguish. Since we do not put a time limit on it, the number of iterations is taken to infinity. $h_\sigma(F)$ then measures the exponential rate of growth of the number of discernable messages. Its motivation is the same as that behind topological entropy.

Kolmogorov and Tihomirov [1959] linked information-theoretic considerations to an arbitrary set A in a metric space (with some conditions on compactness) by either viewing A as a set of all possible messages, or all possible signals. This leads to two parametric frameworks with three main notions, all functions of error - or effective coarse-graining strength - ϵ. It was mentioned in Adler et al. [1965] as being the inspiration behind the notion of topological entropy.

If A is the set of messages, then any $x \in A$ is considered recoverable from another point at most ϵ away. Thus every point in a neighbourhood is obtainable from some reference point.

- The ϵ-spanning set is a set such that every element of A is *at most* ϵ away from some element in the set. Let $\mathbb{N}_\epsilon^R(A)$ be the minimal cardinality of ϵ-spanning sets of A. Define $\mathbb{H}_\epsilon^R(A) = \log_2 \mathbb{N}_\epsilon^R(A)$, which is thus the (number - 1) of different binary signals there should be in order to recover any element of A given some embedding space R. This is called the ϵ-entropy of A w.r.t. R.
- An unreferenced notion of ϵ-entropy of A, $\mathbb{H}_\epsilon(A)$, is obtained by constructing an ϵ-cover, a cover of A by a collection of sets with diameter not greater than 2ϵ. Let $\mathbb{N}_\epsilon(A)$ be the smallest number of sets in an ϵ-cover of A. ϵ-entropy of A is then $\mathbb{H}_\epsilon(A) = \log_2 \mathbb{N}_\epsilon(A)$.

Treating A as a set of signals is equivalent to the framework where any point in a neighbourhood is associated with some reference point (signal). In effect we have an agreed-upon alphabet, similar to a coarse-grained one. The question is then how many distinguishable signals are recoverable? This is answered through the notion of an

- ϵ-separated set, that is, one in which any two points are at least ϵ away from each other. The number of all possible distinct signals is then the maximal cardinality of an ϵ-separated subset of A, $\mathbb{M}_\epsilon(A)$. Its logarithm is the ϵ-capacity of A, and is equal to the length of binary signal that we wish to transmit using the signals available in A (with a subtelty about adding 1).

In (ibid.) the authors' main results are then the relations between the notions.

Adler et al. [1965] introduced topological entropy of a dynamical system through refinement. Refinement of two covers $\mathbb{U}$ and $\mathbb{V}$ is

$$\mathbb{U} \vee \mathbb{V} = \{U \cap V : U \in \mathbb{U}, V \in \mathbb{V}\}.$$

If $\mathbb{U}$ is a subcover of minimal cardinality of the above set X, $F^{-n}\mathbb{U} := \{F^{-n}U \ : U \in \mathbb{U}\}$ for any $n \in \mathbb{Z}^+$ is another, and

$$\mathbb{U}^n = U \vee F^{-1}U \vee ..F^{-n}U,$$

then the limit

$$h_{\text{top}}(\mathbb{U}) = \lim_{n\to\infty} \frac{\log |\mathbb{U}^n|}{n}$$

exists, and topological entropy is defined as

$$h_{\text{top}} = \sup_{\mathbb{U}} \lim_{n\to\infty} \frac{\log |\mathbb{U}^n|}{n}. \tag{1.34}$$

Dinaburg and Bowen apply the notion to metric spaces. Dinaburg [1971] relates the link Kolmogorov makes between topological entropy and the ϵ-entropy mentioned earlier. It is based on extending the notion of a distance between two points to that of the maximal distance between those points during any n iterations of the map. Given some continuous F on the metric space $X_0 = (X, \rho)$, define $X_n = (X, \rho_n)$, where

$$\rho_n(x, y) = \sup_{i \in 0,1,..n-1} \rho(F^i x, F^i y)$$

As before, let

$$\mathbb{H}_\epsilon(X_n) = \log |\mathbb{N}_\epsilon(X_n)| \tag{1.35}$$

Then the limit $\lim_{n\to\infty} \frac{\mathbb{H}_\epsilon(X_n)}{n}$ exists, and

$$h_{\text{top}}(X) = \lim_{\epsilon\to 0} \lim_{n\to\infty} \frac{\mathbb{H}_\epsilon(X_n)}{n}.$$

Bowen [1971] extends Kolmogorov's notion of ϵ-separated to (n, ϵ)-separated set, which, intuitively, is just a subset of X (as defined above) whose every pair has at some (possibly different) point in n time steps separated by at least ϵ. Write maximal cardinality as $\mathbb{M}_{n,\epsilon}(X)$. Similarly extending the ϵ-spanning set gives $\mathbb{N}^R_{n,\epsilon}(X)$. It can then be proved that

$$h_{\text{top}}(X) = \lim_{\epsilon\to 0} \lim_{n\to\infty} \frac{\log \mathbb{M}_{n,\epsilon}(X)}{n}, \tag{1.36}$$

as well as

$$h_{\text{top}}(X) = \lim_{\epsilon\to 0} \lim_{n\to\infty} \frac{\log \mathbb{N}^R_{n,\epsilon}(X)}{n},$$

The n-cylinder construction led to the notion of a set of orbits which, after some time, can be distinguished with a certain resolution. Given a certain relation between a partition P of X and ϵ, we see that this set is exactly the (n, ϵ)-separated set. Hence the exponential rate of growth of trajectories distinguishable by cylinders, as defined in eq. (1.33), should be equivalent to Bowen's quantity ((1.36)), which is just the the topological entropy of X. Note that whilst the topological entropy as defined through counting n-cylinders involves taking a supremum over all partitions of the state space, the two definitions above contain a limit of small ϵ. Since ϵ is effectively inversely proportional to the cardinality of the partition, and since increasing that latter cannot decrease the number of resolvable orbits, then to all ϵ however small correspond partitions, and the given limit of small ϵ is equivalent to taking the supremum over all ϵ.

Link to Information Theory Consider a dynamical system with some topological entropy h_{top}, realised by a finite partition P of the metric space X. Vieweing the orbit as message received with some error in the signal the system becomes a stochastic process, and the orbit in the coarse-grained space is a sequence, with P defining the alphabet. The number $\log N(n)$ of all possible sequences of length n can be thought of as entropy of some distribution that assigns equal weight to all sequences. The topological entropy of (X, F) can be computed as

$$h_n = \log N(n) - \log N(n-1). \tag{1.37}$$

Practical estimation of topological entropy would involve constructing sequences with some initial condition. Implicit in the definition is the requirement that these are sampled uniformly from X. Thus h_T can be thought of as a specific quantity, with an existence of some more general notion that would depend on the initial distribution.

These are precisely the considerations behind the Kolmogorov-Sinai, or metric, entropy. In fact, in Adler et al. [1965] the functional entropic form is said to be "merely a delicate method of counting the number of sets in a partition in such a manner that the measures of the sets are given their appropriate weight in the tally". Adler et al. [1965] conjectured that given a dynamical system (X, F) with regular Borel measures μ, invariant w.r.t. the map,

$$h_{\text{top}}(X, F) = \sup_{\mu} h_{\mu,F}(X, F), \tag{1.38}$$

$$h_{\mu,F} = \sup_{P} h_{\mu}(X,F,P), \tag{1.39}$$

and

$$h_{\mu}(X,F,P) = \lim_{n\to\infty} \frac{H_{\mu}(\mathbb{U}^n)}{n}, \tag{1.40}$$

given the H_μ is the Shannon entropy of measure μ. Invariant measure of X is such that the measure of each measurable subset of X is equal to the measure of its preimage under F. The metric entropy of (X,F) is computable (Crutchfield and Packard [1983]) through

$$h_{\mu} = H_N - H_{N-1}, \tag{1.41}$$

given block entropy of lenth N defined on the *generating* partition of X, which could in turn be defined through this. Any other partition would correspondingly give be a lower bound.

Effective Measure Complexity and Excess Entropy

Crutchfield and Packard [1983] introduced the term excess entropy in the context of noisy symbolic dynamical systems, as a relative difference between deterministic entropy rate (metric entropy) and its finite-length noisy approximation. The marginal scaling at either infinite-length sequences or no noise both scale as power-laws with the noise and convergence exponent respectively; these can be estimated for various dynamical systems (and would be dependent on the partition of the state space).

The same terminology was used in Crutchfield and Young [1989](p.213) to define the measure of fluctuations in free information, $H(L) - hL$ (h is the dynamic entropy, which could be understood as the entropy rate), though no follow up on this definition, or examples of its uses, were given. The relationship between names and the appropriate quantites is also not made very clear.

When considering prediction measures on stationary system producing strings drawn from a finite alphabet (exactly the stochastic process described here), Grassberger [1986] asks about the additional information needed to predict a new symbol, given the previous N are known already:

$$h_N = H_{N+1} - H_N. \tag{1.42}$$

This can be shown to be the same as eq. (1.15), and is interpreted as the apparent randomness of strings of size N. Since addition of information about the past, i.e. lengthening of the block, can only decrease the uncertainty involved in predicting the new symbol, h_N should decrease with N. Hence the entropy rate can be defined as the limit

$$h = \lim_{N\to\infty} h_N, \tag{1.43}$$

where it exists. Grassberger [1986] also defines a related quantity, called Effective Measure Complexity (EMC):

$$EMC = \sum_{N=1}^{\infty} (h_N - h). \tag{1.44}$$

EMC is the normalised Riemann sum of h_N, the finite approximations to the entropy rate. It exists if h_N converge to h at an exponential rate, whereas an infinite EMC is suggestive of other, for example power-law, scaling (see Grassberger [1986]).

Consider the graphical representations of excess entropy as is used in the more recent publications of J. Crutchfield, D. Feldman and C. Shalizi, shown in fig.1.9. In their framework entropy rate is often referred to as entropy density. Using the monotonic growth property of H_N, $h \geq 0$. It is seen to be the asymptotic slope of block entropy as it grows with block size. The slope at any finite N is h_N, the randomness left when factoring out information present in strings of length N. The limit, entropy rate h, is interpreted as the irreducible randomness present in the information source; the inherent unpredictability per symbol of a string.

Figure 1.9 clearly demonstrates that the quantity E referred to by the authors as excess entropy is equivalent to Grassberger's EMC, as the process is that of equation (1.44). When Feldman and Crutchfield [2003] differentiate between types of excess entropy they single out the E_C, the excess entropy from (1.9), as a measure of convergence. It measures the randomness only apparent due to considering parts of the system - essentially not taking into account all the possible correlations. This is the randomness that can be 'explained away' as the entropy rate of finite-sized blcoks converges to its true value.

The reason why excess entropy appears in both subfigures is because for one-dimensional systems (i.e. systems with an unambiguous way of increasing block size by one) E_C is equal to E_S, the *subextensive* excess entropy, given by

$$H(L) = E_S + hL, \tag{1.45}$$

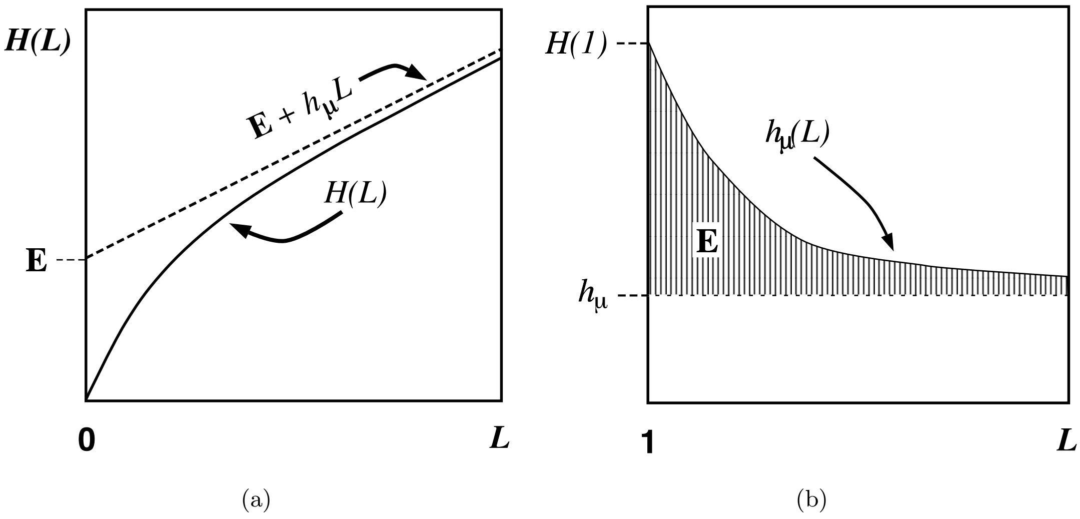


Figure 1.9: Block entropy, entropy rate and excess entropy E, from Feldman and Crutchfield [1998]. Here $H(L)$ is equivalent to H_L, and h_μ is h (the μ in the subscript refers to the measure we have as P).

in the limit of $L \to \infty$.

This interpretation sees excess entropy through the assumption of convergence of block entropy to linear behaviour with block size. The linear slope is given by the entropy rate; the height of the intercept has some information about the how much or how fast the monotonically increasing block entropy would have changed.

The third interpretation of excess entropy, E_I, was given in Li [1991], where the author notes that since (1.45) holds, then for two blocks of size M and N, excess entropy is just

$$C = \lim_{N,M \to \infty} [H_M + H_N - H_{M+N}]. \tag{1.46}$$

Here C stands for complexity, which is the framework in which in Li looks at symbolic sequences. Both blocks can be infinite, and since there is no overlap, the original sequence is actually assumed to be bi-infinite.

Hence excess entropy, or complexity, measures the information about one half of the sequence stored in the other. If the first half is termed the past, and the second the future, then excess entropy is the total information the past has about the future (and the other way around). Using the terminology developed above,

$$C = I(\overleftarrow{S_0}, \overrightarrow{S_0}). \tag{1.47}$$

In Feldman and Crutchfield [2003] C is seen to be different from E_I when considering 2D nearest and next-nearest Ising model. Whereas E_I is found by simply partitioning the plane into two halves, there is an ambiguitiy to the way block sizes increases (of which only one possible way was examined).

Here we talk about introducing causal measures by separating conceptual past from conceptual future: Alternatively, without being so restrictive, we can allow the past, or the future, or both, to spread over a set of times. In fact since T is one-dimensional and hence can be ordered, any time interval $I \subseteq T$ would by definition contain a selection of pasts and futures and hence their causal relation.

This separation leads on to the notions of predictability, which implies existence of link between excess entropy, or the EMC, and the ϵ-machines. A method to compute the EMC from this deterministic Markov model of the process is given by Grassberger et al. [1988]. In fact Grassberger [1986] proves

$$EMC \leq SC. \tag{1.48}$$

In Kolmogorov [1965] it is mutual information that is given precedence over entropy as a useful quantity. Entropy can be infinite and thus uninformative; whereas mutual information is more likely to be finite since it is bounded by the extent of connections between the systems in question. In this vein in the computational mechanics literature EMC is simply called complexity. There it is common to plot complexity - entropy diagrams for a variety of systems, though the results are interpretationally confusing - see 1.10(b). This is done using symbolic dynamics of the logistic map, however there is not much investigation into the effect of changing the (single) point of (binary) partition. Consequently the diagram for the excess entropy for the logistic map is rather uninformative (1.10(a)).

A variety of other Venn ‘information diagrams’ is now in existence, the same authors attaching a range of conceptual meanings to the various subsets. In the same manner statistical complexity is being calculated for a range of systems - but here, too, it seems to be more of potential categorisation tool.

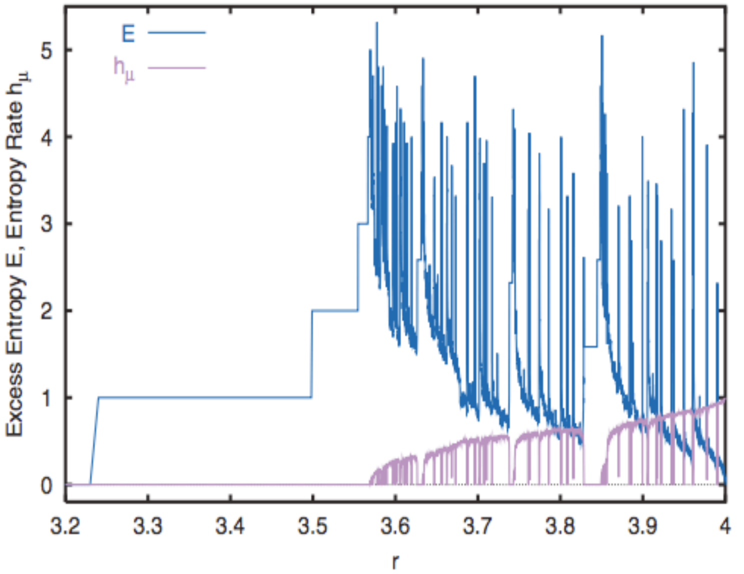


FIG. 1. (Color) Excess entropy **E** and entropy rate h_μ as a function of the parameter r. The top curve is excess entropy. The r values were sampled uniformly as r was varied from 3.4 to 4.0 in increments of 0.0001. The largest L used was L=30 for systems with low entropy. For each parameter value with positive entropy, 1×10^7 words of length L were sampled.

(a) Excess Entropy

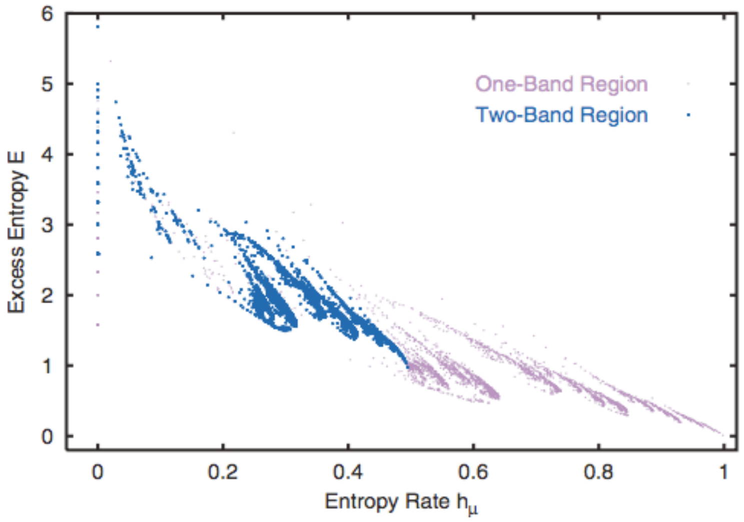


FIG. 2. (Color) Entropy rate and excess entropy (h_μ,**E**)-pairs for the logistic map. Points from regions of the map in which the bifurcation diagram has one or two (or more) bands are colored differently. There are 3214 parameter values sampled for the one-band region and 3440 values for the two-band region. The r values were sampled uniformly. The one-band region is $r \in (3.6786, 4.0)$; the two-band region is $r \in (3.5926, 3.6786)$. The largest L used was L=30 for systems with low entropy. For each parameter value with positive entropy, 1×10^7 words of length L were sampled.

(b) Excess Entropy v Entropy Rate

Figure 1.10: Measures of complexity using symbolic dynamics generated by the logistic map, from Feldman et al. [2008].

Chapter 2

Persistent Mutual Information

2.1 Settings and Definitions

Consider as before a discrete time stochastic process. Let $\mathbb{A}$ be the system alphabet, a countable finite set. Let S_i be the variable taking values in $\mathbb{A}$ and corresponding to the state at time i.

We anticipate potential correlation between the past and the future. For that reason, similar to the procedure behind several statistical mechanics measures, we consider two distinct subsets of $T = \mathbb{Z}$. Specifically, let $p, f \subset T$ be non-empty and non-overlapping, and $\min(f) > \max(p)$. The quantity of interest that we choose to look at is a probability measure μ supported on a σ-algebra of some Ω_p x Ω_f, where some Ω_b is the space of outcomes noted consecutively at elements of p or f.

With this in mind we first consider the case where the past $p = \{-T_1 + 1, .., 0\}$, and future $f = \{1, .., T_2\}$ for some $T_1, T_2 \in \mathbb{Z}^+$. Using the notation introduced in section 1.2.3, the 'past' variate is hence $S^0_{-T_1+1}$, while the 'future' becomes $S_1^{T_2}$. The 'joint' variate is easiest denoted by S_J where $J = p \cup f$.

In this framework the mutual information between the system's past and its future is

$$I(S^0_{-T_1+1}, S_1^{T_2}) = H[S^0_{-T_1+1}] + H[S_1^{T_2}] - H[S_J]. \tag{2.1}$$

If the alphabet $\mathbb{A}$ is a finite, countable set, then H is the block (discrete) Shannon entropy, the same as defined in the section earlier in the context of stochastic processes. If the outcomes are continuous variates H becomes an integral with respect to some measure. Taking the limit of semi-infinite block lengths,

$$I(\overleftarrow{S}_1, \overrightarrow{S}_1) = \lim_{T_1, T_2 \to \infty} I(S^0_{-T_1+1}, S_1^{T_2}) \tag{2.2}$$

which of course is the excess entropy as defined by eq. (1.47).

Here it should be mentioned that in line with the stochastic setting in all these cases we assume stationarity: the probability of seeing a sequence is invariant under time-shifts. This allows moving either the future forward or the past backward to be an arbitrary matter of choice.

There are several ways in which a measure over the 'past' can be defined. In maps with a clock we randomise the start time of the measurement. If the systems we look at are ergodic this will be equivalent to the initial measure having equal weights on all elements

of the attractor. In this sense it is conceptually welcome since it maximises our uncertainty in the sense described by Jaynes. Alternatively one could work with a flat measure over the initial state space, but that would depend on the choice of coordinates, and for example in the logistic map could reflect properties of the map such as the relative weight of the basins of attractions. This would result in a different measure over the attractor and hence necessarily lower the PMI. Another issue is the absence of the attractor altogether. In the standard map we consider the flat initial distribution over the state space. For the Double Pendulum, however, we sample from the microcanonical ensemble.
We now propose introducing a time gap between the future and the past. For that purpose we keep the reference point of 0 and define 'remote future' as $\{\tau+1,..,\tau+1+T_2\}$. Consider the mutual information between some past and future separated by a gap of size τ:

$$I(\tau, T_1, T_2) = I(S^0_{-T_1+1}, S^{\tau+1+T_2}_{\tau+1}). \tag{2.3}$$

We define Persistent Mutual Information (PMI) as

$$I(\tau) = \lim_{T_1, T_2 \to \infty} I(\tau, T_1, T_2), \tag{2.4}$$

where the limit exists (though if the limit is infinite we can still talk about 'PMI' in the context of how the argument is changing with parameters). At $\tau = 0$ PMI is equal to excess entropy. It is, however, a more general quantity, since the τ parameter imposes an effective minimum on the length scale of correlations we pick up on. This is particularly useful for discovering the global causal structures that exclude short term dependencies.
The properties of PMI can all be traced to properties of mutual information. The only differences between MI and PMI stem from the 'temporal' position of the marginal distributions, which as such are not detected by mathematics. Consequently, like MI, PMI can detect nonlinear inter-relation and in this sense is an improvement on covariance. Like MI it is zero only when absolutely no correlation can be found (given some resolution) - provided of course that we do not ask after inter-relations occurring in the gap between the past and the future. PMI is thus a parametric measure of nonlinear dependence.

Consider starting from a uniform *distribution* over the state space. Although in reality only a finite number of copies of the system is available, nevertheless we must admit the possibility of an infinite number, and assume that they can sample a continuous probability distribution. Starting with the latter also makes more sense when studying evolution of points in some state space where a uniform initial distribution is possible.

The H in eq. (2.1) is then the continuous entropy. Consider *defining* the joint prior as $\rho^p_J = \rho^p_p \rho^p_f$. Then the priors cancel from the mutual information expression, and eq. (2.1) can be rewritten in the usual Kullback-Leibler form of

$$I(\rho_p \rho_f | \rho_J) = \int_{x,y \in E} dxdy \ \rho_J(x,y) \log \left[\frac{\rho_J(x,y)}{\rho_p(x)\rho_f(y)} \right].$$

Here the marginals are $\rho_p(x) = \int_{y \in E} dy \rho_J(x,y)$ and $\rho_f(y) = \int_{x \in E} dx \rho_J(x,y)$. The fact that priors can be made cancel render PMI much less dependent on the specifics of the underlying spaces - the particulars of these can be made to not influence the result, which after all is only about the extent of correlation between the past and future. We thus view it as a necessary part of the PMI definition, since otherwise it is possible that PMI would not give zero for independent variables.

Incidentally, we see that Mutual Information corresponds to the entropy of the joint distribution where *the product of the marginals functions as a reference measure*, i.e.

$$I(\rho_p \rho_f | \rho_J) = -H \left[\rho_J | \ \rho_p \rho_f \right]. \qquad (2.5)$$

Or, in terms of marginal and joint distributions of discrete random variables,

$$I(\mu^p \mu^f | \mu^J) = \sum_{i,j} \mu^J_{(i,j)} \log \left[\frac{\mu^J_{(i,j)}}{\mu^p_i \mu^f_j} \right], \qquad (2.6)$$

where i, j are indices over the elements of the past, future, and joint partitions.

Graphical Interpretation Figure 2.1 shows the subject of PMI, which is the joint distribution. If the shapes of the joint support are taken to somehow represent the weight of the joint, then it is clear that the picture on the left is indicative of a much more random process than the figure on right, since the initial condition does not seem to constrain the future outcome in any way.

This graph points to other variables that can potentially differentiate between the pictures. One is the dimension of the joint distribution. It is clear that the most 'causality' is present when the joint is as little spread out as possible, which in the conservative system of this example means that its dimension has to be at least equal to the dimension of the

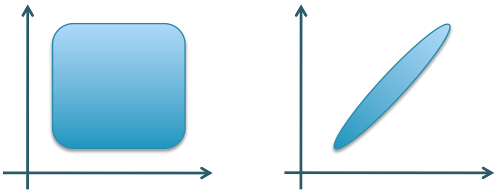


Figure 2.1: The two axes here are to be understood as the marginal spaces with some underlying metrics. The shapes represent the support space of the joint distribution in the product space. Although both the marginals in the two cases can be the same, the joint, if one assumes a proportionality to the support space, would be different, giving a much larger joint entropy for the first case.

marginals itself - ‘a point for a point’. On the other end of the scale is the completely random process which would produce the joint distribution that reaches to all parts of the product space, and as such has the dimension equal to the sum of individual dimensions of the marginals.

An even more complex picture is when the joint distribution is fractal or even multifractal. Unless imperfect resolution is assumed this case is not pictured in the figure above. On the face of it the object here drawn in blue would change shape depending on the resolution of the image. We deal with this case in the next section where for the period-doubling accumulation point of the logistic map the estimated PMI is seen to increase indefinitely, or at least as far as practical resolution can take us. We will see that as long as the visibility is limited by resolution PMI can appear to increase even if the joint is not fractal, in the limit of infinite resolution.

2.2 Persistent Mutual Information in Dynamical Systems

PMI is fundamentally a probabilistic notion. The need for something fulfilling its role arises naturally in the context of stochastic systems, where descriptions of states at different times are done on the level of measures over the state space. The future is not fully determined by the present, and so uncertainty enters the system through the evolution rule. It is this factor that invites exact statements and leads to quantities such as PMI, entropy rate, excess entropy, and others.

Deterministic systems, on the other hand, do not allow for any doubt in evolution. In order for PMI to make sense in this setting we need to let some aspect of the system admit uncertainty. On the more philosophical ground this introduction of probability can be interpreted as working with incomplete knowledge about the given aspect.

The usual definition of a dynamical system as a state space combined with an evolution rule gives at least two levels where this uncertainty may enter (Crutchfield and Packard [1983]), plus a combination of the two. The first leads to 'noisy' systems defined by evolution rule supplemented with an error, studied in for example White et al. [1981]. This could also, of course, be interpreted as incomplete resolution of the state space. We take up a similar idea, but consider this partial knowledge as being a feature of the observer, and not the map. In short, we ask the question of what information a limited resolution of the initial state can provide about the final outcome, *defined with the same level of uncertainty.* In this second level uncertainty enters the dynamical system at the level of knowledge of the initial condition. This notion is supported in Farmer et al. [1980]: "prediction must be discussed in terms of ensembles of initial conditions rather in terms of the behaviour of individual points".

We now define these concepts more rigorously. Let (X, F) be a (discrete) dynamical system. Let P be a partition on X into M cells C such that $P = \{C_i : i = 1..M\}$ and

$$X = \bigsqcup_{i=1}^{M} C_i. \tag{2.7}$$

With some suitable σ_P define a measure space (P, σ_P, μ).

We also admit a prior measure μ_p. This allows the definition of measure of the evolved system as follows.

Let

$$\mu(C_i) := \mu_p(C_i). \tag{2.8}$$

The evolved measure μ_p^τ is defined for all $A \subseteq X$ through

$$\mu_p^\tau(A) := \mu_p\left(F^{-\tau}A\right). \tag{2.9}$$

Then

$$\mu^\tau(C_i) := \mu_p^\tau(C_i) \tag{2.10}$$

We define the joint measure μ_J through the conditional, such that for any $(A, B) \subseteq X$xX,

$$\mu_J\left(A, B\right) = \mu(A)\mu_p^\tau\left(F^\tau(A) \cap B\right). \tag{2.11}$$

In the cases we study the state space $X \subset \mathbb{R}^d$, and prior measures will be the Lebesgue measures. This means that the joint prior is indeed the product of the marginal priors, and the two will cancel. PMI will then be a function of the 'past' marginal only.

We begin with N i.i.d. points $X_i^0 \sim \rho_0$ (where ρ_0 is the density associated with 'past' measure defined *at* a certain resolution), and evolve each with F^τ to obtain X_i^τ (below we talk about the methods used to estimate PMI using the set of $X_i = \left(X_i^0, X_i^\tau\right)$, $i = 1..N$ as data).

In this methodology we essentially reduced the semi-infinite block defined by the stochastic process approach as the 'past history' onto a single variable. It is possible because PMI being the function of entropies, it does not manipulate values from support spaces, rather the measures of the subsets defined on the latter. Here the fully-deterministic system ensures that a point $X_i^0 \in X$ is associated with a unique orbit which, if a symbolic block variable is required, can be rewritten in terms of indices of cells housing its consecutive elements, in a manner similar to the process of finding the metric entropy described in the Introduction. Metric entropy and other functions of blocks of variables are based on the fact that there is *not* a unique correspondence between the initial block and consecutive blocks. Here, on the other hand, we rely on the block corresponding to the initial *point*. This ensures that the measures we sample by considering only the initial and final points are the same as we would have sampled had we considered sequences of points or their symbolic representation.

Interpretation Fig. 2.1 earlier provides a visual explanation of what it means to discuss PMI in the ostensibly deterministic context of dynamical system. Without loss of generality we can view the graphs by imagining the pictured axes as discretised state spaces of some deterministic map, the very setting of symbolic dynamics (and the 'effective' symbolic dynamics of our method). Consider a system whose marginal spaces are the same, and with the same partitioning, such that the flat past measure induces a flat future measure - as will be the case with the standard map. Simply looking at the marginal measures gives no indication of the extent to which the map loses initial information. This is exactly what the joint captures (or rather a function related to the lower limit of the rate of information loss. Orbits can be different in the intervening times but close together at some time τ, whereas if orbits differ at τ any differences in the intervening times will not 'lessen' the difference noted by PMI).

Consider a subset of the past marginal with some measure. We populate the subset with points whose relative number is defined by the measure of that value, and is either equidistributed, or, if a further partitioning exists, subdivided again. The points then get evolved by the map for τ times, and their final positions go towards contributing to the 'future' marginal measures. If the motion is somehow predictable or 'causal', then the points that were close together will tend to stay close together. Their relative distances will not decrease. That is indeed the case in the right hand side of the figure, under the assumption that the thickness of the line is somehow indicative of the size of cells.

If, on other hand, trajectories diverge in a chaotic manner, so that any knowledge of the initial condition is lost, it is likely that there will be a much greater variation in the distribution of the points initially in one subset. That is what the first subfigure shows. So even when evolution is deterministic it is still possible to ask the question of how drastic a small inexactitude in the initial condition will, on average, turn out to be.

In the framework usually employed by Tsallis *et al* (see the next few citations for example), PMI can be viewed simply as an aggregate related to entropy production. In works such as Baldovin et al. [2003], Añaños et al. [2005], incidentally also focusing on the standard map, a number of points start equidistributed in a cell and their evolution is traced. Their positions at some τ is then added to make up the overall distribution at that time, obtained by also averaging over the location of the initial cell, which is made to be arbitrary in the state space. Evolution of the individual cell then corresponds to the horizontal movement in the figure given. The difference between approaches is also clear - whereas the authors marginalise by averaging in the *horizontal* direction to get the future distribution, here we look at the evolution without losing track of the relative location of the joint points.

2.3 Permanently Persistent Mutual Information

We would now like to draw a distinction between the variables related to the transient process of settling down and *existing* in some stationary state. Works like the ones just mentioned by Tsallis study the evolution in the entropy of the marginal distributions. The information in these measures, being related to the differences in mean values of observations, thus concerns the change brought about by the actual settling process. On the other hand the causal inter-relations between the past and the future are only preserved in the joint, however small its dimension (with PMI we of course disregard the intervening time bulk). PMI is a function of the joint distribution, and so includes hidden information in the settled dynamics.

We choose to view the difference using the terminology of weak v strong emergence. The original, emergentist definition is due to Broad [1925]:

> We must wait till we meet with an actual instance of an object of the higher order before we can discover such a law; and [...] we cannot possibly deduce it beforehand from any combination of laws which we have discovered by observing aggregates of a lower order.

The implication behind this is that looking at a collection of histories in some ways smoothes the particular features of each one, and that conversely by *not* doing so we gain something of the forecastability of the individual realisation.

Linguistically one could also make the distinction between predictability and forecastability, and use the most natural example these words conjure up. Predictability usually refers to the extent to which global behaviour is understood after seeing the system multiple times, e.g. this is a concept related to climate. Asking after the future of a specific ‘instance’ is the action of forecasting, which is naturally associated with weather. The implication in the latter being that to make the forecast in an optimal manner one exploits the data from the *recent* past of this *specific* instance.

In modern phraseology this is summarised by Chalmers [2006]:

> We can say that a high-level phenomenon is *strongly emergent* with respect to a low-level domain when the high-level phenomenon arises from the low-level domain, but truths concerning that phenomenon are not *deducible* even in principle from truths in the low-level domain.

In the context of chaos it is common to consider time separation as an analogue to the conceptual level separation. Thus lower-level and higher-level domains become trajectory positions in the past and the future respectively. The flat initial measure over some partition of the support space *of the attractor* represents the truth known about the initial condition: it can be in any of the sets in the partition with equal probability. Here we consider the truth about an initial condition to be equivalent to the statement of our knowledge of the initial condition. This knowledge gives us some information about the system, which we can use to deduce the final position. This latter truth will also take the form of a measure. The information common to these two measures can then be viewed as the information from the past that remains relevant and constraining about the far future. We call this quantity Permanently Persistent Mutual Information. We also propose that in its guise as the ultimate lower limit of forecastability that is possible given a specific realisation of a system, PPMI thus measures the extent of strong emergence[1].

Specifically, Permanently Persistent Mutual Information is

$$I(\infty) = \lim_{\tau \to \infty} I(\tau), \tag{2.12}$$

where $I(\tau)$ is given by eq. (2.1). Note that the limits are taken to correspond to first examining the bi-infinite sequences, and only *then* separating them. It is an interesting and perhaps to some extent philosophical question of whether changing the order of limits has an effect on $I(\infty)$. In practice it is perhaps easier to take the τ limit first, since it is the length of data one misses out on.

[1] Here we relax the notion of 'not deducible even in principle', since such a definition precludes any possibility of quantification.

2.4 Estimating PMI

We wish to estimate mutual information between two densities, the 'past' and the 'future'. Our dataset consists of $i = 1..N$ pairs of d-dimensional points $X_i = \left(X_i^0, X_i^\tau\right)$, $d \in \mathbb{Z}^+$. X_i^0 are understood to be realisations of X^0, distributed according to the 'past' ρ_p; X_i^τ of $X^\tau \sim \rho^\tau = \rho_f$, and X_i of X. The distribution of X is the joint distribution ρ_J. Because the initial samples are i.i.d, X_i are i.i.d. as well.

The mutual information is a function of the ρ densities. As we do not have direct access to these, we need to use estimators which take as input variables sampled from the underlying distributions. The different ways of expressing mutual information means estimations can be performed on a variety of levels, and outcomes manipulated algebraically to get the answer.

Approximating the Measures At the more straightforward end of the spectrum is approximating the distributions by flat-intervalled versions based on some partition of S and $S^{'}$ (the support space of X^τ), defined indirectly through requiring uniform sampling of the attractor with some given number of points N. The measure of each partition element is associated to the relative number of points that fall within that cell, so for example the measure of the $(i,j)^{th}$ cell of the partition of Sx$S^{'}$ is the number of X_i whose first element is the i^{th} cell of the partition of S *and* whose second element is the j^{th} cell of the $S^{'}$ partition. With this the formula for mutual information becomes the usual discrete Shannon entropy version given in eq. (2.6), where μ become defined on the partitions by frequency counting. As long as the underlying distributions are smooth enough the estimate converges in the limit of first, large N, and second, small cell size. There is some scope of variation in this method, rooted in the motivation behind the partitioning. The two more common approaches are division of S into cells of the same linear size, and cells of the same measure.

Estimating Entropy The next level up involves estimating individual entropies and finding the deficit of the outcomes to arrive at mutual information. However, as we shall see in the future sections, one of the cases we will be looking at will involve constant - at least analytically - marginal entropies, so PMI will only be dependent on the joint entropy. We are therefore also interested in estimating entropy in its own right.

Kozachenko and Leonenko [1987] introduced an estimate of continuous entropy from a set of vectors in a metric space of arbitrary dimension d based on statistics of *nearest* neighbour distances $\epsilon_i, i = 1..N$. For the Euclidean metric the authors prove that

$$H = \lim_{N \to \infty} \frac{d}{N} \sum_{i=1}^{N} \ln \epsilon_i + \ln c_1(d) + \ln \gamma + \ln(N-1) \tag{2.13}$$

is an unbiased estimate of the continuous entropy, and that the mean of H_N converges to the true entropy as well. Here $\ln \gamma$ is Euler's constant and $r^d c_1(d)$ is the Euclidean volume of the unit sphere in d dimensions. The bulk of the paper contains the proof.

The basis of the estimate is in the switch between the framework with a probability distribution over the position of points to one with a distribution describing interpoint distances. Here we follow the methodology as expostulated by Kraskov et al. [2004]. We will call the latter estimate the K-G estimate to emphasise that here the depth of probability resolution can be changed by considering the distance to k^{th} nearest neighbour.

Consider again the formula for the continuous entropy of a distribution ρ of a variate X taking values x,

$$H(X) = -\int \rho(x) \log \rho(x) dx. \tag{2.14}$$

As this can be interpreted as the mean of $\log(\rho)$, $\hat{H}(X)$ is then an unbiased estimator of H:

$$\hat{H}(X) = -\frac{1}{N} \sum_{i=1}^{N} \widehat{\log \rho(x_i)}, \tag{2.15}$$

where $\widehat{\log \rho(x_i)}$ is some unbiased estimator of $\log \rho(x)$.

Let $p_i(\epsilon)$ be the 'weight' of some i^{th} ball of radius ϵ, $p_i(\epsilon) = \int_{x \in \text{ball}} \rho(x) dx$. Then by the mean value theorem there exists x_i s.t.

$$p_i(\epsilon) = \rho(x_i) V_d, \tag{2.16}$$

where the volume of the ball $V_d = \epsilon^d c(d)$, and $c(d)$ is the (not necessarily Euclidean) volume of the unit ball in the support space of ρ.

Therefore the expected value of $\log \rho(x)$ is, from (2.16), just

$$\widehat{\log \rho(x)} = \widehat{\log p(\epsilon)} + \log c(d) + d\widehat{\log \epsilon}. \tag{2.17}$$

The next key point is that p_i the distribution of 'weight' of the i^{th} ball, can be used to express the probability $P_k(\epsilon)d\epsilon$ of a point having its k^{th} nearest neighbour in the thin shell of linear size $d\epsilon$ centered at the radius $\epsilon/2$ away from it. Interest in this quantity was motivated as far back as 1940s Chandrasekhar [1943]. Bhattacharyya and Chakrabarti [2008] gives a clear explanation of the methods by which $P_k(\epsilon)d\epsilon$ can be derived. A simple consideration that yields it is combinatorial in nature: we look for the number of ways in which $k-1$ out of $N-1$ points can be arranged strictly *inside* the ball and the way in which the remaining $N-1-1-(k+1) = N-1-k$ points can be arranged outside it, to end up with the trinomial formula. Thus $P_k(\epsilon)$ is a function of N, k, and $p_i(\epsilon)$, and

$$\mathbb{E}\log p_i = \int_0^\infty d\epsilon P_k(\epsilon)\log p_i(\epsilon), \tag{2.18}$$

yielding

$$\mathbb{E}\log p_i = \psi(k) - \psi(N), \tag{2.19}$$

where $\psi = \Gamma^{-1}(x)d\Gamma(x)/dx$ is the digamma function. Since the estimate of entropy, equal to negative the expected value of $\log\rho(x)$, is just

$$\hat{H}(X) = -\psi(k) + \psi(N) - \log c(d) + \frac{d}{N}\sum_{i=1}^{N}\log\epsilon_i. \tag{2.20}$$

Here we are looking at the probability with cells of resolution defined by the k^{th} nearest neighbour, whose distance from point i is $\epsilon_i/2$. According to Kraskov et al. [2004], the errors are maximum of order k/N, but naturally vary as the distributions deviate from uniform and eq. (2.16) becomes a less accurate statement. The K-G estimator can be seen to coincide with eq. (2.13) for $k = 1$, since $\Psi(1) = -\gamma$, and for large N, in whose limit these converge, $\Psi(N)$ aligns with $\ln(N)$.

Estimating Mutual Information Errors stemming from the marginal and joint entropies may be of different order, and so may not necessarily cancel, leading to systematic deviation in mutual information. Problems like this are common when, for example, the same *linear* cell size is used for both the product and marginal spaces. Over/under-sampling could then be different for these probability distributions, leading to an imbalance that may have an effect on the estimate of mutual information. This is exactly what happens when eq. (2.20) is used to estimate both the marginal and joint entropies *using the same* k, which

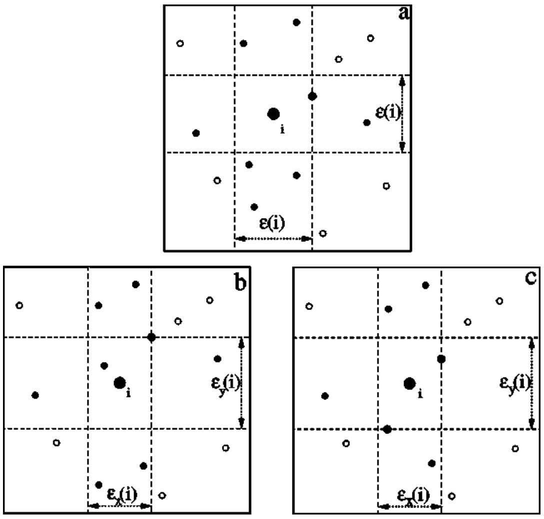


FIG. 1. Panel (a): Determination of $\epsilon(i)$, $n_x(i)$, and $n_y(i)$ in the first algorithm, for $k=1$ and some fixed i. In this example, $n_x(i)=5$ and $n_y(i)=3$. Panels (b),(c): Determination of $\epsilon_x(i)$, $\epsilon_y(i)$, $n_x(i)$, and $n_y(i)$ in the second algorithm for $k=2$. Panel (b) shows a case in which $\epsilon_x(i)$ and $\epsilon_y(i)$ are determined by the same point, while panel (c) shows a case in which they are determined by different points.

Figure 2.2: From Kraskov et al. [2004] illustrating the relation between ϵ and number of nearest neighbours in the marginal space, where the authors have used x for p, and y for f.

would be the natural place to start in order to lessen the computational load. (Although we have to add that here the reverse happens. k by definition controls the estimate of probability according to (2.4), so it would not be that a different weight is sampled, but rather that the uniformity assumption might have to be lessened for the joint more than the marginals. Nevertheless the effect of disproportionate errors is the same.) Having said that, there is no problem adapting eq. (2.20) to work in the joint space, the only difference being that the volume of the unit ball is now the product of the respective volumes in marginal spaces; and that the multiplicative factor from the average interpoint distances is the sum of the dimensions of the marginals, $d_J = d_p + d_f$.

Kraskov et al. [2004] introduces a mutual information estimator that finds the entropic deficit by combining the K-L estimates where the marginals and the joint entropies have *variable* resolution. As a result it is less prone to non-uniformity based errors. The aim is to cancel, through the addition and the subtraction, the $\frac{d}{N}\sum_{i=1}^{N}\log\epsilon_i$ term. It is possible since, as we have just noted, the dimensions are additive. All that is required is that the linear cell size ϵ is kept the same for the marginals *and* the joint.

Recall that the ϵ_i is defined *through* the number of neighbours the i^{th} point contains within a ball of that diameter. Therefore, if it is kept constant, the k in the K-G estimator of the marginals can be rewritten as $n_p(i) + 1$ (for the number of neighbours situated within that ball in the 'past' marginal) and $n_f(i) + 1$ for the second marginal. We thus have an estimator effectively parametrised by the resolution of the *joint* distribution. Fixing a k we then compute ϵ_i by looking at the less dense distribution of points in the joint space, and from this distance find the marginal numbers n_p and n_f of points falling into those cells. Thus the k in the marginal formulae vary with individual points and the first terms have to be replaced by $-1/N \sum_{i=1}^{N} \psi(n_{p/f} + 1)$. Thus, from eq. (1.2.2) in the previous chapter, the estimator for mutual information is given by:

$$\hat{I}(X, Y) = \psi(k) + \psi(N) - 1/N \sum_{i=1}^{N} \left(\psi(n_p + 1) + \psi(n_f + 1)\right). \qquad (2.21)$$

Computational Method Implicit in the K-G-based estimators of entropy and mutual information is the requirement to find the k^{th} nearest neighbour of a vector of arbitrary dimension. The mutual information estimator requires, in addition to that, to do the reverse and find other vectors *given* a certain distance from the first. These searches form the basis of the computational load. Both problems can be solved by a 'dumb' search, but that begins to be unfeasible for any reasonable parameter range. Interestingly enough, both problems are actually also tractable in a simple and related manner as functions of a kdTree[2]. We now briefly outline the possible methods.

Any 'dumb' algorithm of the kind will be of order N^2. Any possible improvement will involve a balance in the difficulty in implementing a new algorithm and the range of parameters which we wish to use. Methodologies which begin be advantageous towards the higher end of the reasonable parameter range often do so at the expense of structures which require some minimum setup time. This is exactly what happens with the kdTree. For small values of (d, N, k) it thus makes more sense to use the most primitive N^2 search (hereafter called the 'dumb' method) which will - though arguably for unusably small parameters - be faster than the more advanced methods.

We considered three possible methods, each one offering some advantage depending on the perspective. At the simplest and the most easily (double loop) implemented end of the scale is the 'dumb' method. Optimised for large N is the kdTree setup. For the range in

[2] A kdTree is a nested way of storing data that yields logarithmic, rather than exponential, search times. See Üngör [2013](url in references) for a tutorial - alternatively, Press et al. [2002] for implementation and by necessity an introduction to the subject.

between we introduce the recurrent method.

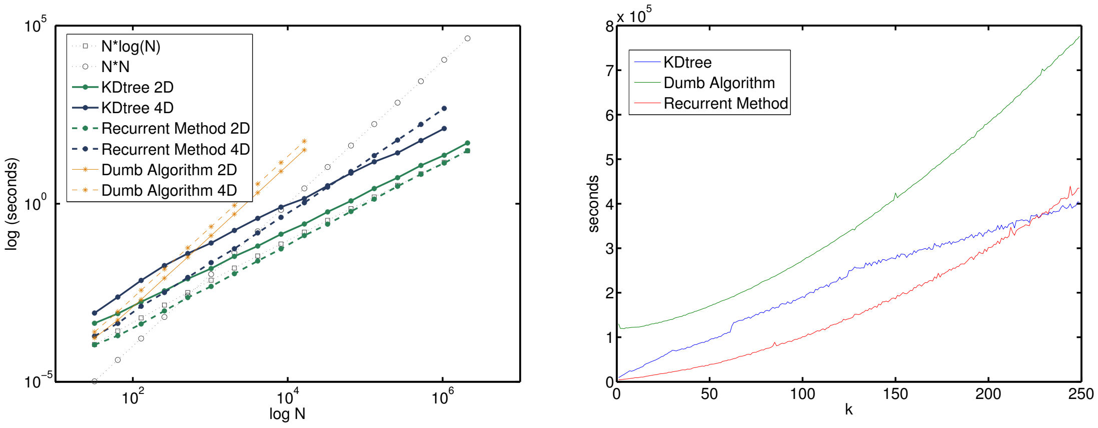
Recurrent Method The recurrent method is optimised for a nearest neighbour search of a given index. The algorithm first looks at the number of points, and if it is less than a certain threshold a, performs the 'dumb' search to find the knn (k^{th} nearest neighbour) distances. For a large number of points the dataset gets split into two *based on the median in the direction with the largest range.* The process then continues recursively until the subset has fewer than b points left (here we take $b = a$). It is at this level that the knn search is actually performed, here using the simple 'dumb' method.

The key feature of this algorithm is the error checking that becomes necessary when one considers that for a point near the boundary of the split only the points *to one side* of the boundary are checked for their distances, and as such some overestimation is bound to happen. Points contained in the other set may actually lie closer to a point near the boundary. The saving feature is that that should only be the case for points from the other set that are themselves close to the boundary. Not all the points from the other set need to be checked. The natural way to simplify this is to sort the two subsets first and then error check simply by going along the indices.

Thus the algorithm recursively splits the set into subsets small enough to apply the standard search procedure, and once that is done, begins the reverse process of combining the subsets together by pairs, building up the original set. Each recombination involves an error check for knn distances for points on the boundaries, for which the pair needs to be sorted along some direction. The sorting needs to be repeated with every step backwards, since the split may have been done in different directions for each of the subsets. The metacode for $d = 2$ is presented below. The algorithm is easily adaptable to periodic settings and higher dimensions. Comparison for running times between the three methods is shown below.

```
template <int DIM> void NND (myPoint <DIM> *list, int size_in, int *part1, int k_in, double Period, double* distances_address, int TRIVIAL_IN) {
  if (size_in < TRIVIAL_IN) {
    dumb_NN(list, size_in, k_in, Period, distances_address); //The Dumb Algorithm

    //part1 controls by which variable the list was sorted - here set them as OPPOSITE since no sorting is actually done in the dumbNN method; this ensures lists are definitely sorted later
    *part1 = (*part1 + 1) % DIM;

  } else {
    int direction = 0; //default for 1st dimension
    int list_length = 0;

    splitlist(list, size_in, &list_length, &direction); //USE THE SAME SPLITTING PROCEDURE FOR PERIODIC DATA

    int l1 = direction;
    int l2 = direction;

    NND( list, list_length, &l1, k_in, Period, (list -> get_p_kd()) , TRIVIAL_IN);
    NND( list + list_length, size_in - list_length, &l2, k_in, Period, ( (list + list_length) -> get_p_kd()) , TRIVIAL_IN);

    //sort the two parts of the list in the same direction - if they are not sorted in that direction already
    comp_dir = direction;

    if (l1 != direction) sort( list, list + list_length, cmp <DIM>);
    if (l2 != direction) sort( list + list_length, list + size_in, cmp <DIM>);

    *part1 = direction; //variable that keeps track along which dimension the two parts are now sorted

    errorcheck( list, list + list_length, list_length, size_in - list_length, direction, k_in, Period);
    errorcheck( list + list_length, list, size_in - list_length, list_length, direction, k_in, Period);
  }
}
```

The more common way to do operations on relative distances between a set of points is to set up a kdTree. Conceptually this is an object containing a vector of original points (not an array, but something which allows points to be added and removed) supplanted on a partitioning of the space that these points induce, along with information about the structure of the resulting boxes and their locations within each other. Computationally, along with a kdTree object and some object holding a d-dimensional point (and in our



(a) $k = 3$, $a = 10$

(b) $N = 1000$, $a = 10$, $D = 2$

Figure 2.3: Running time of the search of k^{th} nearest neighbour in a sample of N points equidistributed in a box of dimension D. The lines of proportionality, N^2 and $N \log N$, were each scaled down to fit in the graph.

code we store metric as a functor), it requires a box and a boxnode. A box is a rectangle defined by two corners, so called because the partitioning is along the directions defined by the dimensions, and a boxnode is there to store indices of boxes and allow for the retrieval of parent and daughter boxes. The smallest box contains two points. The knn search is a simple traversal upstream that starts by opening the box containing the upper bound on k. This gives a candidate for $\epsilon(k)$. Further boxes are 'opened' and its points checked for proximity to the point in question only if the distance to another box is smaller than the candidate. Similarly, finding the number of points within a given distance would mean opening all boxes that are within the value given. This is where the functor comes in useful. This is a general procedure applicable in any dimension. It does not require knowing the boundaries of the space, though in all the cases we look at this is in fact known. It is easily adaptable to systems with periodic boundary conditions.

Press et al. [2002] contains the basic algorithm for $C++$, which is the platform we used. Figure 2.3 shows the comparison for running time for all the three knn search algorithms for N points on a non-periodic box of dimension d. Since we never seriously consider the point-by-point search the only reason for comparison is to show that there are parameter regimes where the recurrent method fares better.

From figure 2.3(a) the first thing to notice is that the obvious method does indeed run as N^2, and that changing the dimension merely gives it a different proportionality constant. The same multiplication is evident when the kdTree method is run for higher dimensional data, keeping both roughly proportional to $N * \log N$. That is not the case for the recurrent method. There changing the dimension alters the manner in which the curves increase, so that increasing the sample size in a higher dimension would give a disproportionately larger running time than the same increase in a low-dimension space. This causes there to be a crossover between the plots for kdTree and the recurrent method in $4D$: in that space after increasing N beyond a certain point it makes sense to switch to the kdTree method. The point this comes in at is, still, towards the larger end of the average scale. In terms of sample size the recurrent method is seen to behave well, being even more optimal than the kdTree in $2D$, though showing some signs of a potential crossover with the kdTree 2D method.

The main problem, of course, is that the error checking procedure should be relatively short when the distribution is uniform as is the case above. The moment there is a change from uniformity the recurrent method may not be faster than the kdTree. In practice, however, the recurrent method on the joint distribution of the standard map performed in reasonable times.

Figure 2.3(b) shows precisely the computational price for having to set up the kdTree structure in order to compute distances indexed by k. For k less than approximately 200, at least in $2D$, it is actually faster to use the recurrent method (for this very small N). Only at smaller resolutions does the kdTree structure begin to pay off.

In practice we began our work by computing PMI for the logistic map using the simple binning strategy, for both equidistance and equidistributed bins (see next section). The main hurdle turned out to be not the computational length but rather a high sensitivity to under and over-sampling, which we were able to see for parameters where an analytical answer was known. Yet when the underlying distributions became fractal the outcomes using these methods were telling in so far as the fractal dimension was concerned. For the standard map with a four-dimensional joint we used the K-G entropy estimator and the kdTree method. Once it was setup, it became straightforward to continue with the method even for low-dimensional systems and abandon binning altogether.

Chapter 3

Persistent Mutual Information and Permanently Persistent Mutual Information in the Logistic Map

Persistent Mutual Information is a measure of correlations that persist above a time gap. Varying this can shift the focus from short-term causality to finding the trends less affected by the repeated application of this (discrete) map. In dissipative systems that possess an attractor, like the logistic map, the short-term behaviour of any trajectory will in some sense invariably involve movement towards that attractor. Our investigations into long-term correlations correspond therefore to analysing the dynamics on the attractor.
The individual systems we will be looking at are therefore attractors of the logistic and the tent maps. The setup in which we analyse it presupposes an ensemble of these, with a uniform initial ensemble distribution ρ_0. Under F^τ iterates it evolves to ρ_τ, and, keeping track of the conditional probabilities, we compute the mutual information between the two to obtain the Persistent Mutual Information $I(\tau)$.
This is done through feeding the dataset (sampling the joint pdf) generated by this setup into a mutual information estimator. It forms the beginning of our discussion. This ad hoc approach forgoes for the moment discussions of whether, for example, the attractor admits a probability distribution - and the MI estimator will simply assume a distribution behind the input set of points (since in a lot of cases the distribution will simply not exist, this discussion will also prove a test of how well the estimator deals with these situations). In this we anticipate potentially starting with a time series from an unknown source, and asking what features PMI picks up. We do at some point cheat by using our knowledge of the attractor at a given logistic Map parameter r, but only for the purposes of getting the most out of the computational setup.

The logistic map is a good toy model since for some regimes we can analytically compute the PMI. This allows us to assess behaviour of both different estimators and various parameters. We therefore structure the discussion as follows: first we find the optimal methods for computing the PMI by comparing our results to the analytic predictions. We then investigate variation of PMI with τ. We note that some regimes admit PMI that increases indefinitely with resolution, and derive an expression for this variation. Lastly we compute Permanently Persistent Mutual Information (PPMI) as a candidate for the measure of strong emergence in both the logistic and tent maps.

3.1 Persistent Mutual Information in the Logistic Map

Let $X \subseteq \mathbb{R}$ be the state space of the logistic map defined by

$$x_{n+1}= f(x_n) = rx_n(1 - x_n), \tag{3.1}$$

where $x \in X$. We confine our analysis to $0 \leq r \leq 4$, for which $X = [0, 1]$.

The main behaviour of the map was discussed in the Introduction. Briefly, for some $r < r_c$ the attractor A_r consists of a finite number of points $p = |A(r)|$ that doubles successively as r increases to r_c, $r_c \approx 3.57$. For $r_c < r \leq 4$ there is a reverse process of halving the global periodicity down to 1. After the period-doubling accumulation point r_c the motion becomes chaotic, with only occasional periodic windows. But global periodicity is still present in chaotic motion, where it enters through the chaotic bands, so that when A_r consists of p nonzero-sized bands, the trajectory visits each band every p^{th} step: and the description of chaotic refers to the effective motion within the bands themselves.

Analytical Limit Consider the logistic map for some value of r that results in regular motion with period p. At any point in time the trajectory can then be found in one of p points. Let μ^p be the measure on A_r giving every element a weight of $1/p$. As before, let μ^f be the evolved measure at some time τ on A_r, and μ^J the corresponding joint measure. PMI is then given by

$$I(\tau) = \sum_{i,j} \mu^J_{i,j} \log \left[\frac{\mu^J_{i,j}}{\mu^p_i \mu^f_j} \right], \tag{3.2}$$

where the indices $i, j = 1..p$.

The evolved measure clearly retains the same weights and the joint measure only has p nonzero elements, so that each must have the of weight $1/p$. There are hence p nonzero terms in the double sum, each of which is equal to $(1/p) \log p$, and so for period-p motion the PMI in the logistic map is

$$I(\tau) = \log p. \tag{3.3}$$

The same solution applies for regimes when the attractor contains non-zero intervals with the same periodicity, i.e. when $r > r_c$ and motion is *not necessarily* regular.

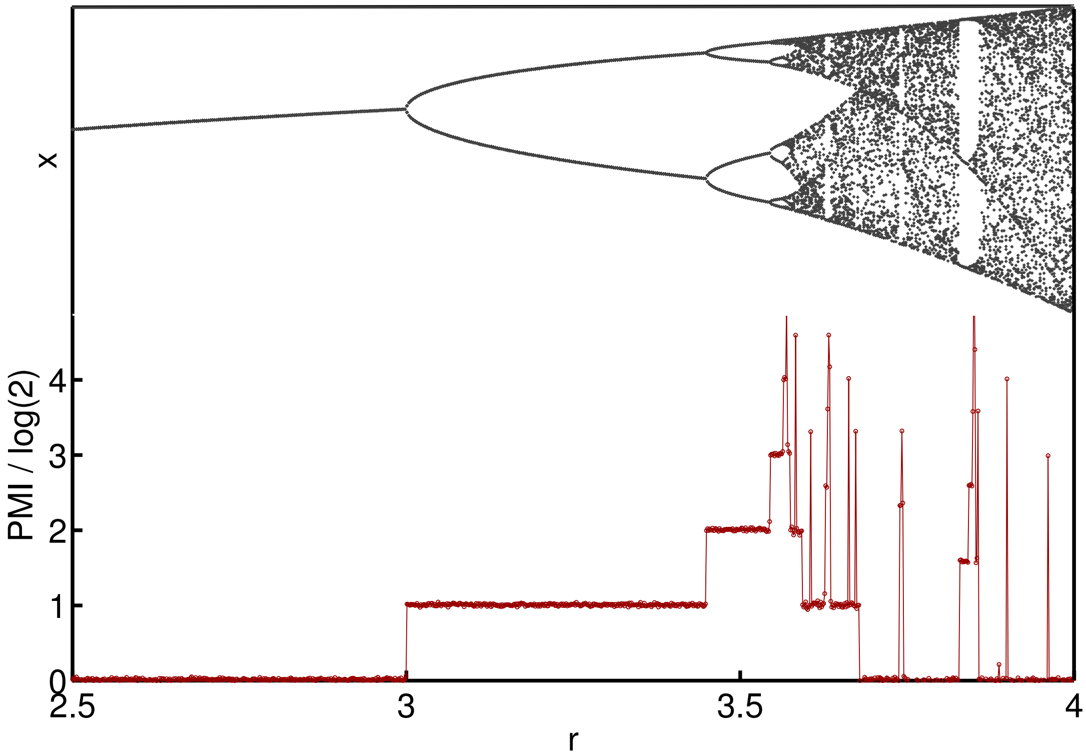


Figure 3.1: Kraskov-Grassberger estimate of PPMI for a typical and computationally optimal set of parameter values (analogous to the ranges in the subsequent graphs), for the logistic map $F(r)$ with added noise of order 10^{-12}.

PMI in the Logistic Map Figure 3.1 shows the Persistent Mutual Information for the logistic map for some typically large τ, and the corresponding bifurcation diagram, as a function of the parameter r. Just as predicted PMI increases at period-doublings in steps of size $\log 2$ until the accumulation point, after which the global trend is to decrease. As the bands merge PMI becomes a background of zero *in no way different to the fully-developed chaos*. Periodic regimes show themselves as structures visible on top of the band layers.

There are two parameters that are responsible for what is seen in figure 3.1. First, this graph is computed with a finite resolution - eq. (3.5) puts a limiting value on the observed resolution in terms of the number of points N. There thus exists a set of r values for which the PMI at perfect resolution could be larger than what is visible (and may be infinite).

The second is the set of r for which PMI is computed in the first place. The set of r in which to detect higher periodicities is small. Around r_c, for the given set of values we see a period of order 10, and then chaotic motion with a similar trend, simply because of the relatively large size of $\Delta r = 0.001$. 'Regular' (nonchaotic) peaks at $r > r_c$ are finite for the same reason, that is, the resolution of r rather than k/N. The finiteness of the abscissa resolution thus makes the range of observed PMI finite and the graph readable.

3.1.1 Methodology

To compute PMI at r and τ we require a set of N data points sampling the joint distribution,

$$X = \{\big(X_i^0, X_i^\tau\big)_{i=1}^N\}, \tag{3.4}$$

$X_i^0 \sim \rho_0, X_i^\tau = F^\tau X_i^0 \; \forall i$, ρ_0 uniform over the attractor.

Consider uniformly sampling the unit interval N times and evolving the outcomes for some 'settling time' $t = t_s$. Let ρ_{ts} be the resultant distribution. As $t_s \to \infty$ the support space of ρ_{ts} begins to get closer to the attractor A_r. And yet unless the basins of attraction of different components of A_r are equal in size, which in all likelihood will not be the case (and is definitely known to not be the case for some r), then ρ_{ts} will not approximate a uniform measure μ^p over the attractor.

Instead we make use of the fact that in the logistic map the attractor is ergodic. The time average of a single trajectory over the attractor will produce a uniform distribution over its elements, since the time spent in each point will just be inversely proportional to the total number of points, $p = |A_r|$. We can therefore sample μ^1 by taking the time average through an uncertain t_s. In terms of measurements this means that the first element x_0 (which will always be 0.5) is iterated for some large, $t = t_s$ time steps, and only then do the data points start being recorded. The fact that for $\tau < N$ this means looking at two overlapping sets in which points in the 'future' also double as 'pasts' only becomes a problem when the attractor is fractal (see later section).

Problems can arise if the attractor does not admit a density and yet we still want to use an estimator that assumes a sufficiently smooth distribution. Suppose the points are located ideally on the attractor. If the preferred strategy is binning then there is no problem since making the switch between probability distribution over a set and its measure is implicit within the procedure itself. If, however, the data is not perfectly converged, there will be some finite distances between the data points. A large number of bins or simply the equiprobable binning strategy with the wrong bin parameter will not see the 'true' future measure, and lead to oversampling errors.

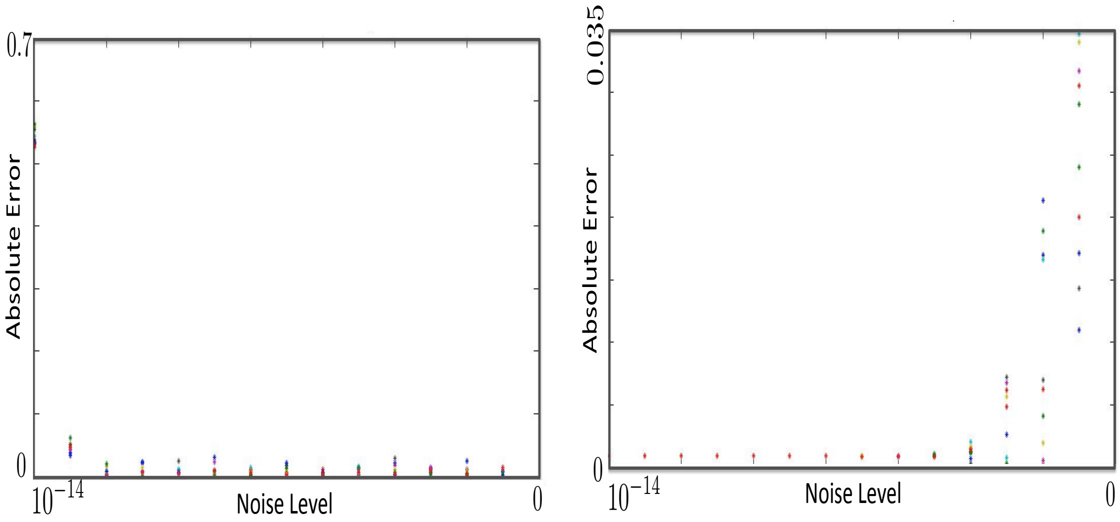
We also consider the Kraskov-Grassberger (K-G) estimator that requires distances between k nearest neighbours. If some of the data is on top of each other the distances will be zero, and it would depend on the arbitrary choices made in the way these cases were defined in the algorithm. In terms of the theory behind the estimator, this is a statement about the smoothness of the distribution which coincident points thus break. At all events the answer should not vary with k unless k is large enough to be greater than N/p.

For that purpose we dilute our data with noise. This solves the problem, since here the K-G estimator is seen to work well, giving correct $\log p$ answers with a small enough error. The noise is added after all the evolution has finished and before distances are computed; it is distributed uniformly across a small interval.

In reality the extent to which the data has converged depends on x_0 and t_s (as well as τ and p). Without adding noise it *is* possible (and realistic enough) to run into problems where the points are indistinguishable from each other as far as the *double* machine-epsilon is concerned. Practically, due to only a finite number of values accessible to a finite-precision computer, after a long time trajectories will settle on either being scattered on, or fluctuating among, a few points around the elements of the attractor. The number of these available points is small enough to lead to data points being incidental. But because of this finiteness it does not take a large enough k to decrease the PMI closer to its true value. For instance the average of the first error in fig. 3.2(a) goes down by half when $k = 20$ is considered instead of $k = 4$.

Fig. 3.2(a) shows that for the K-G estimator there are in general three error regimes depending on the magnitude of noise. When too little noise or no noise is added then for the periodic regimes PMI begins to significantly deviate from its true value, since the estimator relies on the assumption on smoothness of the density. Banded chaos is insensitive to small noise as expected. When there is too much noise it threatens to smooth over the geometry of the attractor. For period 2 its magnitude is 0.1. It effectively turns points into whole regions but PMI still gives the correct answer (albeit with the small systematic error easily attributable to N) since it looks only for the global periodicity. In this case the gap between the attractors was greater than 0.1.

The danger to close the gap is greater for banded chaos where a large proportion of the state space is already occupied, and errors set in at much smaller values, as shown in figure 3.2(b). Zooming in it is easily seen that errors begin even earlier. The visibility of this error



(a) Period 2 ($r = 3.4$)

(b) 2-Band Chaos ($r = 3.65$)

Figure 3.2: Absolute error v noise level of PMI computed using the Kraskov-Grassberger estimator with $k = 4$, plotted for 10 realisations. $N = 5000$, $t_s = 10^{10}$, $\tau = 10^5$. The machine-epsilon is circa 10^{-13}.

will increase at smaller noise levels as bands of higher periodicity might be closer together. Figure 3.3 shows how PMI changes as two strips of random numbers move closer (to be read right to left) and begin to overlap. Increasing k would make the change occur further to the right of zero. The plots show the effect that an increasingly large region of higher density has on the joint distribution.

We thus use a noise level of 10^{-12}, a reasonably small value higher than machine precision that makes periodic motion tractable, and yet low enough to still detect periodicities in banded chaos of a higher order.

At the opposite end of the spectrum is the issue of stationarity that arises when t_s is not made high enough. Here points are so far away from each other that there is little indication of the details of the actual attractor. This happens when convergence is particularly slow, for example when r is just above the period doubling points. t_s should therefore kept large, and in our examples it goes up 10^{10}. This problem is not related to estimation but rather to the question of whether the initial data actually samples the desired distribution. We mention it here to justify the need to dilute the data with noise - since this then allows us to raise t_s as much as needed.

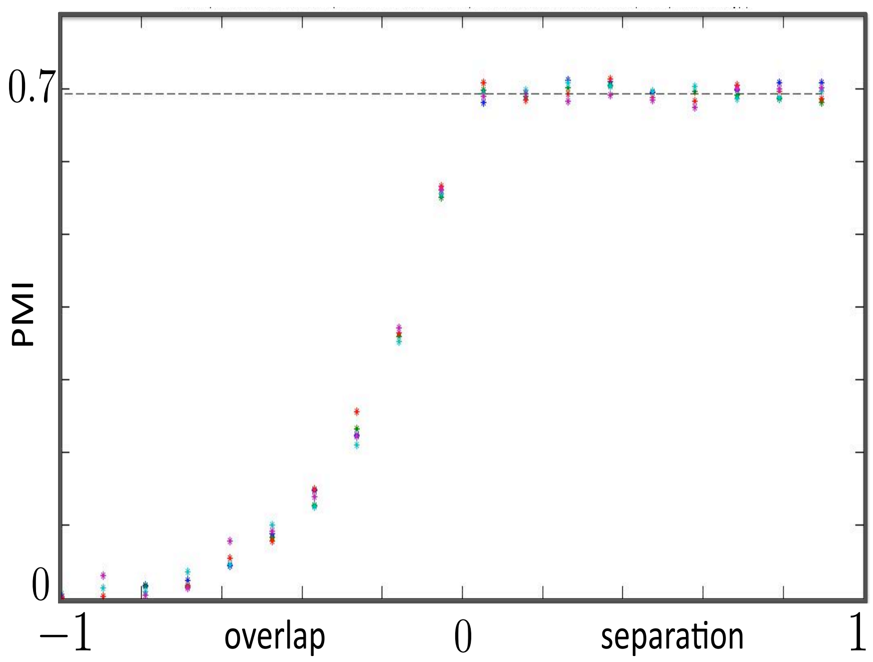


Figure 3.3: PMI between bands of uniformly distributed random numbers plotted as a function of their relative separation. Negative numbers correspond to an overlap. The dotted line is drawn at $\log 2$. PMI is computed using the K-G estimator, $k = 4$, $N = 5000$.

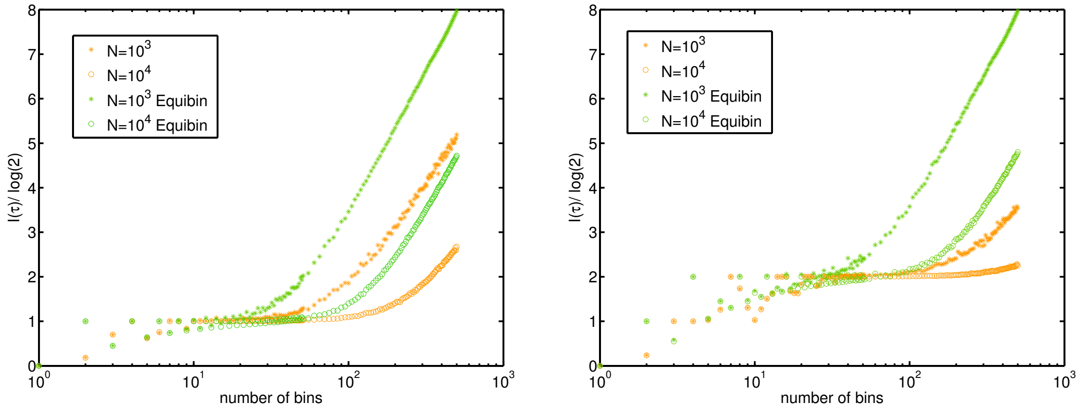
It is clear that *any* numerically-computed $I(\tau)$ will be bounded from above by the logarithm of the maximal resolvable period. The resolution used to compute PMI effectively introduces a partition that defines the observed measure. We will not be able to detect interdependency between the past and future when the motion is inside that partition. This limits from below the spatial resolution of dependency in a way that any temporal ones are limited by τ.

For the K-G estimator with k nearest neighbours this limit is the effective number of cells, N/k. Periodicities $p > N/k$ will be left undetected. Therefore for periodic motion with period p the *measured* PMI can be written as

$$I(\tau) = \log\left(\min\left[p, N/k\right]\right). \tag{3.5}$$

We now test two estimator strategies for regimes with known global periodicities p, where the analytical value of PMI is given by the logarithm of p.

Binning We first try the default method of binning the marginal and joint state spaces to compute PMI directly. We consider two strategies of partitioning with n bins: the first



(a) 2-Band Chaos ($r = 3.65$)

(b) 4-Band Chaos ($r = 3.58$)

Figure 3.4: PMI in the Banded Chaos logistic map regimes v number of bins used to estimate PMI through eq. (3.2). Normal bins have same linear size, and equibins contain the same amount of 'probability'. Settling time and τ equal to 10^4.

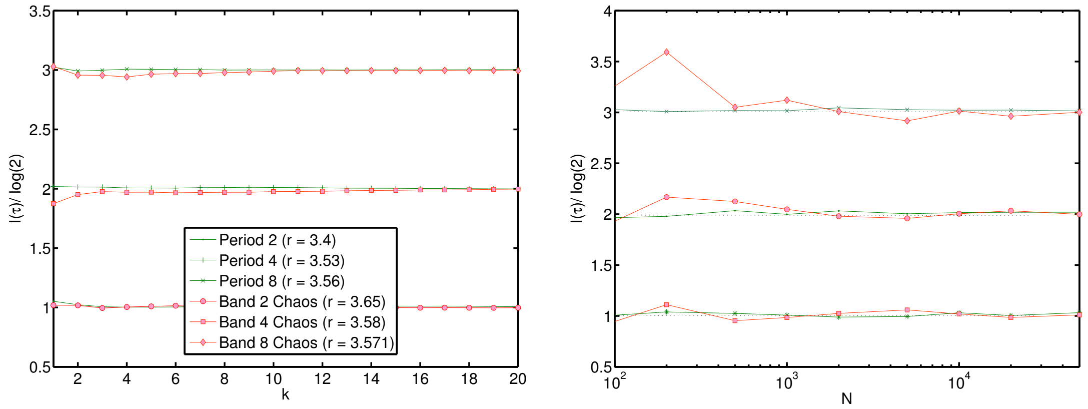
is based on each bin having an equal linear size equal to $1/n$, and the second, equibinning, results in bins of equal weights.

Equibins are seen to do worse. Errors will come in if the number of bins or the number of points is not a power of 2. Another possible reason is that some bins might be forced to straddle more than one chaotic band. It is possible to circumvent that to some extent by having a hybrid criteria putting a limit on the distance at whose expense equal frequencies are maintained, but we simply use a different method.

Figure 3.4 shows some sources of error the binning method is prone to. For this range of parameters there is a small region of n, the number of bins, where the graphs plateau before beginning the systematic increase. These plateaus happen at values equal to the logarithm of the overall periodicity, and we associate them with what the 'true' PMI should be.

From these graphs we see that there is a small range of (n, N) values that give the correct results. These depend on the binning strategy used as well as the underlying behaviour of the map. Thus, the higher the periodicity of the map, the higher needs to be the sample size N in order to be able to resolve it. The plateaux are much more cleanly defined in the 2-band chaos (fig. 3.4(a)) than in the 4-band chaos (fig. 3.4(a)). Notice the latter case also sees a low-end N increasing at an ever-growing rate and not going through a plateau at all.

The location of the optimal n range also tends to shift. For low values of n the equibinning strategy gives correct answers when n is a multiple of the periodicity, but other than that, low n is simply not able to resolve the true PMI, and is very sensitive to changes by every



(a) (b)

Figure 3.5: Kraskov-Grassberger estimate of PMI across $\tau = 10^6$ iterations for 6 regimes of the logistic map after settling time of 10^{10}, with noise of order 10^{-12}. Varying nearest neighbour index k was done at fixed sample size $N = 10^4$(a), while the reverse held $k = 3$ (b).

extra cell.

The higher end of the n scale is characterised by undersampling which almost instantly produces a drastic systematic increase in the estimated mutual information. It begins when the total number of bins (especially in the joint, as there we have n^2 bins) is large enough to render any statistics done with that finite number of points essentially meaningless. Increasing N shifts the offset to higher values of n.

Undersampling sets of at roughly the same values of n for both normal and equibinning methods, but its effects are felt more drastically in the latter case where the errors shoot up with increasing n at a higher rate.

We see that with binning, apart from the large running time, there are systematic errors that begin at parameter values that are related to the map behaviour. This, therefore, is not the optimal methodology to use for blind regimes. It does yield correct answers if N is pushed to the limit (and with a reasonable $n \approx 30$), and it suffices here, but for systems with $d > 4$ the methodology would have to be reconsidered.

Kraskov-Grassberger Method We now compute the K-G estimate of PMI, adding to final data a small noise of order 10^{-12}. Figure 3.5 shows how PMI at different regimes varies with estimator parameters of sample size N and nearest neighbour index k.

We see that for all regimes the K-G works rather well, though with more fluctuations at smaller values of N. There does not seem to be as strong a preference for k. Chaotic motion results in higher fluctuations, but that is to be expected, since the deviation from the uniform distribution is much larger in those regimes where the addition of uniform noise does not change the relative location of points to such an extent. For the range of (N, k) values we do see a systematic small bias but the absolute error is so small to render it invisible on the scale of the graphs above. This bias could become more evident at larger periodicities, but practically we do not graph results with p roughly of order greater than 10.

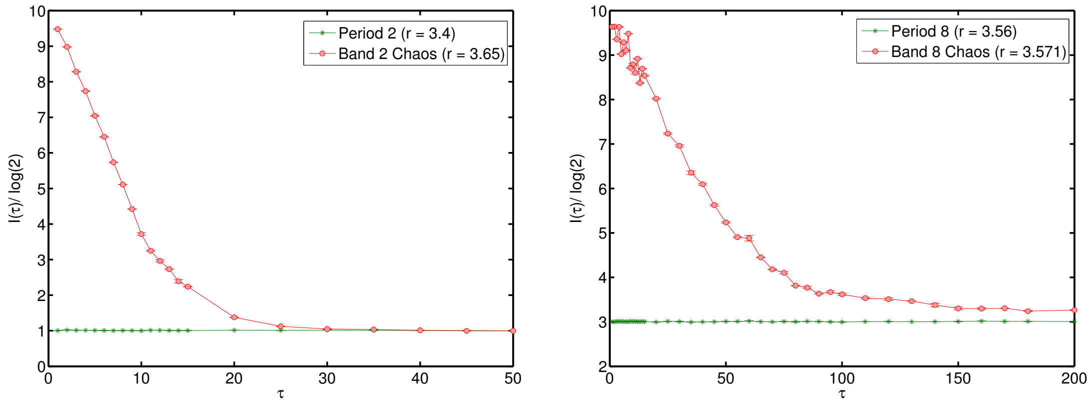
We conclude that the K-G estimator fares better than the binning strategy, and works well for both periodic and chaotic motion. For reasonable (N, k) values it does not contain over/under sampling and invariably picks up the correct periodicity (a good test is the analytic value of PMI at fully-developed bandless chaos at $r = 4$. Whilst binning with large n undersampled, the K-G estimator gave the correct answer of zero). It is optimal in terms of computation time and can easily be adapted to other systems. We therefore use it for both the logistic map and the tent map.

Thus we compute PMI using the K-G for low values of k and and N of order at least 10^3. We also note the necessity to be very careful with parameters. The given choices put an effective limit on the resolution. That means that any PPMI graph will be of finite height. Jumping ahead, it could also potentially contain peaks of different character: if t_s is not high enough there will be peaks for $r < r_c$ that will result in not sampling from the attractor - as opposed to 'true' peaks where the settled system has a good memory. Both are then limited by the resolution, which is in our case the estimator parameter.

3.1.2 PMI v τ

Here we investigate behaviour of $I(\tau)$. We assume the system is settled, otherwise if t_s is too small some of the settling will happen during the time gap, and $I(\tau)$ will pick this up. We also assume the limit of perfect resolution and leave the variation of I with τ at accumulation points for a later section.

Increasing τ raises the effective upper limit on the timescale of visible correlations. For periodic motion with finite p there is nothing to remember other than periodicity, so we expect $I(\tau)$ to be equal to $\log p \ \forall \tau \in \mathbb{Z}^+$. This is indeed supported by the graphs below where plots in green show $I(\tau)$ for periods 2 (3.6(a)) and 8(3.6(b)). Indeed we expect



(a) Global Periodicity of 2

(b) Global Periodicity of 8

Figure 3.6: Variation of PMI with τ for the logistic map parameters corresponding to motion with different periodicities. PMI computed using with $k = 5$, with $N = 10^4$, settling time of 10^5, and noise of order 10^{-12}. (Here there are no error bars as this is a single run. However, from numerous experimentation there is no reason to suppose this shows anything but the average behaviour. In addition to that, future plots estimate the gradient of the initial descent for varying r, and from figure 3.7 and later analysis we see that any variation in the individual runs will not overwhelm the general trend - though this might not be true for higher band chaos.)

this to be the case with any periodic motion independent whether it happens after the period-doubling accumulation point.

Chaotic motion is a different case. As the trajectory moves through the bands of some global periodicity p it still retains some information about its location within the band. We expect short term correlations to be present, but to die off as $\tau \to \infty$ and $I(\tau) \to \log p$. This is indeed the case and the same periodicities as in the regular case are shown in figure 3.6 above.

Looking at the way $I(\tau)$ approaches the asymptotic value in the chaotic cases one can conjecture existence of a region where $I(\tau)$ varies linearly with τ (periodic r can then be considered as slopes with gradient zero).

We estimated the slope using a small constant τ interval up $\tau < 10$ as a function of r. The result, without error bars, can be found in figure 3.7.

Notice the striking similarity with the shape of the Lyapunov exponent, shown in figure 3.8. In both plots troughs occur when there are windows of periodic motion. The plot in 3.7 is computed at slightly lower r resolution, and so does not contain that many troughs.

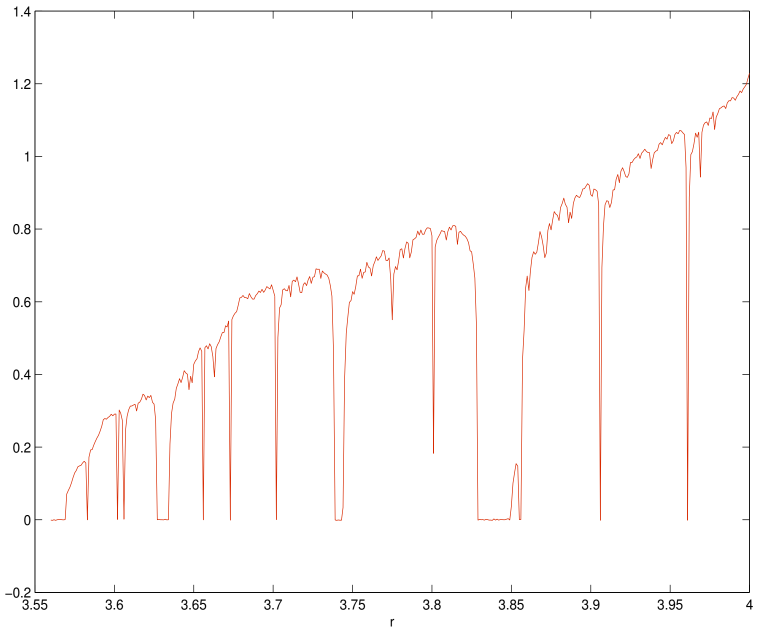


Figure 3.7: Linear approximation to the gradient of PMI (multiplied by -1) in the interval $0 \leq \tau \leq 5$ v the logistic map parameter r. $N = 10^4$, $t_s = 10^5$, $k = 5$, $\delta r = 0.001$.

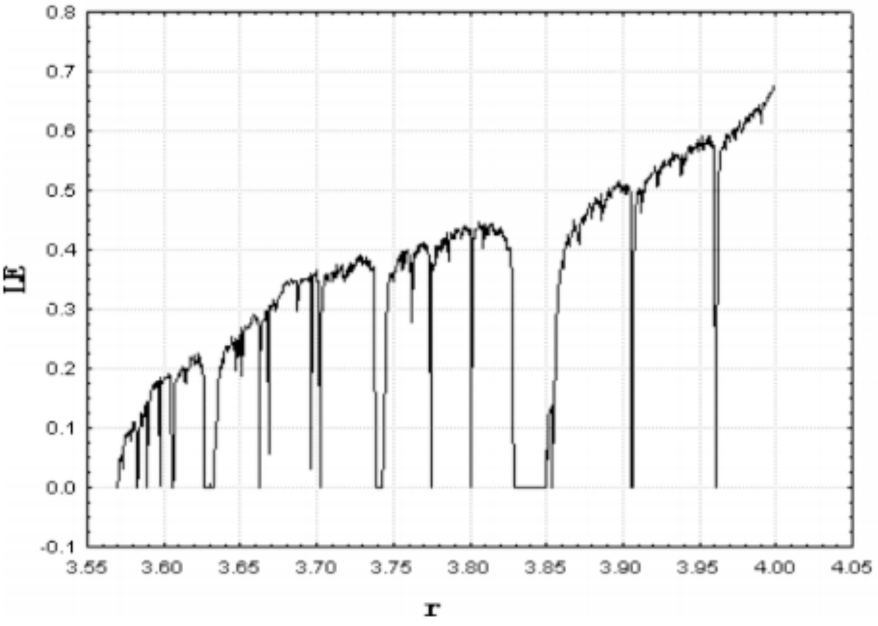


Fig. 4. The Lyapunov exponents of Logistic map (set LE = 0 for LE < 0).

Figure 3.8: Lyapunov Exponent for the logistic map, taken from Luo et al. [2009].

Notice also that the range of the abscissa in fig. 3.6(b) is four that times that of fig. 3.6(a) and yet $I(\tau)$ has still not converged to its expected value that is in line with the green plot.

Let $\Delta I = I(\tau) - I(0)$. Then using the entropy deficit expression for PMI,

$$\Delta I = H_p^\tau + H_f^\tau - H_J^\tau - H_p^0 - H_f^0 + H_J^0, \tag{3.6}$$

where the indices are self-explanatory. By definition $H_p^0 = H_p^\tau$, so

$$\Delta I = \left(H_f^\tau - H_f^0\right) + \left(H_J^0 - H_J^\tau\right). \tag{3.7}$$

In equilibrium the distribution defined by the single trajectory would render the entropy of the past equal to that of the future, so the first term disappears.
We now express the second bracketed expression using the K-G entropy estimator, $H \approx \psi(N) - \psi(k) + d\mathbb{E}[\log 2x]$, where d is the state space dimension (here 2 in the joint space), $x = \epsilon/2$ is the distance to k^{th} nearest neighbour, and here we write $\mathbb{E}$ for the average over N points. Then for a fixed k, canceling the factor of 2 in the logarithm and using the K-G estimator,

$$H_J^0 - H_J^\tau \approx 2\mathbb{E}[\log x_0^J] - 2\mathbb{E}[\log x_\tau^J]. \tag{3.8}$$

Distance x^J in the joint at $\tau = 0$ is just x_0, the distance between two k n.n. in the past. We now conjecture that there exists an interval of τ where the distance in the joint will be realised almost exclusively by the distance in the future, x_τ. If the motion is chaotic with Lyapunov exponent λ,

$$\log x_\tau \approx \lambda\tau + \log x_0, \tag{3.9}$$

so

$$H_J^0 - H_J^\tau \approx 2\mathbb{E}[\log x_0] - 2\lambda\tau - 2\mathbb{E}[\log x_0], \tag{3.10}$$

and hence

$$H_J^0 - H_J^\tau \approx -2\lambda\tau. \tag{3.11}$$

So $\Delta I \approx -2\lambda\tau$. Approximating the gradient of I with τ at $\tau = 0$ we derive the bound

$$\frac{d(I(\tau))}{d\tau} \approx \frac{\Delta I}{\tau} \approx -2\lambda. \tag{3.12}$$

In other words, the negative linear approximation should be roughly equal to twice the Laypunov exponent, *for small time separations* τ. That is what we indeed see when comparing figures 3.7 with 3.8.

There is an interesting point to be made here. The less global periodicity there is, the faster PMI converges to its asymptotic value. Here we used the same range of τ to estimate the gradient, independent of r. Therefore the errors at relatively small r values are significantly larger (compare subgraphs of 3.6). This analysis could therefore be made much more precise by simply estimating the gradient from the entirety of the linear range. Since this involves finding the upper limit of the latter it could also then be used to investigate the speed of convergence as a function of r.

3.2 Resolution-Dependent PMI

It is clear that at all times $I(\tau)$ depends on the resolution. We have seen how it effectively puts an upper limit on the observed periodicity. It does not matter so much in the three simple cases we have observed: periodic behaviour, fully-mixing behaviour at $r = 4$, and the hybrid case. However, for the accumulation points PMI, which we have seen to be the logarithm of the overall period, thus becomes simply infinity. New way of interpreting the results is needed. The manner in which $I(\tau)$ changes with resolution is directly related to the information dimension of the underlying spaces. Here we derive this and express our results in terms of a new quantity, the Information codimension, which we introduce in order to express the resolution in terms most appropriate to the preferred estimator.

PMI for Fractal Measures

The differential entropy H is defined as

$$H[\rho] = -\int_{x \in E} \log \rho(x) \rho(x) d^d x, \tag{3.13}$$

where E be Lebesgue-measurable, and ρ is a normalised continuous measure density on E. This implies that E is a subset of $\mathbb{R}^d$, and is either bounded or has finite measure. We can also partition E into cells of size $v = \int_{x \in \text{cell}} d^d x$, and define a discrete measure μ on the partition P through

$$\mu_i = \int_{x \in C_i} \rho(x) d^d x,$$

where $P = \{C_1, C_2, ..C_m\}$.

The number of such cells is then $m = \int_{x \in E} d^d x / v$.

The Shannon (discrete) entropy of μ is then

$$S(\mu) = -\sum_{i=1}^{m} \mu_i \log \mu_i. \tag{3.14}$$

To emphasize the fact that μ is a result of a partition, and that hence S depends on the partition P, we will sometimes write S_ϵ, where $\epsilon = v^{-d}$.

Linking Discrete and Continuous Entropy forms

By the Integral Mean Value Theorem, assuming ρ is continuous, there exist points $x_i \in C_i$ such that $\mu_i = \rho_i v$, where $\rho_i = \rho(x_i)$. For $v > 0$

$$S(\mu) = -\sum_{i=1}^{m} \rho_i v \log(\rho_i v) = -\sum_{i=1}^{m} (\rho_i v \log \rho_i + \rho_i v \log v)$$

Then from (3.13), for v small we have

$$S(\mu) \approx H[\rho] - \log v.$$

If, however, ρ is not assumed to be continuous, then $\mu_i = \rho_i v$ *defines* the effective value of ρ_i as an approximation of μ_i for that box. We can then *define*

$$H[\rho] = S(\mu) + \log v.$$

Since in a d-dimensional space v gets replaced by ϵ^d, this can be written as

$$H[\rho] = S(\mu) + d \log \epsilon. \tag{3.15}$$

Thus $H[\rho]$ can diverge with resolution.

Entropy in terms of resolution

Information dimension D of a distribution ρ is defined as

$$D = \lim_{\epsilon \to 0} \frac{\sum_i \mu_i \log \mu_i}{\log \epsilon}$$

or

$$D = \lim_{\epsilon \to 0} \frac{-S(\mu)}{\log \epsilon}. \tag{3.16}$$

Factorising (3.15),

$$H[\rho] = -\log \epsilon \left(\frac{-S(\mu)}{\log \epsilon} - d \right),$$

so that, substituting in (3.16), we get

$$H[\rho] \approx (d - D) \log \epsilon, \tag{3.17}$$

as a statement of the manner in which $H[\rho]$ changes with resolution for small ϵ limit. The same is obtained through

$$H[\rho] = -\sum_i \mu_i \log\left[\frac{\mu_i}{\epsilon^d}\right] = d\log\epsilon - D\log\epsilon + \text{const.} \tag{3.18}$$

Recall Persistent Mutual Information is defined as

$$I(\tau) = H[\rho_0] + H[\rho_\tau] - H[\rho_{0,\tau}]. \tag{3.19}$$

Let the support spaces of the marginal distributions ρ_0 and ρ_τ be partitioned by boxes of linear size ϵ, as above. Then

$$H[\rho_0] = d_- \log\epsilon - D_- \log\epsilon + \text{const}, \tag{3.20}$$

$$H[\rho_\tau] = d_+ \log\epsilon - D_+ \log\epsilon + \text{const}, \tag{3.21}$$

and

$$H[\rho_{0,\tau}] = d_{-+} \log\epsilon - D_{-+} \log\epsilon + \text{const.} \tag{3.22}$$

Then

$$I(\tau) = (d_- + d_+ - d_{-+})\log\epsilon - (D_- + D_+ - D_{-+})\log\epsilon + \text{const.} \tag{3.23}$$

Here d is the box-counting dimension of the embedding space. Since $d_{-+} = d_- + d_+$, we have

$$I(\tau) = -\left(D_- + D_+ - D_{-+}\right)\log\epsilon + \text{const.} \tag{3.24}$$

Hence PMI scales with the logarithm of the characteristic partitioning size of the support spaces.

There can potentially be some ambiguity in both notation and concepts for this case when PMI increases with resolution indefinitely. One option is to say that the actual PMI is then not defined, and $I(\tau)$ merely characterises the *manner* in which PMI tends to infinity. Another is to consider the limit of $I(\tau)/\log(\epsilon)$, which does exist. It is perhaps easiest to do the former, especially since the marginal and joint D as defined in the limit of infinite resolution will always be equal, an not very interesting limit. Therefore we keep in mind when talking about $I(\tau)$ that we actually mean $I(\tau, \text{resolution})$, and that the information dimensions merely express the manner in which quantities increase.

This can be rewritten in terms of another partitioning that better reflects the process through which we obtain the results - through equipartioning of the probability distribution.

Entropy in terms of probability resolution

In the sections above we partitioned the space E into cells of equal volume. Since ρ is arbitrary the μ_i need not be equal. Alternatively we can require the μ_i to all be equal, and partition $redE$ accordingly. The measure of every cell is then the reciprocal of the total number of cells m, $\mu_i = \hat{\mu} = 1/m \; \forall i$.

Since cells are now allowed to vary in size, $\epsilon = \epsilon_i$, and eq (3.17) for entropy does not hold. Recall that it was

$$H[\rho] = -\sum_c \mu_c \log\left[\frac{\mu_c}{\epsilon^d}\right] = d \log\epsilon - D\log\epsilon + \text{const} \qquad (3.25)$$

It can be rewritten in terms of $\hat{\mu}$ through inverting eq (3.16):

$$\log\epsilon \approx \frac{-S(\mu)}{D}.$$

Since μ is now an equidistribution, its discrete entropy $S(\mu)$ is equal to the logarithm of the number of underlying cells, $\log m$, and so

$$\log\epsilon \approx \frac{-\log m}{D},$$

leading to

$$H[\rho] \approx \frac{-(d-D)}{D} \; \log m.$$

The number of cells m can be though of as controlling the resolution of probability ρ.

PMI through Local Probability Resolution

It follows that

$$I(\tau) = \left(\frac{D_- + D_+ - D_{-+}}{D_{-+}}\right) \log m + \text{const}. \qquad (3.26)$$

Here m is the number of cells that contain equal probability. The K-G estimator we use for Shannon entropies has for a parameter the number k of nearest neighbours to which each point looks. Thus a set k corresponds to an effective partitioning of the probability distribution into N/k cells of weight k/N. Hence PMI can be rewritten as

$$I(\tau) = I(\tau_0) + \Gamma \log\left(\frac{N}{k}\right), \tag{3.27}$$

where we define

$$\Gamma(N,k) = \frac{(D_- + D_+ - D_{-+})}{D_{-+}} \tag{3.28}$$

as the information codimension. Note that here we admit the possible dependence of the relevant information dimensions on τ, which implicitly defines the underlying measures.

Thus for systems whose joint information dimension is not just the sum of the marginal ones Persistent Mutual Information should scale logarithmically with probability resolution. This is in contrast to 'simple' regimes of dissipative systems like logistic map studied above. When the attractor consists of a finite number of points as is the case for period-p cycle, information dimensions of all the support spaces are the same (which also happen to be zero). For any τ, which due to our definition of PPMI does not have to be finite, higher sample sizes only ensure PMI converges to $\log(p)$ in a manner specific to the estimator used.

This can be interpreted in terms of ensembles. Flat initial distribution sampled with N points can be considered as being equivalent to starting with N closed systems. By the optimal 'lack of information' argument we then assume that the distribution out of which the systems were picked was flat. PMI then corresponds to the average information about the future that would be obtained should one of the systems be examined. For non-fractal attractors increasing the number of samples becomes, after some N^* e.g.$>p$ pointless, in the sense that this average value would not change. PMI dependency on sample size of this form would manifest itself in the average information about the possible future state increasing without end at a logarithmic rate.

3.2.1 Resolution Dependency at the Accumulation Point

We now compute the PMI at the period-doubling accumulation point r_c. Figure 3.9 shows the result for a variety of resolution ranges that we control by varying k (computationally faster than increasing N, it at the same time lowers the errors).

The expected slope is that of unity. The attractor is a Cantor set with some infor-

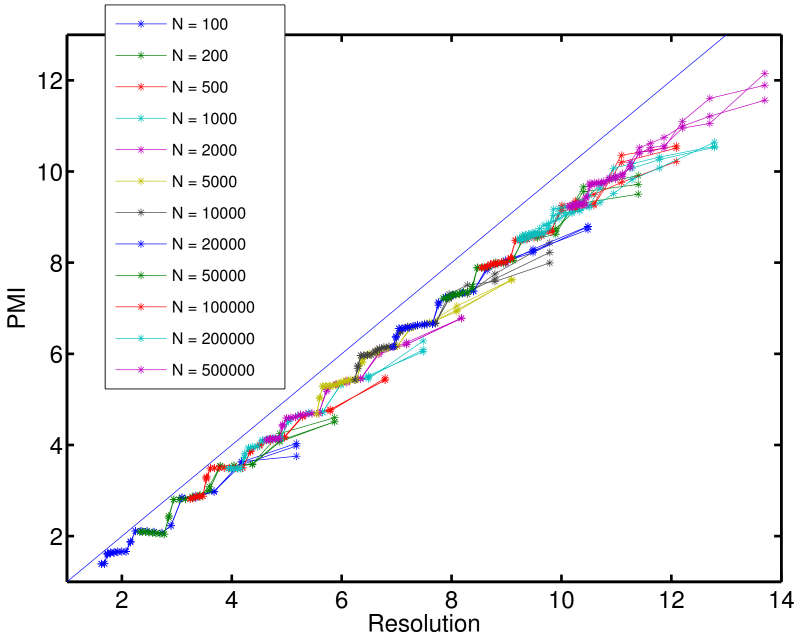


Figure 3.9: Kraskov-Grassberger estimate of PMI as a function of the resolution $\Psi(N)-\Psi(k)$ at the accumulation point r_c of the logistic map, with added noise of order 10^{-12}. Settling time and time gap τ are all 10^4; nearest neighbour index $1 \leq k \leq 20$.

mation dimension D, $D = D_- = D_+ = D_{-+}$. Therefore the information codimension Γ, which controls the slope, is simply unity.

We do indeed see that the trends follow the slope of unity, but then begin to decline. This is unexpected in that an apparent decline can be interpreted as the start of convergence towards some finite PMI value. By definition at r_c the periodicity is infinite and thus PMI should not stop increasing.

A possible reason is that the (floored) finite precision value with which we approximate r_c will necessarily give a finite periodicity. Yet when this is tried for the very low approximations to r_c, associated with periodicities visible on the scales given above, PMI converges in an abrupt manner, very different to the one observed in figure 3.9. Also here r_c is given to 94 decimal places, which by trial and error we know to give the period (whether regular or chaotic) higher than the range of observed ordinate values.

Neither is it the case that our resolution limits the 'visible' periodic dependencies, since then PMI would settle with resolution in the same sudden manner as described above.

In order to understand what factors effect the change in slope we vary several parameters. Figure 3.10 shows results for higher values of t_s, and τ. Plots are seen to follow the expected slope for longer, and from examining further variations we conclude that the

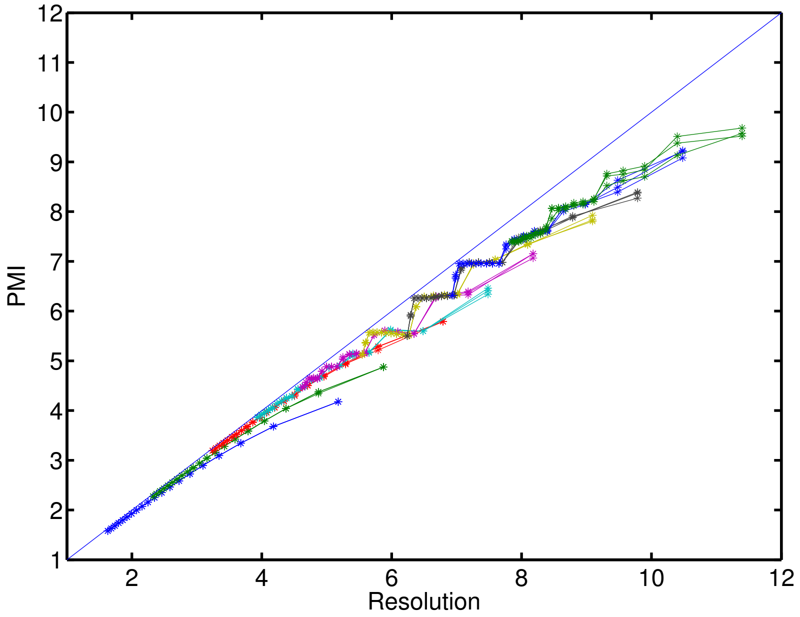


Figure 3.10: Same details and legend as for figure 3.9, but settling time and time gap τ are all 10^9.

main cause of this is τ.

The reason why increasing τ leads us to resolve higher periodicities lies in the specific way we collect data. The methodology section above justified using a single trajectory and taking consecutive time steps as independent initial positions on the attractor. We thus have at most $(N + \tau)$ sequential datapoints, of which we collect, again, at most $2N$.

The way the trajectory arranges itself on the attractor is related to its fractal nature. Every second point of the trajectory will be in some portion of the state space. Every fourth point will come back closer. Every eighth point will be even closer. Thus to detect higher periodicities a longer and longer trajectory is needed. As a result the maximum resolvable periodic will be a function of $(N + \tau)$. The further the 'past' and 'future' are separated, the better will be the resolution of the underlying attractor. In order to see the plots begin to deviate from the expect slope a higher resolution range is needed for a higher τ.

It is also interesting to see the step-wise manner in which the plots increase for the low end of the resolution scale. To some extent this is equivalent to the oversampling part of the plots when PMI was computed using the binning strategies. Here increasing the resolution only changes the PMI when the effective neighbourhood size is small enough to only resolve the higher periodicities. The fact that the jump appears discontinuous indicates that there is a spatial gap between points that are near to each other every p^{th} step and those that are near to each other every $(p+1)^{th}$ step. Especially in the first figure 3.9 it is possible to see that the jumps correspond to $I(\tau) = \log(p)$, as expected from the period-doubling behaviour.

3.3 Permanently Persistent Mutual Information

PPMI, or permanently persistent mutual information, is defined as

$$I(\infty) = \lim_{\tau \to \infty} I(\tau). \tag{3.29}$$

It represents the information that does not get eroded away but is ultimately preserved across time. We consider PPMI in the context of measures of (strong) emergence. This section relates the observed PMI for the logistic map to PPMI, and concludes with the corresponding analysis of the tent map as another 1D example.

For the majority of the logistic map regimes $I(\tau, t_s, \text{resolution})$ decayed with τ to some constant asymptotic value. The speed of this is varied but was generally much slower at values of r corresponding to chaotic bands of high periodicity. Also, unless the settling time was high enough PMI would display (otherwise transient) peaks after period-doublings. Convergence speed also varied between chaotic and regular regimes - being almost instantaneous in the latter. We conclude that for most regimes the limit defined in the equation above does in fact exist, though when the underlying measures are fractal the definition above needs to be supplemented by some (finite) resolution at which the infinite τ limit is taken. We make the same assumption for the tent map.

3.3.1 Example 1: the Logistic Map

We associate the plot in figure 3.1 with PPMI, since it is done for a τ value large enough for PMI to have converged (checked heuristically). As expected, for each r it reflects the extent of the overall periodicity. Figure 3.11 shows the main qualitative result in the PPMI of the logistic map: the symmetry with which periodicity is picked up on both sides of the period-doubling accumulation point. We clearly see the doubling of the period as r approaches r_c. Equally well we see PPMI decreasing in steps of the same magnitude. These represent the bands merging together.

PPMI also picks up an interesting feature in the manner these mergers happen. Looking at the bifurcation diagram it is not unreasonable to assume that there exists an overlap region between one merger and the next where more and more of one band pair covers the same state space as another. In other words, that the range of trajectory motion will, for that band, increase with r.

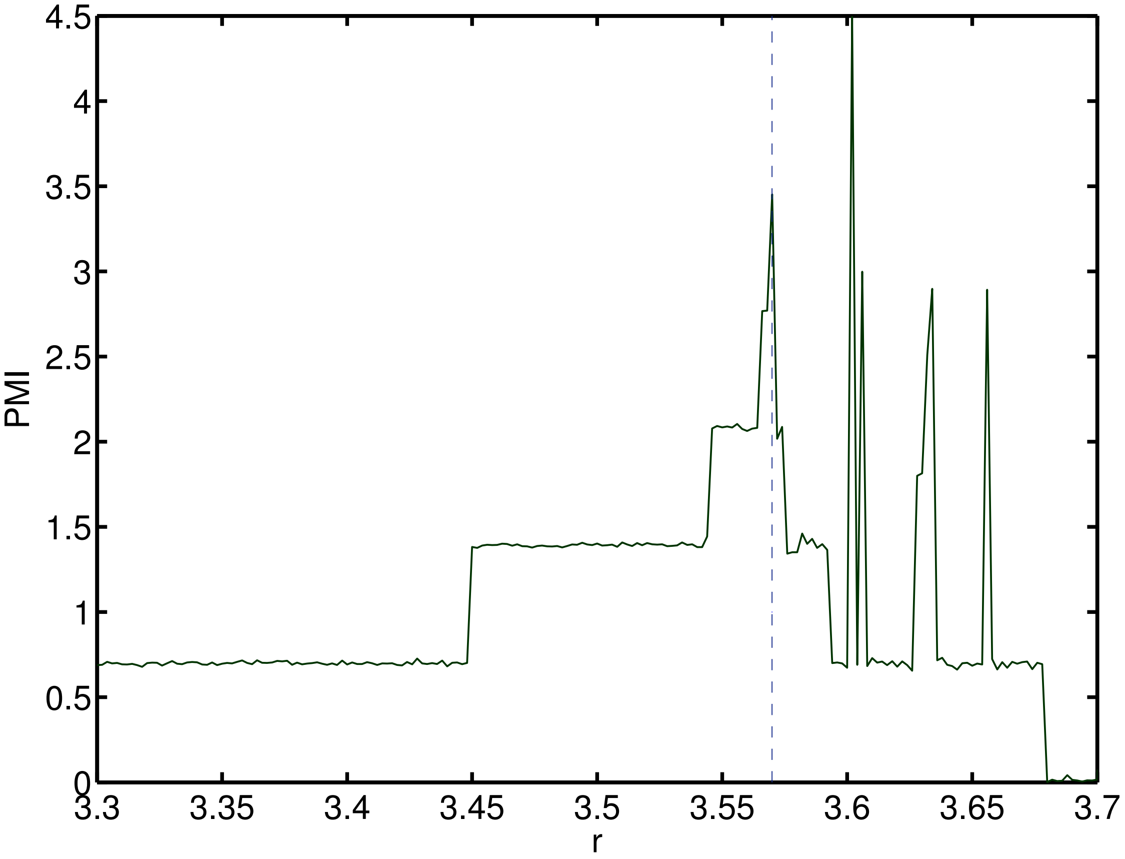


Figure 3.11: Kraskov-Grassberger estimate of PMI across $\tau = 10^5$ iterations for the logistic map after settling time of 10^5, with noise of order 10^{-12}. Sample size $N = 5000$, estimate done at nearest neighbour index $k = 4$. The dotted line is drawn at $\approx r_c$, the onset of chaos.

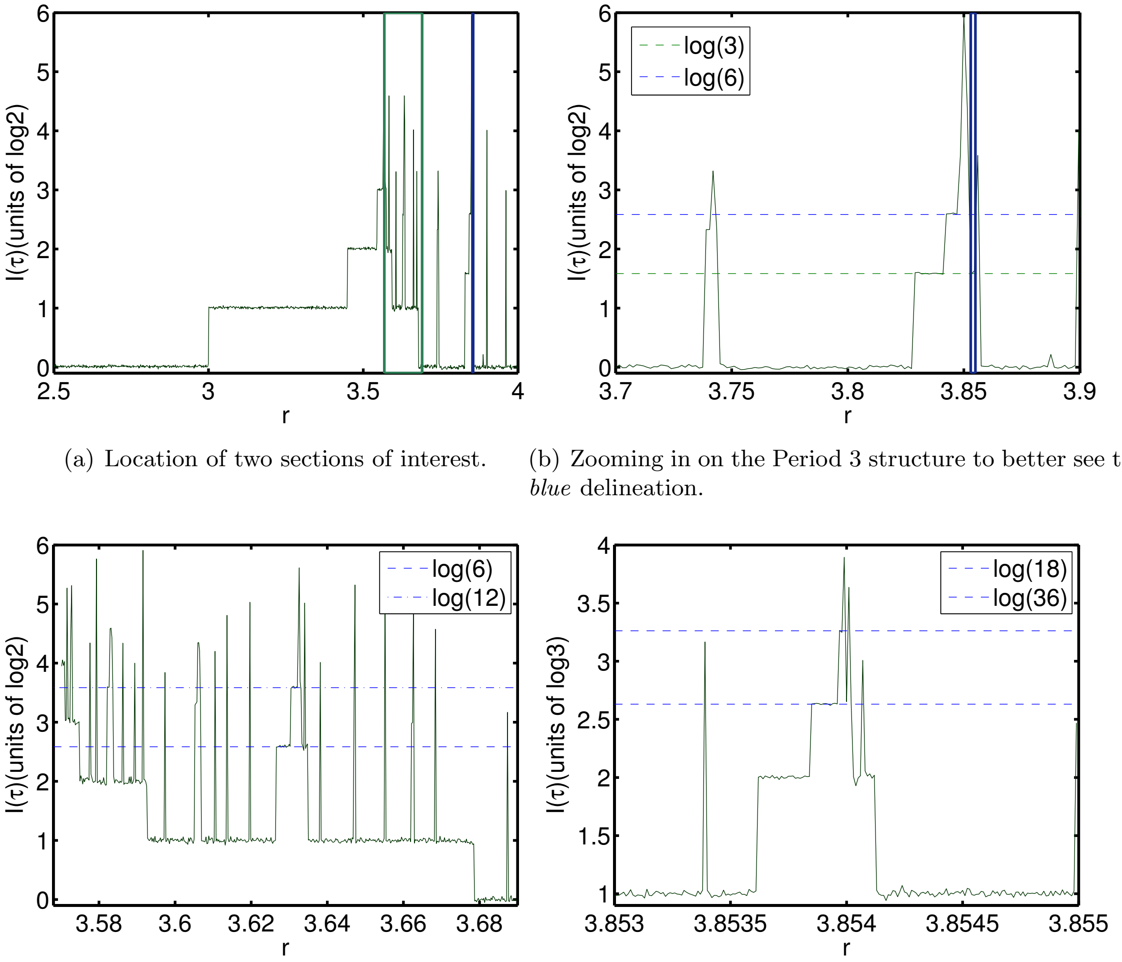
Figure 3.3 shows that when two bands (albeit of random numbers, but here PMI treats chaos as noise) increase their overlap the PMI changes to zero also smoothly. Analytically this is accounted for through a region of the joint supports the overlap region and hence has double the weight. That is not what happens after bands merge together. PMI immediately jumps to a value corresponding to half the periodicity, indicating that the amount of new 'space' available to the trajectory is, if not the entire other band itself, at least constant throughout that r range.

This analysis also allows us to see that the chaotic regime is infinitely rich in its behaviour. Figure 3.12 illustrates what PPMI can pick up by focusing on two ranges of $r > r_c$ values.

The green section is very narrow. To appreciate its position fully we show it again in fig. 3.12(c), which is a more detailed picture of the chaotic regime. The period-three structure is clearly visible. It is now plain that our section of interest lies on the right-hand, 'chaotic' side of the period-three structure. Before moving on note that the left-hand side of the structure displays period doubling, jumping from 3 to 6 (in a periodic manner, though that of course is not evident from PMI). Figure 3.12(d) shows PMI of the section in question, normalised by log(3). The background of 1 corresponds to the period-three regime. We then observe one period *tripling* (to log(9)) followed by two period doublings (to log(18) and log(36)).

This trend of periodicity tripling on the right-hand side of a peak is also observed in fig. 3.12(c), which shows a structure built on a band-two chaos. The two initial steps correspond, just as above, to a tripling followed by a doubling of the period.

PPMI is thus a powerful tool for detecting such periodicities. There is no increase in computational cost, the only limit being the width of the increment δr. Its lower (unobtainable) bound is given by the machine-epsilon, but in practice the numerical nature of each step in the algorithm that makes the sampled map many-to-one will somewhat raise it.



(a) Location of two sections of interest.

(b) Zooming in on the Period 3 structure to better see the *blue* delineation.

(c) Inside the section delineated by *green* above

(d) Inside the section delineated by *blue* in the two plots above.

Figure 3.12: Kraskov-Grassberger estimate of PMI across $\tau = 10^6$ iterations for 6 regimes of the logistic map after settling time of 10^{10}, with noise of order 10^{-12}. Unless stated otherwise the drawn lines are normalised by the same unit (log(2) or log(3)) as the data, making the argument in the logarithm equal to the periodicity.

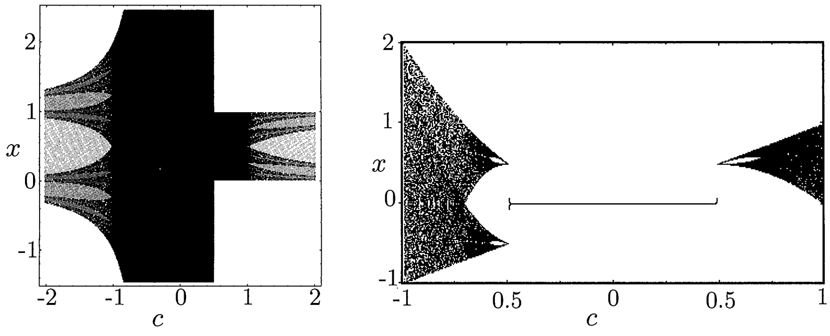


Figure 8. Divergence and Feigenbaum diagrams for the tent maps.

Figure 3.13: The Divergence and the Bifurcation Diagrams for the tent map, taken from Rickert and Klebanoff [1999]. Here $\mu = 2c$.

3.3.2 Example 2: the Tent Map

The tent map is a linear approximation to the logistic map that displays some similar features such as period-doubling.

$$x_{n+1} = \begin{cases} \mu x_n, & x_n < 1/2 \\ \mu(1 - x_n) & x_n \geq 1/2, \end{cases} \tag{3.30}$$

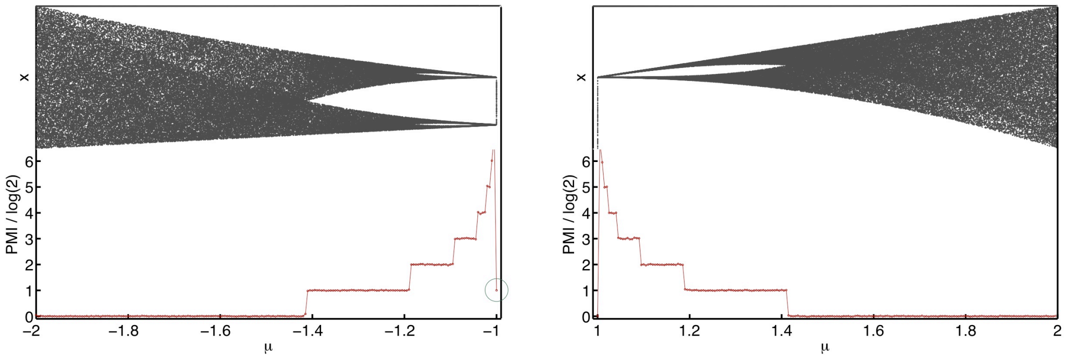
The parameter μ can be positive or negative. For $0 \leq \mu < 1$ all orbits are attracted to zero, at $\mu = 1$ the attractor is $[0, 1/2]$. For $1 \leq \mu < 2$, $x_i \in [0, 1] \forall i$. Excluding 1 also excludes any periodic motion, and until $\mu = 2$ the periodicity of the bands halves in the same manner as in the logistic map. At higher μ trajectories are no longer confined.

Negative μ shows a qualitatively similar picture of period-doubling, with two key differences. The first is that for $-2 < \mu \leq -1$ all orbits are contained within $[\mu/2, \mu^2/2]$, which is easy to see by considering the cone that represents that map and is produced by negative μ. The second difference is that now at $\mu = -1$ there is a set of points that converges to zero, and the second set that is periodic with period 2.

Here we study values of μ for which the trajectories do not diverge. This corresponds to the regions that the Divergence diagram in figure 3.13 shows in black.

Figure 3.14 shows PMI for the tent map for a range of positive and negative μ-parameter values where trajectories do not diverge. As expected it picks up the global periodicity p, rendering $I(\infty) = \log(p)$.

When μ is in the interval between the two figures, $-1 < \mu < 1$, all orbits are attracted to



(a) $-2 < \mu \leq -1$

(b) $1 \leq \mu < 2$

Figure 3.14: Kraskov-Grassberger estimate of PMI for the (symmetric) tent map $F(\mu)$ with the respective bifurcation diagrams. The latter computed at $t =$ settling time $= 10^4$. Time gap $\tau = 10^4$, sample size $N = 10^4$, PMI found as nearest neighbour index $k = 5$, as an average of three runs (errors miniscule compared to plot). Added noise is of order 10^{-12}. PMI $I(\tau, \mu = -1) = \log(2)$ is circled. Note the relatively small number of points used for each μ in the bifurcation diagrams rendered the $\mu = -1, 1$ attractors as having holes on the visible scale, which is not the case. The size of the attracting domain varies.

fixed point at 0, and PMI (as logarithm of a unit period) would, respectively, be also equal to 0. It is also 0 when all the chaotic bands have merged together and any periodicity is no longer resolvable. We do not anticipate *existence* of any periodicity at such, since it must then, as $mod(\mu)$ increases, occur suddenly and be of at least $\log(N)$.

Consider behaviour of the tent map at $\mu = -1$ and $\mu = 1$. There the bifurcation diagram shows seemingly similar behaviour, yet PMI values differ. The would-be continuous lines covering different intervals do in fact result from two different behaviours: at $\mu = 1$ almost all points below $x = 1/2$ are attracting points. This lack of periodicity gives the observed PMI of 0. $\mu = -1$, on the other hand, forces the existence of an (observably large) range of points with period 2, which corresponds to a $\log(2)$ we see in the PMI plot below. Other than that, there is an almost exact correspondence between the PMI for positive and negative μ - notice that the magnitude of μ at stepping values coincide.

As an aside, the $\log(2)$ point at $\mu = -1$ can be thrown away if for negative μ the map gets substituted by two consecutive iterations. The result in shown in figure 3.15. First, all the period-two behaviour now gives a PMI of zero. Second, we obtain extra evidence for the conjecture that once the bands merge there is mixing on the full-scale - that the entirety of the state space accorded to the other 'arm' is no available to the original trajectory. This creates the non-smooth change in the bifurcation diagram.

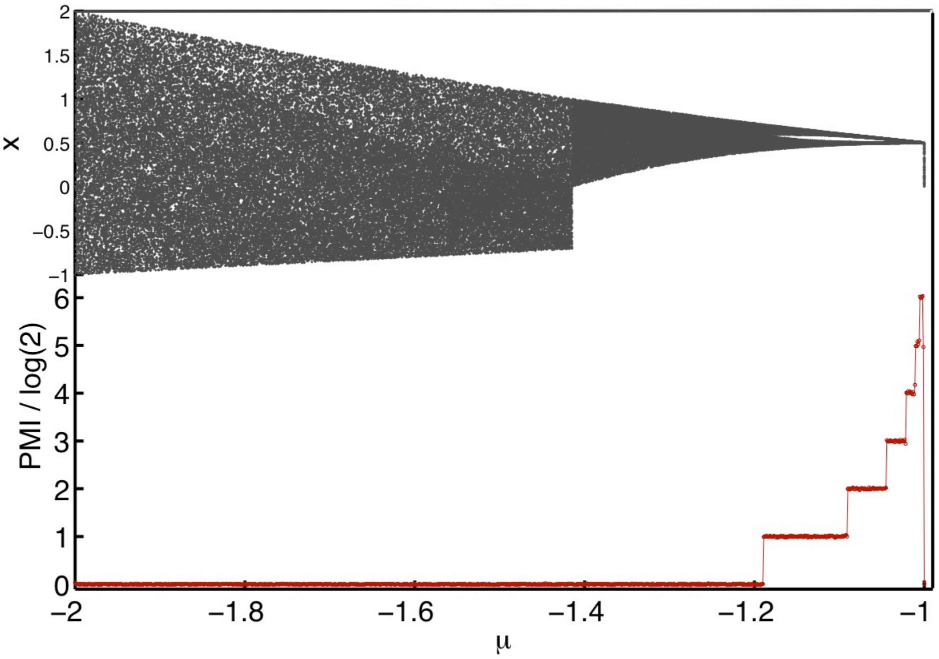


Figure 3.15: Same as figure 3.14(a) but only registering every second iteration.

This is the key feature that separate the bifurcation diagram from the PMI - the former is essentially a marginal quantifier, whereas PMI is able to pick out the dependencies present in the attractor.

We can associate the PMI shown in the figures above to PPMI. The caveats here are the same as in the logistic map case, that is, the range of μ values for which PMI is computed does not cover the higher periodicities (close to $\mu = -1, 1$), and so to associate the graphs with PPMI the steps have to be, in logarithmic fashion, mentally continued up to infinity. Variation of parameters such as settling time, τ, or resolution does not alter the figures, neither are they obscured by peaks of slow relaxation that were present in the logistic map parameters around period-doubling values.

3.3.3 PPMI as a measure of Emergence

We now discuss the difference between PMI and mutual information between the past and future of a system that is not yet settled, i.e. where the initial ensemble distribution is over the domain of the logistic map. This we do through considering a simplest case of a period 2 attractor.

Let $B(A)$ be the basin of attraction of point $x = A$ at some $n = \tau$, and its complement

$B(A') = X/B(A)$ that of $x = A'$. Then

$$I(\tau) = S(\mu_0\,(B(A))\,, \tag{3.31}$$

where μ_0 is the initial measure, and $S(a) = -a\log(a) - (1-a)\log(1-a)$ is the binary entropy function. If the initial measure is uniform over the attractor - as ranged by a single trajectory - then, as we have seen, $\mu_0\,(B(A)) = \mu_0\,(B(A')) = 1/2$, and $I = \log 2$ independent of τ. If, however, our initial state of absolute lack of knowledge is about the *unsettled* system, the initial distribution will be over the whole map domain. As such, it will depend on the Borel measure of the basins.

In the logistic map $\mu_0\,(B(A))$ does not in general equal to $\mu_0\,(B(A'))$; given absolute inital uncertainty more orbits will end up in one phase than another. As such $I(\tau)$ will, in accordance with the concave S, decrease to below $\log 2$.

As such, for any period-p regime, $I(\tau)$ would not give the overall periodicity, but rather - especially if $p > 2$ - a complex entanglement of the weights of the basins of attraction.

On the one side this is a valid measure for this scenario. It is our choice to focus on the settled system, and more importantly to presume to extract the information from a single trajectory only. This is in line with the definition of strong emergence we choose to adopt, that is, forecastability rather than predictability. On this level the periodic behaviour is directly comparable to the banded chaos regimes, since what they have in common is the phase.

Chaos is sometimes defined as motion that loses information about the initial condition in a very specific manner. The way it happens a finitely resolved past should hold absolutely *no* information about the future that is removed by some finite τ_l. PMI, with its emphasis on distributions over ensembles, places an uncertainty on the initial condition, effectively changing a perfectly resolved past into points with finite resolution. PPMI then considers the future removed further than τ_l. We know that causal correlations that are only the result of chaos will not persist for longer than this limit. Therefore PPMI, independent of the resolution, does not see chaos, and treats it as noise. Resolution begins to matter when there are structures the trajectories remember for all times.

In contrast to the bifurcation diagrams PPMI is thus directly informative about the clock. In the tent map, it differentiates between the ostensibly similar $\mu = -1$ and $\mu = 1$ cases, but does not do so for a broader set of μ values. Contrast PPMI to the clearly different bifurcation diagrams in figure 3.14 which neverless, by this measure, turn out to be of systems with the same forecast power. We conclude that in the Tent map PPMI would be a good measure of strong emergence.

In general, in the maps where the attractor essentially introduces a phase difference, it is exactly that information that could be potentially of use in order to forecast the future. Any information about a finite initial condition will be lost after a finite number of iterations. Thus, given some uncertainty in the knowledge of the system in the first place, it is the periodicity that renders prediction possible. For any system with no structure in the motion other than regularity/chaos *and* some periodicity, PPMI is thus the logarithm of the total number of available phases.

Chapter 4

Persistent Mutual Information in the Standard Map

4.1 General Behaviour and Error Analysis

Our aim is to estimate Persistent Mutual Information for the system evolving under the standard map. In the Introduction we described the main features of this map that often functions as a toy model, an archetype of area-preserving dynamical systems. Here we conceive of an ensemble of such systems whose states at time t are distributed according to ρ_t. We will assume a uniform ρ_0 and use the entropy deficit expression for the PMI (eq. (2.1)), which is thus defined as

$$I(\tau) = H[\rho_0] + H[\rho_\tau] - H[\rho_{0,\tau}], \tag{4.1}$$

where H is the Shannon entropy of eq. (1.4) and $\rho_{0,\tau}$ the joint distribution.

More formally, let $X = [0, 2\pi)^2$ be the state space of the standard map, where we associate the sides and consider dynamics on a torus. This can be turned into a measurable space, and since X is continuous these measures can be expressed through densities, or probability distributions, and associated with the probability distribution over the ensemble. Let ρ_0 and ρ_τ be the initial and final densities on X,

$$\rho_\tau = F^\tau \rho_0, \tag{4.2}$$

where F is the standard map evolution operator on densities. The joint distribution $\rho_{0,\tau}$ is then obtained through considering the conditional.

Eq. (4.1) is particularly suitable for area-preserving systems such as the standard map. A flat initial distribution stays flat for all times, which means that normalising the linear size of X makes contribution from the marginal entropies disappear, leaving

$$I(\tau) = -H[\rho_{0,\tau}]. \tag{4.3}$$

Thus Persistent Mutual Information in normalised, area-preserving (Hamiltonian) continuous systems is simply the entropy of the joint distribution.

The joint distribution $\rho_{0,\tau}$ cannot be obtained analytically. Entropies and the various other mean values of interest have to be additionally estimated using samples drawn from $\rho_{0,\tau}$. In this sense eq. (4.3) simplifies computation of PMI as far as current estimator research is concerned: unbiased entropy estimators are more common and their properties understood better than estimators of mutual information with their potentially additive errors. Thus,

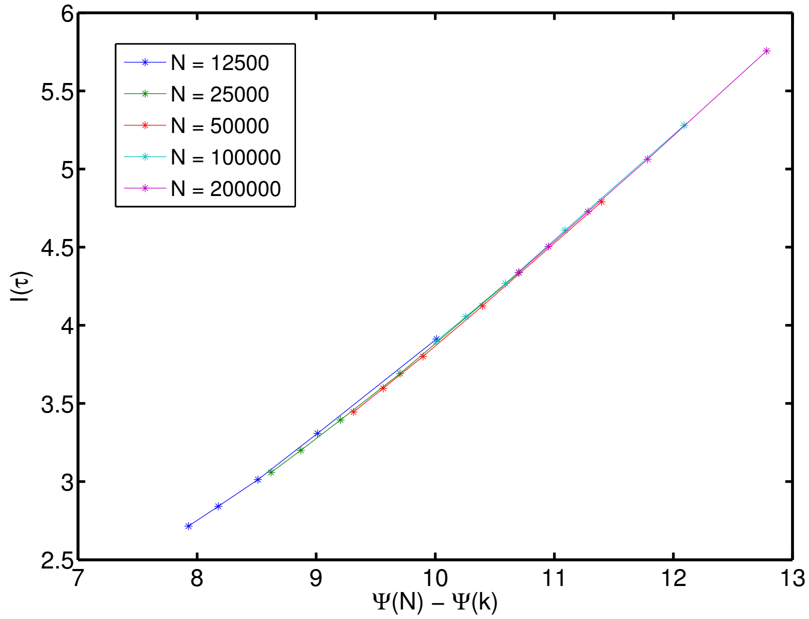


Figure 4.1: PMI found using the K-G estimator with nearest neighbour indices $1 \leq k \leq 5$ for various sample sizes N. Here $K = 0.97 \approx K_c$, $\tau = 100$.

given joint data

$$X^{0,\tau} \sim \rho_{0,\tau}, \tag{4.4}$$

$X^{0,\tau} = X^{0,\tau}(N, \tau)$, and an unbiased estimator $\hat{H}$,

$$H[\rho_{0,\tau}] = \lim_{N\to\infty} \hat{H}(X^{0,\tau}), \tag{4.5}$$

and hence

$$I(\tau) = -\lim_{N\to\infty} \hat{H}(X^{0,\tau}). \tag{4.6}$$

Thus PMI is computed by estimating the entropy from a set of ordered pairs of points and their corresponding τ^{th} mappings.

We estimate PMI at $K \approx K_c$ for several sample sizes N and some τ. Figure 4.1 shows the result computed for the first five nearest neighbour indices k. As resolution is increased PMI does not converge to some asymptotic value but instead increases indefinitely. The framework for situations where this occurs was given in the previous section, where it was found that PMI can be expressed as:

$$I(\tau) = I_0 + \Gamma \log\left(\frac{N}{k}\right). \tag{4.7}$$

The resolution variable in the K-G estimator is expressed as $\Psi(N) - \Psi(k)$, where Ψ is the digamma function - the logarithmic derivative of $(n-1)!$ for some argument n. Here for convenience we adopt the same notation, bringing out the intrinsic dependency of nearest neighbour statistics on resolution, and avoiding errors due to conflicting representations. The plots in figure 4.1 are in line with each other, confirming that $I(N,k) \approx I(\Psi(N)-\Psi(k))$, and hence that eq. (4.7) can be written as $I(N,k) \approx I_0 - \Gamma(\Psi(N) - \Psi(k))$ (the small variation at the lower end of the resolution range is due to small size fluctuations, which makes different realisations with same N deviate further than the minor deviation seen for the different N plots).

The information codimension Γ is the slope of PMI with resolution,

$$\Gamma(N,k) = \frac{(D_- + D_+ - D_{-+})}{D_{-+}}. \tag{4.8}$$

The marginal and joint information dimensions $D_{m/J}$ are defined in eq. (3.16) by assuming a linear relation between the partition-induced (discrete) Shannon entropy and the effective partition cell size. Here the marginal information dimensions will always be assumed to be equal to their box-counting dimensions (in the next section we check that it is reasonable to take the marginal entropies to be zero), and so

$$\Gamma(N,k) = \frac{(4 - D_{-+})}{D_{-+}}. \tag{4.9}$$

Thus for information dimensions D to be defined the PMI has to scale linearly with resolution in the limit of high resolution. If PMI is nonlinear, then D are not defined, and hence neither is Γ. However, we will see that the high resolution limit of PMI scaling is, though indeed linear, entirely non-interesting (corresponding to the fully-causal system), and that all the curious changes occur at finite resolutions. Therefore in this work we compute Γ as simply the linear gradient, whilst being careful to distinguish cases when this assumption is valid to cases when it is not, and actually the PMI scaling is nonlinear. Just as stated before, we take Γ and the information dimensions to be indicative of the *manner* in which PMI increases, and thus we can talk about these quantities for a range of

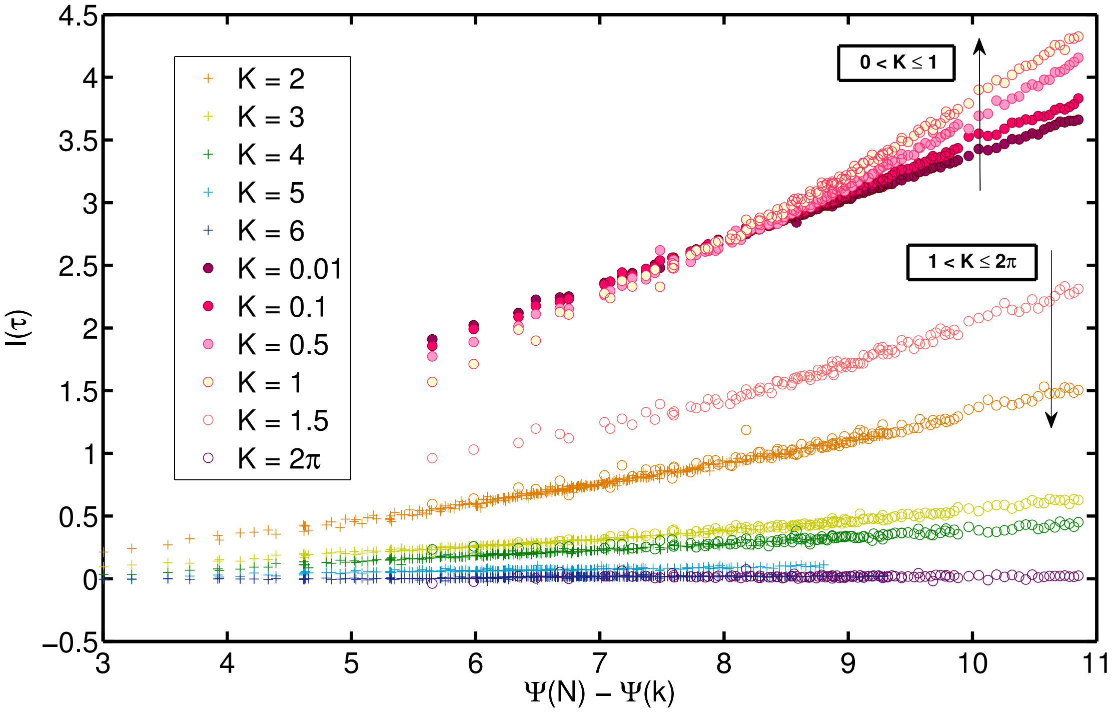


Figure 4.2: Persistent Mutual Information in the standard map at $\tau = 100$ for different nonlinearity parameters K ($K = 0$ corresponds to the fully-integrable case, $K = 2\pi$ to the fully-chaotic case at resolvable scales, and the critical value $K_c \approx 0.97$). The legend is arranged first by '+' (sample size $1000 \leq N \leq 49000$, nearest neighbour count $1 \leq k < 5$), then 'o' (sample size $1000 \leq N \leq 29000$, $5 \leq k \leq 50$). The yellow and green circles are the continuation of the respective '+' lines with $K = 3$ and $K = 4$.

resolutions.

Figure 4.2 shows Persistent Mutual Information in the standard map at $\tau = 100$. A range of N and k values are used, and all coincide to confirm the scaling of PMI with resolution. For this τ the rates with which PMI increases with resolution vary with K. When $K = 2\pi$ the map is, as far as we can see, fully chaotic, and PMI is zero. This is in line with what was observed for the logistic map at $r = 4$, and it implies that at $\tau = 100$ even for largest resolution on the graph the correlations have all decayed to zero.

There does not seem to be a value of $K < 2\pi$ for which PMI converges with resolution. This means that the given scales do not contain any globally stable behavioural trend that would lead to some aspects of trajectories persisting over time. Instead PMI appears to increase indefinitely, with a rate that can be approximated by the logarithm of the probability resolution (corresponding to the effective number of boxes with which the estimator views the joint probability distribution).

The interesting manner in which PMI varies with resolution *for different* K is better seen through considering Γ directly as a function of K. Increasing resolution is equivalent to specifying the past and future positions in less uncertain terms. Better knowledge of the past can only improve the guesses made about the future, that is to say that PMI cannot decrease with resolution. For the chaotic case a $\Gamma(K)$ of zero rightly means that however much one improves the level of resolution with which the initial position is specified, after some finite time it would still not make any difference for the purposes of prediction. By implication, for any K and finite τ there exists a resolution beyond which $\Gamma(K)$ is greater than zero. This would be true for any deterministic dynamical system.

A higher $\Gamma(K)$ means that the system better converts the same gain in the knowledge of initial conditions to information about the future, in other words that it retains predictive information better. Coming back to the standard map, as K increases the KAM tori begin to break down. The size of the chaotic region increases as the number of quasi-periodic trajectories goes down. If one naively associates the ‘amount’ of chaos with the extent of unpredictability we would expect $\Gamma(K)$ to decrease with higher K.

That is indeed what we see when K increases beyond some $K^* \leq 1.5$, when $\Gamma(K)$ falls down to zero. The surprising feature is that as K increases *up to* K^*, $\Gamma(K)$ rises as well. This is equivalent to saying that knowledge about the system evolution obtained from a certain sample size would be greater the more nonlinear a system is. We interpret this statement by recalling the caveat that system evolution refers to the state at a specific time

τ in the future. It is also known that the increase in the level of chaoticity at subcritical K is accompanied by abnormally slow relaxation times. The increase in the number of chaotic trajectories that comes about by breaking up the KAM tori is thus offset by a general mode of stickiness that stops the - lethal to the memory of initial state - exponential divergence. In reality chaotic trajectories spend a long time barely moving apart (being stuck), then diverging with some Lyapunov exponent, then being stuck again.

This apparent peak in Γ that occurs at some $K^* < K_c$ is the first of the two main features that will be the focus of this chapter. The second concerns the linearity of PMI plots itself, i.e. the extent to which our linear approximations capture the more 'in-depth' behaviour of mutual information. In the resolution range of figure 4.2 the slopes of PMI only appear linear for both very small *and* very large values of K. Around K_c - this is better seen in figure 4.1 - PMI is convex. This could of course be suggestive of the existence of more than one linear regime. Moreover, all these features could, and do, vary with τ. PMI can thus be investigated through Γ, which becomes a function of K, N, and τ.

Before proceeding we investigate the errors implicit in our assumptions and methodology.

Methodology and Errors

There are several levels at which errors could come into this procedure, but these will not necessarily be carried through or cause large deviations from the true answers. The first concerns the validity of eq. (4.4), i.e. being certain, to within some error, that the numerically obtained data samples $\rho_{0,\tau}$. This implies an assurance that it is indeed the standard map that is being investigated, and not some other evolution rule (although not strictly true, the implication of eq. (4.4) not holding is that the averages computed with respect to the actual distribution will be different than the averages computed with respect to $\rho_{0,\tau}$). A side product of this failure could be the breakdown of eq. (4.3), though that is not strictly necessary, since two different distributions may have equal marginals. If eq. (4.3) does not hold, eq. (4.4) does not either, but again that does imply that some averages are not equal (hence eq. (4.6) might still stand). The final point concerns the behaviour of the estimator, i.e. eq. (4.5). This includes fractal cases when the Shannon entropy scales as logarithm of the resolution, and broadly speaking concerns predictability of estimator behaviour for the range of distributions considered.

Section 4.1.1 focuses on the validity of statement (4.4). It attempts to clearly identify the assumptions we will take for granted. Subsection 4.1.1 investigates some alternatives, at the same time confirming reasonability of eq. (4.3).

4.1.1 Sampling the Joint Distribution of the Standard Map

The procedure for computing PMI detailed above contains a step that requires a dataset sampled from the joint distribution of the standard map (with the implied dependency on a flat prior and the τ^{th} iterate), $\rho_{0,\tau}$. Yet other steps indicate that the correspondence between $\rho_{0,\tau}$ and the effective distribution being sampled, $\hat{\rho}_{0,\tau}$, need only go so far as to produce comparable entropies (it is unlikely that different datasets with otherwise similar joint and marginal entropies would actually give different entropies if the latter are estimated using a numerical procedure; but would depend on the specifics of the estimator). The marginal entropies of $\rho_{0,\tau}$ are known analytically to be zero, in fact a requirement in eq. (4.3). It is hence possible to check whether the same is true for marginal entropies of $\hat{\rho}_{0,\tau}$. Here, in the event of a successful outcome, the straightforward method of obtaining assurance stops, and in order to understand the extent to which $\hat{\rho}_{0,\tau}$ could be different to $\rho_{0,\tau}$ we must examine in depth the process that generates the dataset.

Using the same notation as for the logistic map, let $X^0 = \{X_i^0 : \ i = 1 \mathinner{..} N\}$ be a set of initial configurations of the standard map F, $X_i^0 \sim \rho_0 \ \forall i$. If the 'future' dataset consists of the iterated points

$$X^\tau(N) = \{X_i^\tau : \ X_i^\tau = F^\tau X_i^0 \ \forall X_i^0 \in X^0\}, \tag{4.10}$$

and the 'joint' of a set of ordered pairs

$$X^{0,\tau}(N) = \{\big(X_i^0, X_i^\tau\big) : X_i^\tau = F^\tau X_i^0 \ \forall X_i^0 \in X^0\}, \tag{4.11}$$

then $X_i^{0,\tau} \sim \rho_{0,\tau}(N, \tau)$.

The two main reasons why the obtained joint data could fail to be distributed according to $\rho_{0,\tau}$ involve first, the set of obtainable initial points, and second, the numerical representation of the process that makes up F. In other words, that the 'wrong' points may be chosen to start with, and then evolved under a mapping slightly different to the original. These two notions, especially the first, are not that problematic if the purpose is to understand how PMI behaves given some data that is at least partially understood, since after all X^0 will be drawn from ρ_0, just as required. These issues only begin to be important if we then wish to relate the observed PMI to features of the map derived *analytically.*

Typicality of Numerical Trajectories

We examine the first of these. Our aim is to use PMI in order to quantify aspects of map behaviour as realised through typical trajectories (the distribution of their starting points being ρ_0, the initial distribution). Yet it is not obvious that the set of actual initial points X^0 is in any way representative of the 'true' trajectories initialised in $X = [0,1)^2$, the map domain. Assumptions have to be made first. Here we review sources of potential differences. Let $(X_r^n)_{n\in\mathbb{Z}}$ be an expanding family of sets contained in X, where X_r^n is a set of rational numbers defined by n decimal places, and $X_r^{n'} \subset X_r^{n>n'}$. Specifics of implementation impose a limit $m \in \mathbb{Z}$ on the 'precision' of starting conditions s.t. the set of initial points X^0 becomes wholly contained within X_r^m,

$$X^0 \subset X_r^m. \tag{4.12}$$

m depends on the choice of available architecture.

There will also be an additional limitation arising out of the particular sampling method used: the set of available starting conditions will be determined by the random number generator:

$$X^0 \subset X_{RNG}^m. \tag{4.13}$$

X_{RNG}^m will vary depending on the seed, the size of X^0, and possibly other parameters. The equality can be exact if $|X^0|$ exceeds the RNG periodicity.

Hence there is a series of nested sets,

$$X^0 \subset X_{RNG}^m \subset X_r^m \subset X, \tag{4.14}$$

where X^0 is some set of realisable initial conditions, and X is the map domain.

The second inequality is unavoidable but perhaps not drastic since the set on the LHS can be changed by changing the RNG used. No single RNG will produce an equality, but a combination is likely to explore a substantial range of X_r^m.

It is the last inequality that is problematic. Forgetting for the time being that the discrete

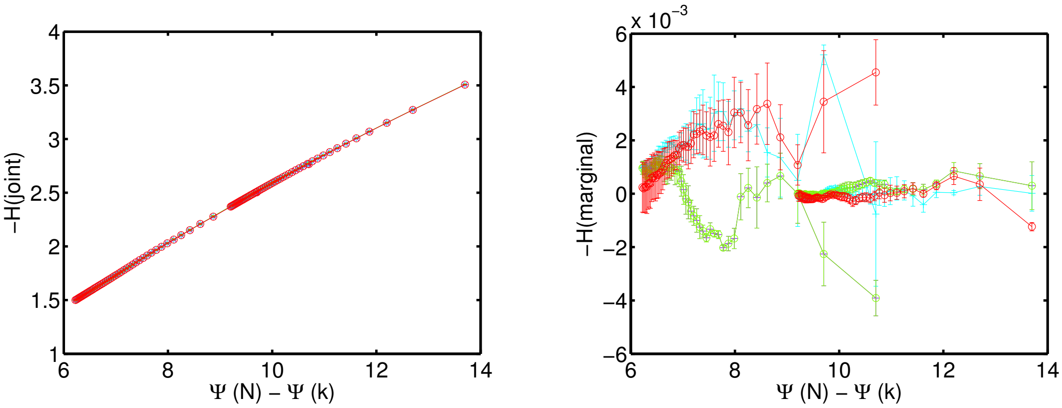
X_r^m does not *admit* a probability density, estimated entropies computed from points sampled uniformly in X_r^m and X are likely to be the same, since the distribution of rational numbers with n decimal places is unlikely to constrain such relatively small sample sizes (N will not go higher than order of millions, whilst n will at all times exceed 6 and will, in fact, be closer to the double precision machine epsilon of 10^{-16}). Yet we do not know whether the selected orbits are typical of the standard map behaviour. Since this representability is the reason behind the flat ρ_0 requirement, what we would in fact end up analysing (had we a perfect numerical representation of F) is a map defined by its typicality as given by X_r^m. This in a way is inescapable and forms a common tacit understanding when using dynamical systems data. In our work we take it as given and proceed to associate these trajectories with ones typical of the standard map. As we shall see this limitation still preserves at least some properties of the standard map, such as area conservation and co-existence of various trajectory types, so the typicality argument does not, at least on first glance, fail the reasonability test.

Precision in Implementation of the Standard Map

The second reason why the data may fail to be sampled from $\rho_{0,\tau}$ is a consequence of the inevitable errors associated with approximating $F : X \to X$ by a map on a set of rationals, $\hat{F} : X_r^m \to X_r^m$. These will be compounded through a large number of iterates τ, and as a result, especially in chaotic regimes, the final iterate may be significantly different to its analytical counterpart. In this section we examine the source of these errors as well as conjecture that possible shadowing properties might still save statistical averages.

We wish to approximate the standard map with double-precision operations. Let $\hat{F} = F_p$ be the standard map machine affected using floating-point arithmetic of precision p. Because of the nature of the map, whose regular regions are interwoven with chaotic ones, we cannot say that out of two values of p the approximation F_p with the higher one will map the point more closely to the true iterate. Such a statement would also fail because of the periodic boundaries that might wrap a large enough error around. Yet it is instructive to see whether the precision p makes a difference to the required averages, and if so, at what number of iterations.

Consider first the entropy of the marginals, which for F should be zero. We examine the marginal entropies for a range of sample sizes N and estimator parameter k that controls



(a) Entropy of the Joint Distribution. Four combinations of double/long double iteration and estimation.

(b) Entropies of the Marginal Distributions. In *green*, superimposed upon *purple*, are entropies of the initial distribution in both double and long double representations. Future entropies are shown in *sky blue* (double) and *red (*long double).

Figure 4.3: Effect of precision on the entropies for $K = 1$, $\tau = 10^4$ and two sample sizes $N = 25000$ and $N = 500000$ as a function of probability resolution where $1 \leq k \leq 50$ is an estimator parameter that defines the neighbourhood in terms of number of nearest neighbours. Results are averages over 3 runs (in 4.3(a) error bars are too small to be visible). Same initial dataset was used for both precisions, correspondingly giving superimposed initial entropies. Estimating schemes used same precision as iterating schemes.

depth of sampling. This is done for $K = 1$, the value around which the Golden KAM torus breaks down. The τ value considered - 10^4 - is located roughly in the middle of the reasonable computational range.

Effect of precision on iteration and estimation was tested by creating two datasets, D_d and D_{ld}, of varying precision. These contain N pairs of initial and final (τ-iterated) points. The set of initial points was chosen using a double random number generator (Mersenne Twister), and depending on the precision of iteration the points were then cast as long doubles and iterated with all the variables recast accordingly. For a given standard map parameter K, sample size N and a τ, three pairs of each dataset are produced. Entropy of each dataset is then estimated twice, using methods of different accuracies.

Figure 4.3(a) shows that precision of the estimating procedure does not play a role (at least for this sample size range). The lack of change in neighbour statistics implies that few points are that close to each other, i.e. some area is still being preserved (at least in places). This is of course backed up by the fact that the 'future' entropies are very close to zero.

We take it as a premise (which will be shown to be true later on) that implementing the map with a higher precision scheme gives a better indication of the true trajectory (at least before the periodic boundaries begin to wrap the error around). This suggests existence of a set of parameters for which F_{ld} should produce a distribution that is closer to $\rho_{0,\tau}$ than the equivalent one produced by F_d. Comparing mean values of the two resultant distributions would give an indication whether loss of information about *particular* trajectories necessarily leads to a change in some global statistical properties. A negation of this statement constitutes additional support for the assumption on which all our subsequent analysis relies on, should we wish to relate our results to the standard map - that shadowing allows $\hat{\rho}_{0,\tau}$ to retain significant information about the underlying dynamics.

This supposition is supported by fig.4.3, from which we can conclude that there exists a regime and a time gap for which, as far as the given means are concerned, two standard map implementation schemes that differ on precision sample some identical variant $\tilde{\rho_{0,\tau}}$ of the joint distribution, and that as this distribution also happens to conserve the area-preserving feature of the map, we might presume on it to do the same with other features of interest. The fact that certain averages taken with respect to distributions modelled by maps implemented with different precisions agree with each other points to some map property that allows for the existence of a family of maps that give same averages.

Numerical route to Exact Solutions

We postulate that there exists a procedure to compute the true iterate of the standard map given some starting conditions that is well defined in terms of the machine being used. This process is computationally intensive and would only work for orbits whose complexity (character combined with its length) does not in some way exceed the available machine memory. We will show that it exists and then use it to assess the accuracy of floating-point iteration schemes (since so far it has not been shown how the roughly three-digit gain that long-double type gives reflects in the final outcomes).

The standard map F consists of a sequence of operations $(O)_j$ on a point in the state space $X = [0,1)^2$. These can be performed to any desired accuracy using arbitrary precision arithmetic. Let $AP^{l,m}$ be an arbitrary precision operator that performs these operations, rounding each outcome to l decimal places, then rounding the final answer to $m \leq l$ decimal places. $F_{AP}^{l,m}$ then corresponds to the arbitrary precision version of the standard map.

The motivation behind rounding is to ensure that results are at all stages reproducible independent of the machine architecture.

Let $Y^{l,m,\tau} = F_{AP}^{l,m,\tau} Y^0$. We conjecture (true if F is continuous) that for all initial condition $X^0 = Y^0$, for any number of iterations τ, the 'true' iterate of the standard map F, $X^\tau = F^\tau X^0$, is given by

$$X^\tau = \lim_{l,m\to\infty} Y^{l,m,\tau}, \tag{4.15}$$

and, moreover, that given some ϵ we can find an m^* such that for any $l = m > m^*$

$$d\left(X^\tau, Y^{l,m,\tau}\right) \leq \epsilon \tag{4.16}$$

for all X^0, where d is any (true if F is again, continuous.) metric on X. We take the diagonal increase of l at the same time as m as the optimal way of taking the limit, given unlimited resources but a cap on l and m. In practice we will only consider $m = l$, so the arbitrary precision version of F can be written as F_{AP}^l.

Some justification for the above conjectures can be found in differences $d(l, X^0)$ between iterates F_{AP}^l and F_{AP}^{l+1} for some initial condition X^0. Figure 4.4 shows the logarithm of d plotted against l, the number of rounding digits after each operation.

From figure 4.4,

$$\log_{10} d \approx \min\left(\mathbf{O}(10^{-1}), -l + c(\tau, K)\right), \tag{4.17}$$

since the slopes appear constant and equal to unity. So at least for the trajectories (initial conditions) that behave in this manner the distance d between the $F_{AP}^{l,\tau}$ iterate and the true solution of the F^τ iterate is, by triangular inequality,

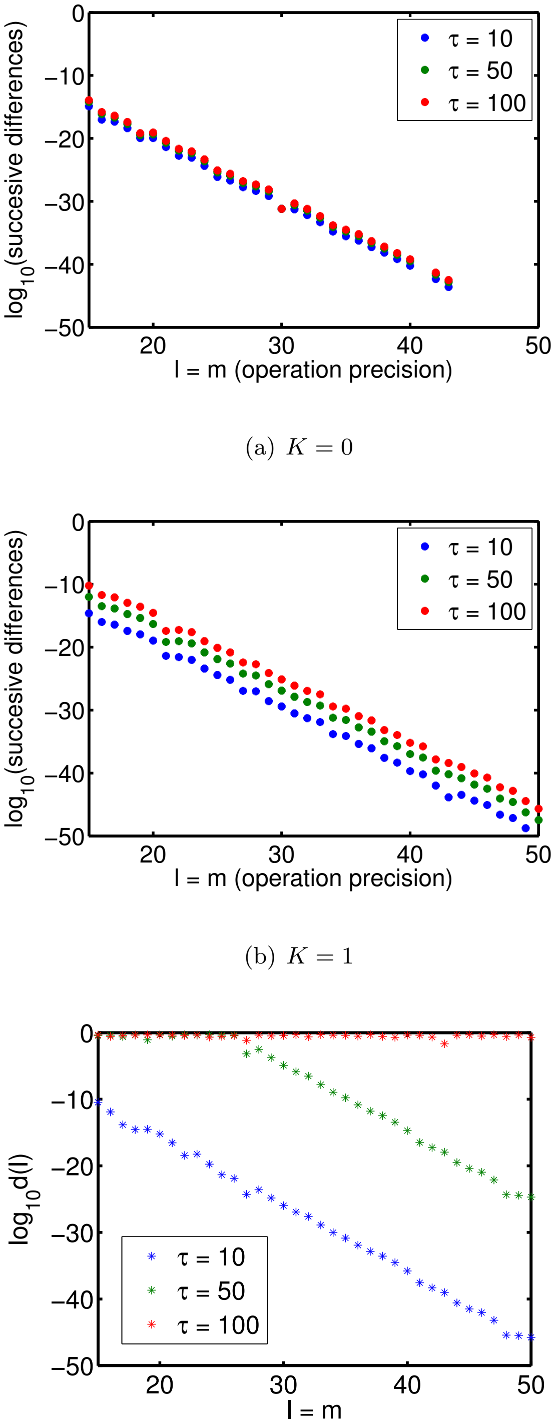
$$d \leq \Sigma_l^\infty d(l) \leq \Sigma_l^\infty 10^{-l} 10^{f(\tau,K)}.$$

Here f is positive and incorporates both the min and c. Hence

$$d \leq 10^{f(\tau,K)} 10^{-l} \ln 10. \tag{4.18}$$

Equation (4.18) suggests that for any positive f a precision l can be found so that the result of the iterative map $F_{AP}^{l,\tau}$ will be within d of some limiting point, which we identify with the true solution.

Before proceeding further it is worth noting that the graphs shown in figure 4.4 already



(a) $K = 0$

(b) $K = 1$

(c) $K = 2\pi$

Figure 4.4: Differences using the maximum metric between outcomes of arbitrary precision methods where operation rounding differs by one digit (lower value on the abscissa). Shown for standard map K corresponding to increasing chaoticity. Each graph corresponds to one initial condition.

provide a hint about orbit complexity. The difference in the final error between iterates with the same l but different τ can be interpreted as the separation of trajectories originally of order 10^{-l} apart. This turns eq. (4.17) into the usual statement

$$\Delta x(\tau) \approx \Delta x(0) + c(\tau, K), \tag{4.19}$$

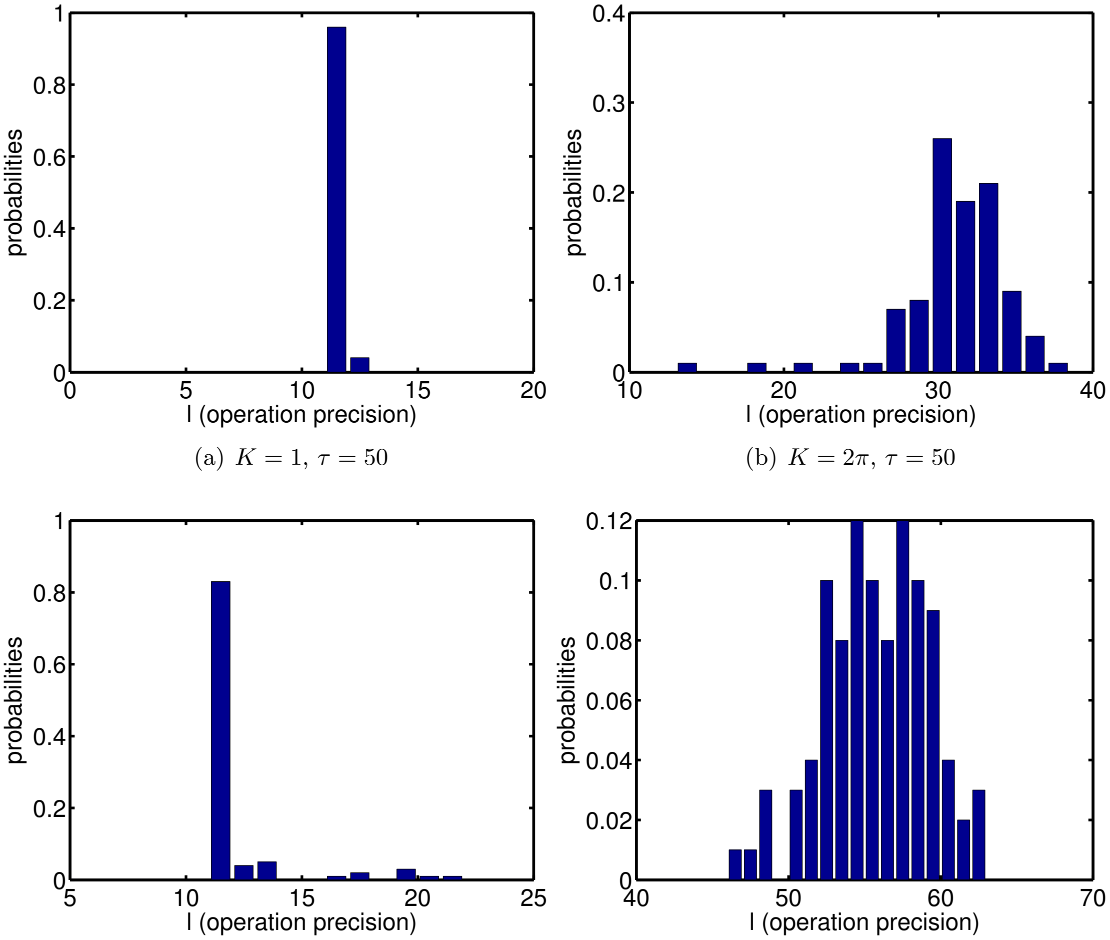
where logarithm of the final separation $\Delta x(\tau)$ is just $\log_{10} d$ and logarithm of $\Delta x(0)$ is l. There is therefore a sense in which $c(\tau, K)$ relates to the speed of trajectory divergence. Qualitatively this statement is indeed supported by the graphs: consider $K = 0$. There the change of c with τ is almost negligible. The trajectory the plots refer to (random initial condition) is not chaotic, and the fact that plots do not collapse shows a potential difference between the Lyapunov exponent and c - the latter takes into account all cumulative algebraic errors, and the former is a statement about the behaviour of the map itself. With higher K the plots start to separate. The intercept, which we identify with c, thus changes with τ, and as a further exercise it would be interesting to see exact rate of change. From eq. (4.19), a linear dependency would mean c is proportional to the Lyapunov the exponent. The main variables to look out for here are hence both the qualitative and quantitative manner of the dependence of the intercept c on τ.

The trends shown in figure 4.4 provide us with an algorithm that for some initial condition and number of iterates outputs the solution that is within a desired error d. The underlying procedure increases l until the remaining cumulative error is less than d. This l is thus dependent on trajectory complexity.

We now use this setup in order to gauge the extent to which the usual floating point arithmetic fails to reproduce trajectories ostensibly associated with the given initial point - and the consequent hope that it actually entails some other, unseen trajectory, thus retaining some fundamental character of the map.

Consider the resolution range for some typical $N = 10^5$, $k = 1$. Under these parameters the logarithm of the average distances between nearest neighbours at some $t = \tau$ is $\log_{10}(1/2) - (5/2)$. So if the distance between a solution and its true solution is $d = 10^{-5}$, then relatively speaking the error is larger than the average interneighbour distance between two points in the 'future'. We ask the question of roughly how large does the arbitrary precision accuracy l need to be in order to have the solution be closer to the true answer than

d. Since the character of the trajectories may differ this can be made general by averaging



(a) $K = 1$, $\tau = 50$

(b) $K = 2\pi$, $\tau = 50$

(c) $K = 1$, $\tau = 100$

(d) $K = 2\pi$, $\tau = 100$

Figure 4.5: Frequency count of precision l at which distance to true solution is less than 10^{-5} for $N = 100$ initial conditions.

over the trajectories. Figure 4.5 shows a sample spread $h(l)$ of precision values required to bring $N = 100$ solutions to within 10^{-5} of their true values.

The range of τ in these examples is limited by the computational effort that increases with l. That in itself is indicative of how fast trajectories deviate from their true values - that largest τ used is 100, far to the lower end of the typical range. At this τ computation of true iterates has to be done with roughly 50 decimal places after each operation implicit in the mapping. In some ways this is the worst case scenario, since here $K = 2\pi$, and there are no regular trajectories. It becomes clear by implication that trajectories estimated using floating point precision with its mere 16 d.p. will after τ that is of order 10 begin to deviate from their true values.

$K = 1$ presents a slightly more optimistic picture. The mean value of l does not change much with τ, or rather it does but in a very slow manner. We will later see that $K \approx K_c$ is characterised by very slow relaxation times, which is what is responsible for the apparent stationarity with l. However, the mean is likely to shift to the right for any reasonable value of τ. Also as we move away from K_c we expect, a posteriori, l to move faster, in both $K < K_c$ and $K > K_c$ directions.

From these results it seems that the only possible gain from switching to long double precision would be offset very quickly by τ. These results point to the conclusion that, assuming floating-point arithmetic does preserve some features of the original map, then the only reason why it would do so is if the deviations from 'true' orbits are somehow *systematic*. After all PMI is only interested in the relative distances and not the absolute values. Preservation of at least such features as the entropies of the marginals leads us to suspect either shadowing, or systematic errors, or simply that the computational standard map F_d is in some ways similar to the original. We thus accept the latter and assume existence of a correspondence with the analytic standard map.

We established the main assumptions behind numerical computations of PMI in the standard map. Throughout this work we will associate the range of behaviour evident in the numerical trajectories with some 'true' system behaviour. This is done in spite of both the finite range of the computationally available initial conditions, and the errors accumulated from finite-precision arithmetic. From working with arbitrary-precision algorithms we see that the accumulated errors of floating-point arithmetic accumulate so fast that the exact precision of the variables makes no realistic differences; and hence that the only reason why some functions of the distributions are conserved has to do with the resultant map somehow having the same characteristics. Therefore we proceed using the double precision, for which at least eq. (4.3) is true. We also use the K-G estimator and assume it is well-behaved so that eq. (4.5) holds, and leave any discussion about that to the concluding sections.

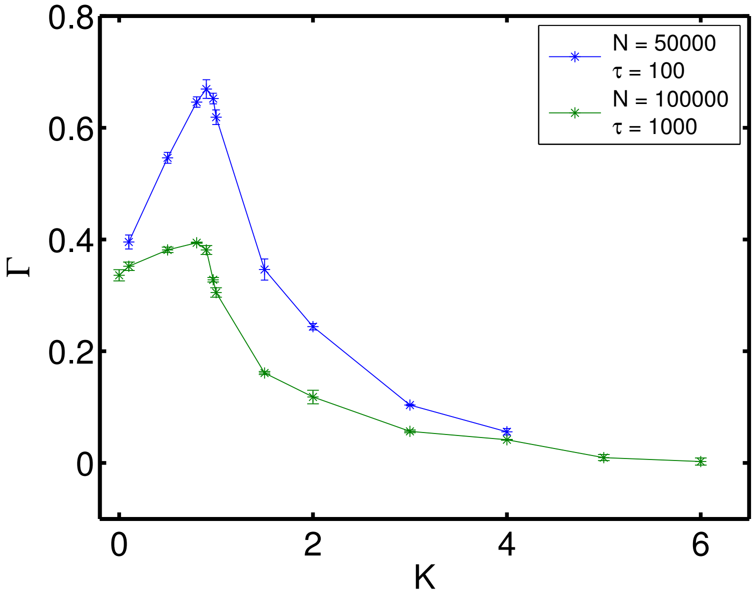


Figure 4.6: Information Codimension Γ in the standard map vs nonlinearity parameter K. Computed from nearest neighbour count $1 \leq k \leq 5$ and averaged over three runs.

4.2 Features of Γ

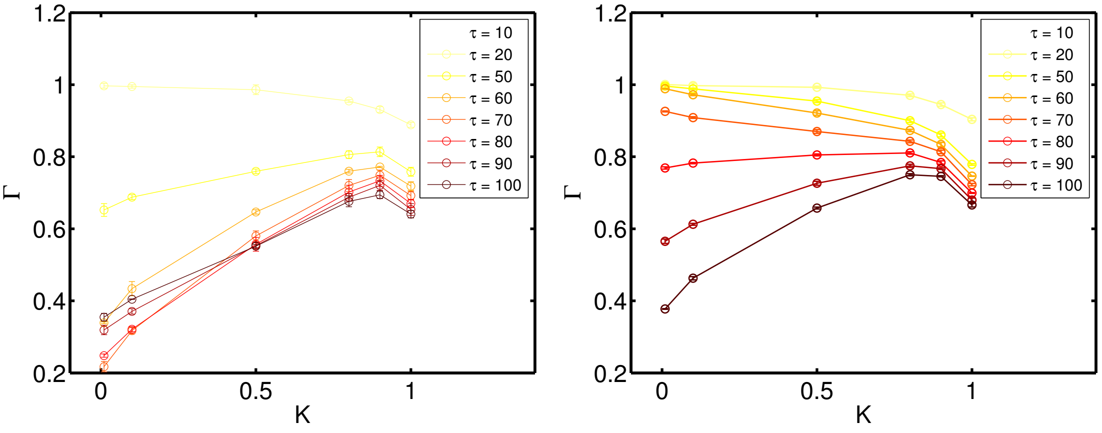
We are now in a position to investigate Γ with some degree of certainty in our computations. We approximate Γ with a linear slope of PMI with $\Psi(N) - \Psi(k)$. The latter can be varied either by increasing k or decreasing N. Although here the effect would be the same, it need not be so (depending on the metric), and we operationally define $\Gamma = \Gamma_k$ as the gradient of PMI with $\Psi(N) - \Psi(k)$ where k is allowed to vary in some fixed range.

A variable slope does not invalidate eq. (4.7) - the scaling of PMI with resolution will be seen later to be broken only by the particular behaviour of the metric, and for clear reasons (figure 4.1 that was used to demonstrate the scaling is actually specifically computed at parameters where Γ is non-linear and PMI transitions from the fully-causal limit to some finite value).

We now compute $\Gamma(K)$ and plot it as a function of K for some (N, τ). Results for two sets of parameter values are shown in figure 4.6.

Γ varies between 0 and 1, which corresponds to the joint information dimension lying between the value for the marginal dimension and their sum. The blue plot has similar parameters to the data shown in figure 4.2, and just as expected we see a peak at some K^*.

The second plot in fig. 4.6 shows the effect of variation of parameters. The peak still exists,



(a) $N = 100000$

(b) $N = 500000$

Figure 4.7: The variation of Γ with τ for the lower range of standard map parameter K for $N = 100000$ (fig. 4.7(a)) $N = 500000$ (fig. 4.7(b)). Γ computed from three runs at $1 \leq k \leq 5$.

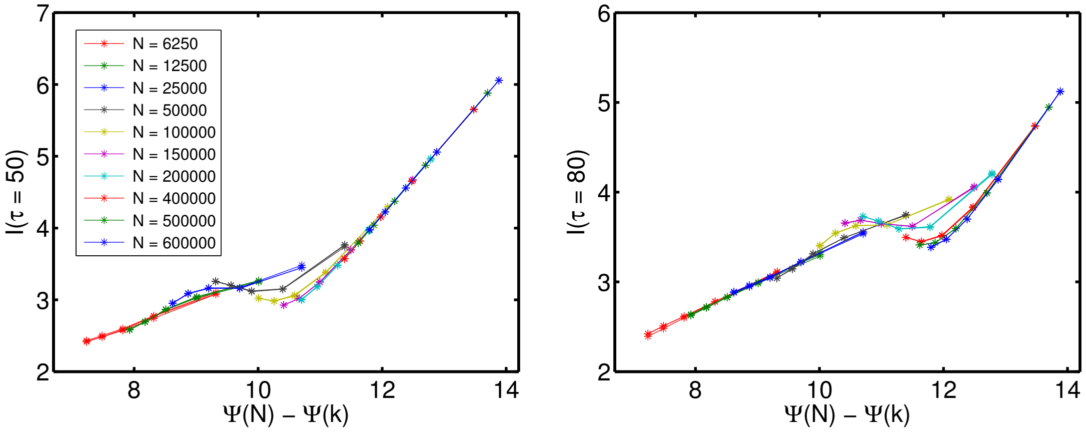
but the result is significantly different to the $\tau = 100, N = 50000$ plot. Time gap and sample size are the only changes that were made. It is reasonable to assume that since there exists a certain scaling with N that the reason for the change is only due to τ.

In order to understand how $\Gamma(\tau)$ changes with parameters we therefore focus on the peak, and investigate $K \leq 1$, which heuristically is a better bound on the possible peak than the $K = 1.5$ guess mentioned above. Figure 4.7 focuses explicitly on the $\Gamma(\tau)$ dependency. In figure 4.7(a) we observe that in the majority of cases increasing τ causes the peak to become lower.

In terms of predictability this is sensible since the higher the number of iterations the more information *from the original resolution* needs to be obtained in order to understand the future in the same way as for a low τ. We now check if this is true for a different range of resolution.

Results are shown in figure 4.7(b). Depending on N, different values of Γ are observed for the same τ. Hence for any τ there does not exist a single unique scaling of PMI with resolution, and $\Gamma(\tau) = \Gamma(\tau, N)$ (all this under the implication that we are actually measuring $\Gamma_k(\tau, N)$).

We also see that in 4.7(a), for low K, $\Gamma(\tau, N)$ actually *increases* with τ. The fact that the fully-integrable case of $K = 0$ is also prone to this behaviour suggest examining $\Gamma(K = 0)$ in order to explain this and disentangle the interdependency.



(a) $\tau = 50$

(b) $\tau = 80$

Figure 4.8: PMI as a function of probability resolution for $K = 0$ for $\tau = 50$ (4.8(a)) and $\tau = 80$ (4.8(b)).

4.2.1 Fully-Integrable Case of $K = 0$

At $K = 0$ the standard map becomes

$$p' = p \tag{4.20}$$

$$x' = x + p, \tag{4.21}$$

(4.22)

where dynamics is once again wrapped around the torus. No chaotic trajectories are present in this fully-integrable case. If viewed on a square, orbits make sideways jumps whose length is proportional to their height (giving $(0, 0)$ as the stable point). In fact all orbits stay on the invariant tori, suggesting that the dimension of the support space of the joint is 3.

Figure 4.8 shows PMI for two different τ values. Both display two distinct scaling regimes at which point Γ is defined in its proper sense (though not in the infinite resolution limit). Making a mental transition between the two τ values would show us a movie where the transition point moves to the right and the screen becomes occupied by the lower, slow plot. Since here τ values are below the $\tau = 100$ plot of fig. 4.2, so we can infer that the slope seen on the latter graph corresponds to the left-most (or lower) of the two seen on the graph above[1].

The result of combining all the information about $K = 0$ is shown in figure 4.9. For all sample sizes N, Γ decreases from 1 to roughly 1/3, dipping to some point below the large

[1] When the slopes are defined Γ_k corresponds to Γ_N

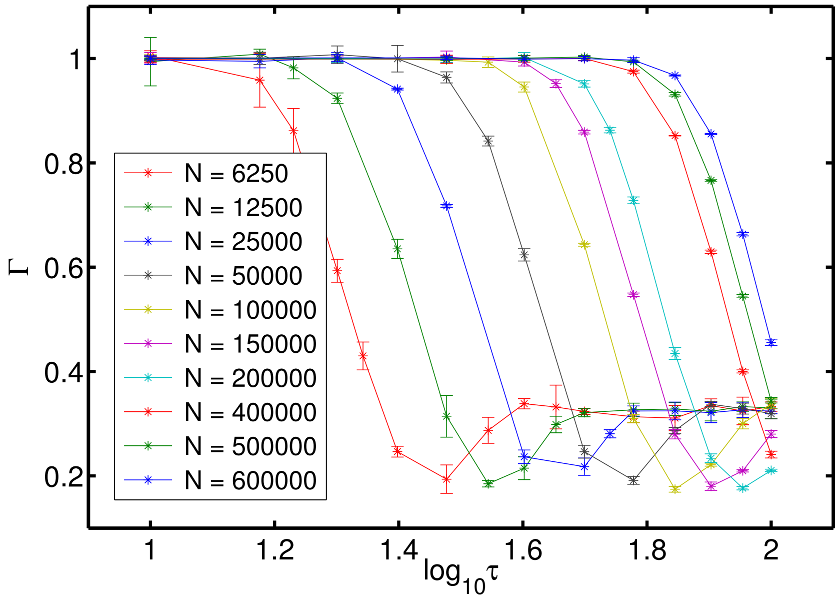


Figure 4.9: τ dependency of Γ for different sample sizes N in the fully-integrable case of $K = 0$. Logarithmic gradients of PMI computed from $1 \leq k \leq 5$ and averaged over 3 runs.

τ limit. The dip is merely a result of the wave-like transition with an overshoot between the two slopes that was observed in figure 4.8. Given that only two values of Γ actually correspond to the linear approximation of the slope, the underlying information dimensions are defined only for the two limits of $\Gamma = 1$ and $\Gamma = 1/3$.

These can be understood in terms of the joint information dimension:

$$D_{-+} = \frac{4}{\Gamma + 1}. \tag{4.23}$$

When $\Gamma = 1$ the information dimension of the joint distribution is equal to the information dimensions of the marginals, i.e. in the limit of $\tau \to 0$, $D_{-+} \to D_{-/+}$. On the other hand $\Gamma = 1/3$ corresponds to $D_{-+} = 3$, the three degrees of freedom associated to (past,future) of regular motion.

Information dimension is a result of entropy scaling with the logarithm of resolution. Lack of change between marginal and joint distributions implies that nearest neighbour statistics stay the same with time (using the framework implicit in the estimator). Points that were close have not yet moved far enough to disrupt the average interpoint distances. Hence the $\Gamma = 1$ limit is one of absolute causality - when the deterministic nature of the map fully defines the future, and uncertainty does not get blown up by iterations.

This framework allows for an explanation of the regularity with which the plots in the figure

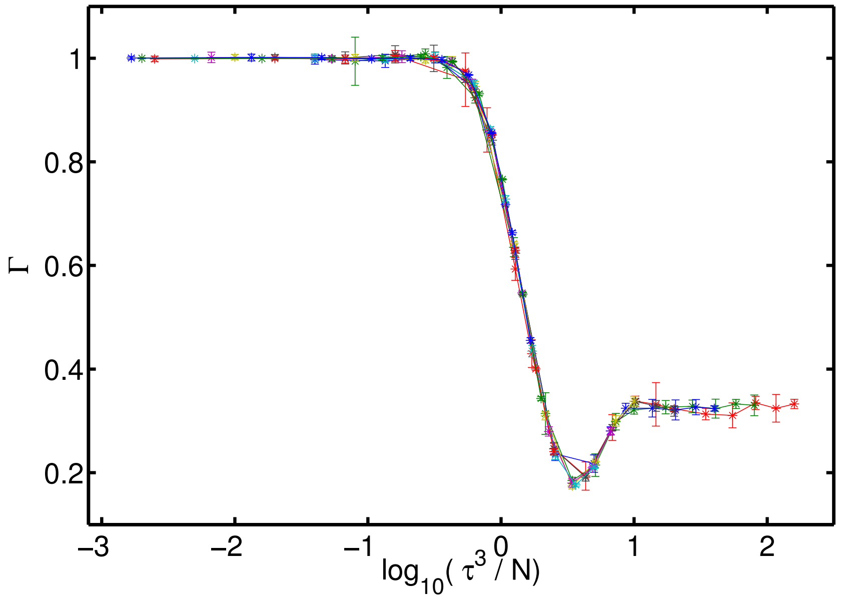
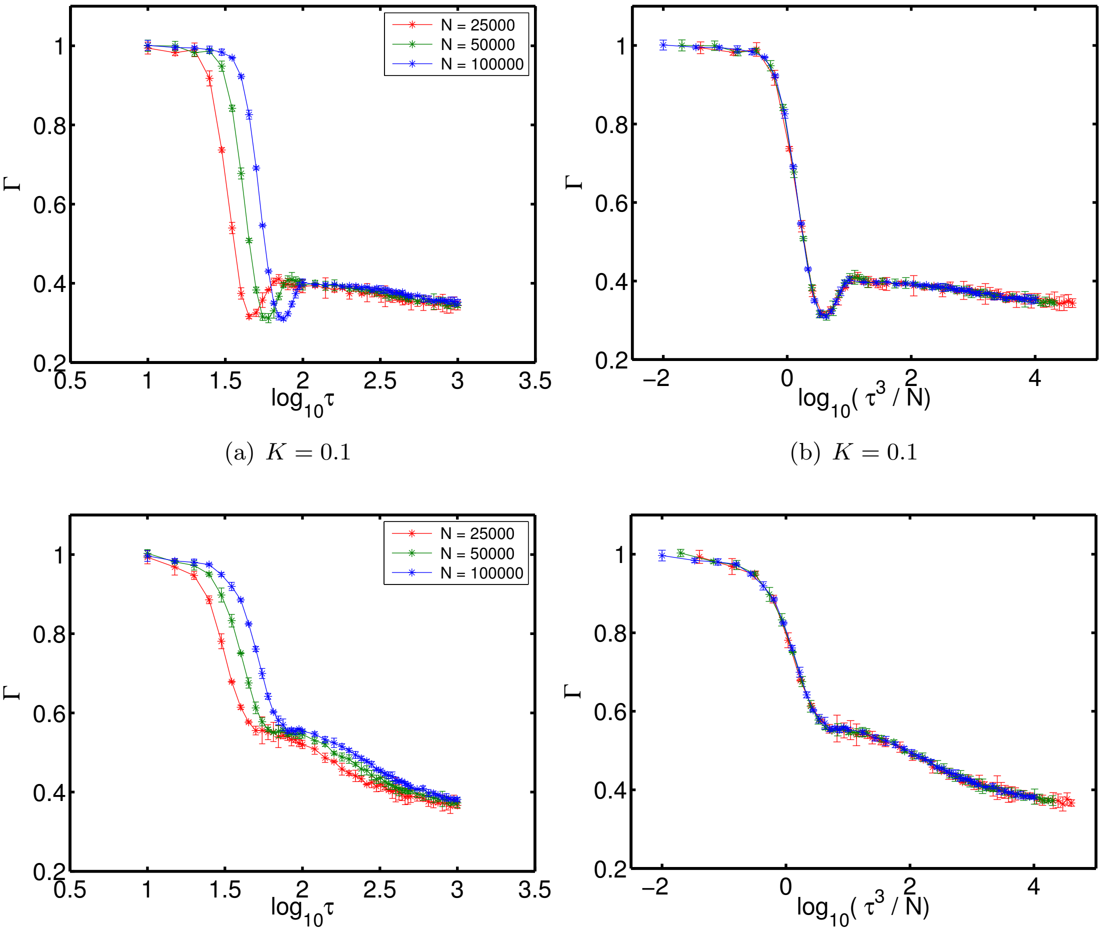


Figure 4.10: $\Gamma(f(N, \tau))$ for the fully-integrable case of $K = 0$. Logarithmic gradients of PMI computed from $1 \leq k \leq 5$ and averaged over 3 runs. The range of N is the same as in figure 4.9, and the legend stands.

translate to the right as N increases. Figure 4.10 shows the collapsed picture, indicating that at least for $K = 0$, $\Gamma(\tau, N)$ has the functional form of τ^3/N. We associate this scaling with regular motion. It can be interpreted through the interpoint statistics: let Δx_0 and Δp_0 be the initial separations in the two directions at $\tau = 0$. Then at τ, $\Delta p_\tau = \Delta p_0$, and $\Delta x_\tau \approx \Delta x_0 + \tau \Delta p_0$. The past is constrained by $\Delta x_0 \Delta p_0 \approx 1/N$, since the information dimension is equal to 2. All Δx_0, Δp_0 and $\tau \Delta p_0$ have to be less than ϵ, where ϵ is the interpoint distance of uniform mixing. Hence $\epsilon \approx (\tau/N)^{1/2}$. When the information dimension is equal to three, $\epsilon \approx N^{-1/3}$, and so $(\tau/N)^{1/2} \approx N^{-1/3}$, or $\tau^3 \approx N$.

4.2.2 Γ at intermediate values of K

We have found that in the fully regular $K = 0$ regime PMI has two distinct linear scaling regimes with resolution (the definition of resolution absorbs the logarithm). The transition between the two occupies a short, finite resolution range that can be expressed as a function of both N and τ. We do not anticipate this to be the case for other values of $K < 2\pi$. The main graph of PMI v resolution showed that for these regimes the plots were distinctly curved. In figure 4.6 Γ is seen to vary smoothly with N and τ, hinting at the lack of linear



(a) $K = 0.1$

(b) $K = 0.1$

(c) $K = 0.5$

(d) $K = 0.5$

Figure 4.11: Information codimension scaling with sample size N and τ for subcritical standard map parameters K. Graphs on the right are rescaled with τ^3/N.

PMI scaling for these parameters. In this section we investigate this in terms of Γ by varying K, N and τ.

There are three main features we wish to bring out. The first is whether, and if so then under which conditions does PMI have a clear linear scaling with resolution and hence a well-defined joint information dimension. The second is to do with the actual values of Γ, particularly at those times, but also generally across (K, N, τ). Recall that Γ indicates the extent of perceived causality. Finally the third aspect is the manner in which those values change across (K, N, τ).

We anticipate qualitatively different behaviour for subcritical K, $K \approx K_c$, and large K.

Small K Figure 4.11 shows behaviour of Γ for two values of K when $K < K_c$. By analogy with the $K = 0$ case we identify the regions of τ where the plots are coincident with a linear

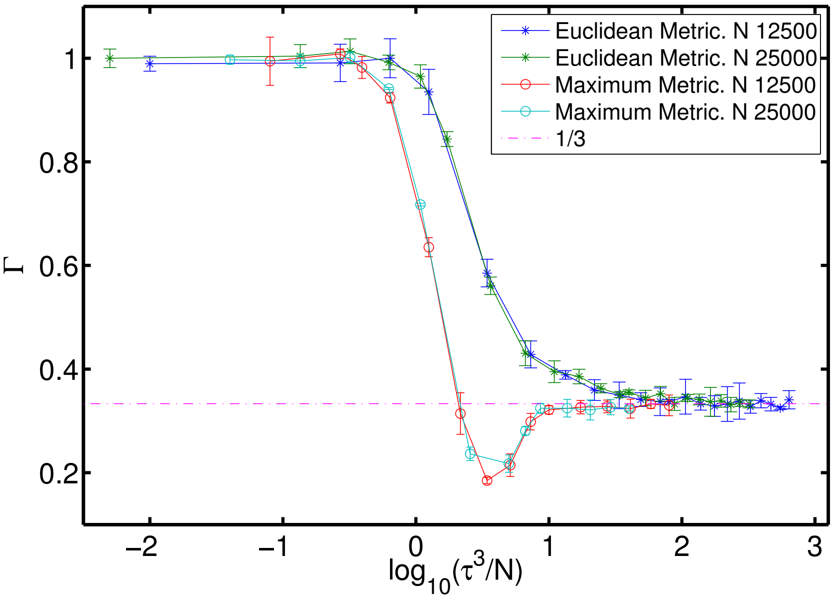


Figure 4.12: Rescaled Γ with τ^3/N for the fully-integrable case of $K = 0$. Logarithmic gradients of PMI computed from $1 \leq k \leq 5$ and averaged over 3 runs.

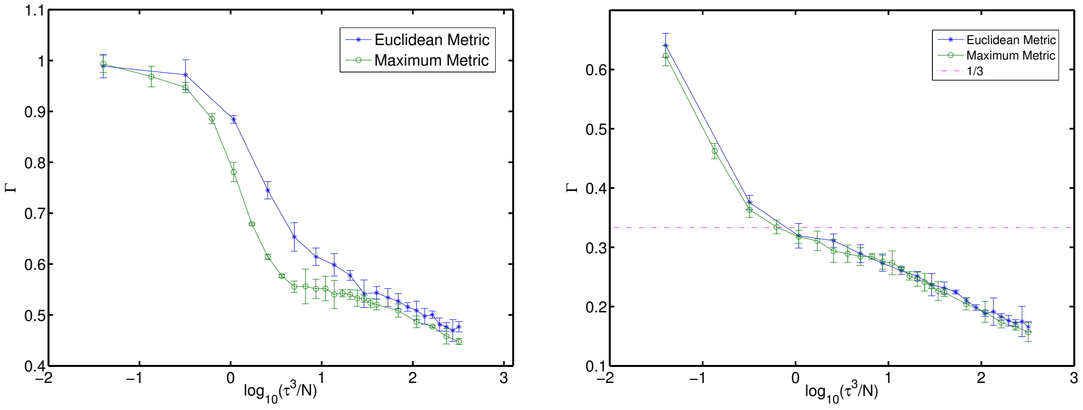
PMI scaling. This happens at small τ, and - we conjecture - in the limit of large τ, but convergence towards these values is slow, and grows slower with K.

Before proceeding to discuss the intermediate scaling we note that the relation between N and τ defined for regular trajectories above continues to hold for subcritical K, the average interpoint distance statistics undergoing a qualitative change when $\tau^3 \approx N$.

Coming back to the apparent pause in the decrease of Γ with τ, it is tempting to identify the intermediate region of τ with another well-defined linear PMI regime. However, in our attempts to explain the Γ dip present at subcritical values of K this kink was found to be a direct consequence of the metric. If Euclidean metric is used instead, all the subcritical Γ no longer looks like it consists of two distinct parts, one for lower and one for higher τ values. It also turns out to be responsible for the dip in Γ, which is deepest at $K = 0$, rising higher while at the same time becoming shallower with higher K, and disappears completely as Γ becomes roughly linear with $\log(\tau)$ at $K \approx K_c$.

Its origins can be found in the non-uniqueness of PMI at those (N, τ) ranges. $I(N, \tau)$ itself depends not only on the resolution but on the metric used to compute interpoint distances. All the graphs above are done with the maximum metric. Recomputing them for $K = 0$ case with the Euclidean metric gives what we claim to be a smooth variation (figure 4.12).

It can be confirmed by doing the same for other values of subcritical K (figure 4.13).



(a) $K = 0.5$

(b) $K = 2$

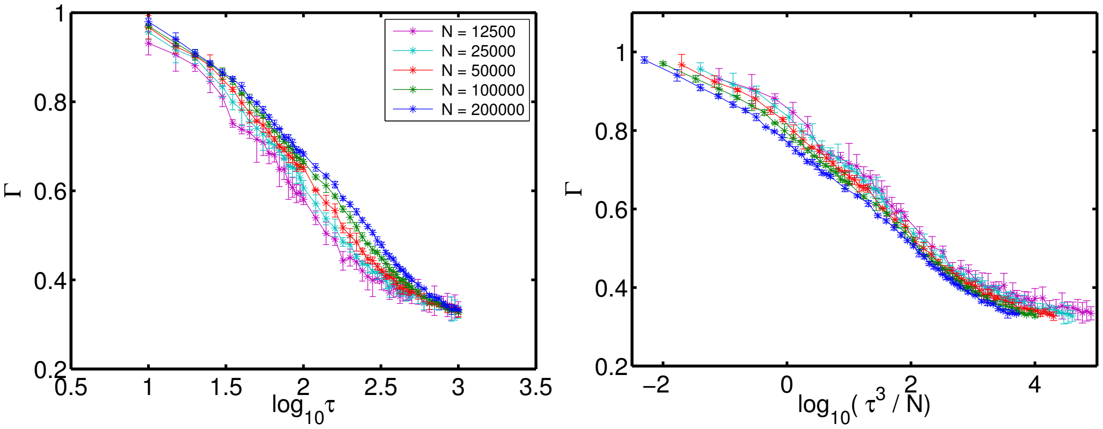
Figure 4.13: Same as figure 4.12, but varying K. Here $N = 25000$.

The dip is hence fully explained by the overshoot that results from using the relatively drastic maximum metric (plotting it for other K values confirms this). There are differences in the way Γ decreases, not just the lack of overshoot, such as the slightly slower convergence of Γ with τ. In fact we continue to use the maximum metric because it is much more computationally efficient. We can see that the metric would not change the global qualitative features of $\Gamma(N, \tau)$.

The dip is thus seen to be the effect of the 'strength' of the maximum metric. The fact that the dip smoothes out with higher K is directly related to the fact that in state space motion is no longer uniformly longitudinal. The extent of this curvature also increases with K, and the PMI computed using the different metrics converges (figure 4.13(b)).

While computationally optimal, the maximum metric can and does fail to give PMI that is uniquely defined by resolution. This is exactly what happens for subcritical K around the region where two linear PMI regimes converge. The dip is a direct outcome of *defining* Γ *through variation with* k. Looking back at figure 4.8 we see that Γ would move between the two limiting values much more abruptly had it been defined through the gradient of PMI taken w.r.t. N, keeping $k = 1$, and looking only in the required direction that changes depending on where N is in relation to τ.

K around K_c It is hard to draw any conclusions from similar graphs around K_c. As K increases to its critical value the (N, τ) rescaling becomes impossible, and indeed it is hard to find resolution ranges for which Γ is well-defined other than small τ and sufficiently large N. Figure 4.14 shows Γ for $K \approx K_c$. As N increases it approaches a straight line, the



(a) $K = 0.97$

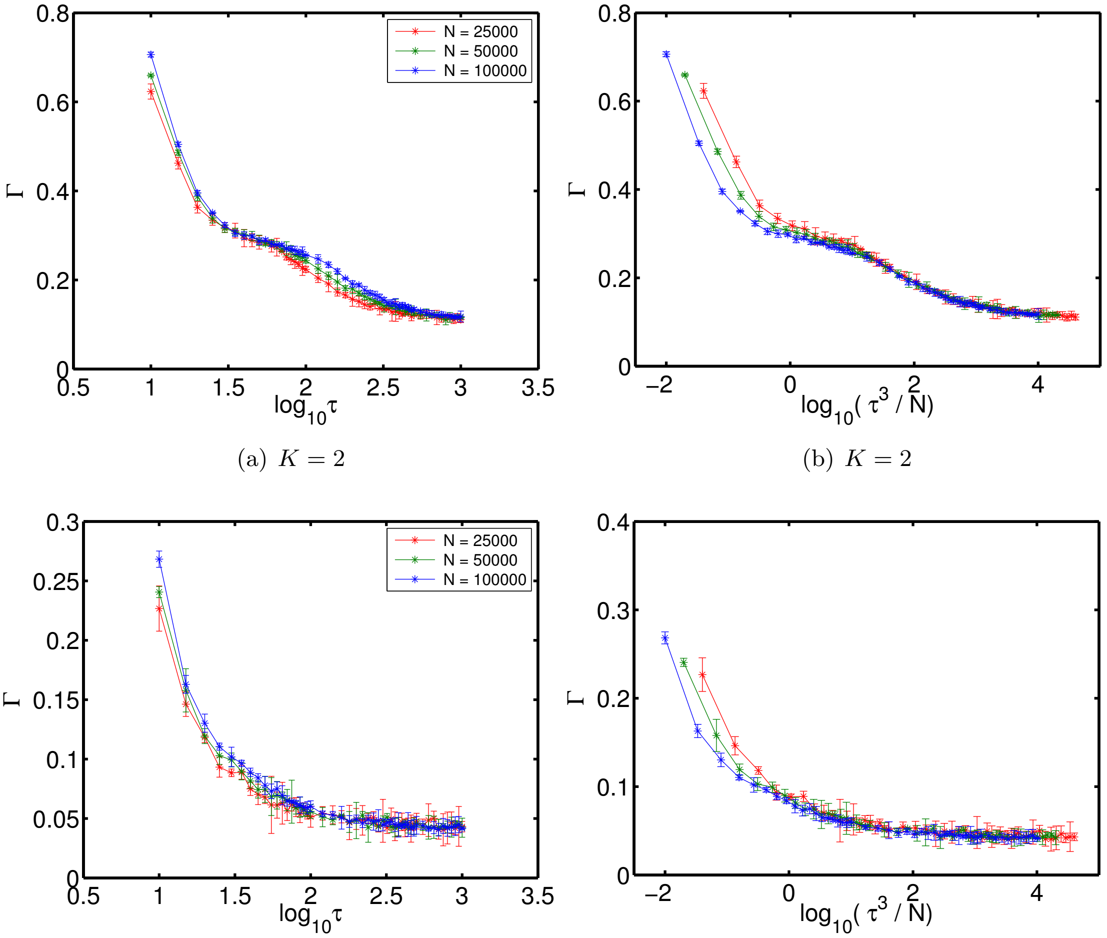
(b) $K = 0.97$, rescaled.

Figure 4.14: Information codimension scaling with sample size N and τ for $K = K_c$. Graphs on the right are rescaled with τ^3/N.

two parts of the earlier graphs becoming similar in shape. There are no longer two distinct regimes as defined by τ^3 less or greater than N. Instead it looks as though for $\tau < 10^3$, $\Gamma \propto \log(\tau^{-\alpha})$ (after some τ this behaviour will stop since Γ cannot go below zero, and any change is assumed to be continuous).

Large K Figure 4.15 shows Γ for large (supercritical) K. This range is characterised by two distinct Γ scaling regimes, and as a result the Γ plots look like a superposition of two parts. It is in the second, *larger* τ range that Γ scales as τ^3/N. This was the scaling related to simple shear, and naturally enough it occurs at a larger τ range than the one that would result from some symmetries in the *chaotic* trajectories. To these we attribute a smaller exponent that one could find heuristically by collapsing the plots.

In terms of PMI scaling we again conjecture existence of some small and large τ linear limits. Yet here the slowing down of Γ decrease that happens between the two Γ scalings also implies that there is a (necessarily) finite range of N during which PMI displays an apparent linear scaling with resolution. Thus for large K we anticipate three finite linear scaling regions. The intermediate one, between two smooth transitions in Γ, suggests an interpretation that is based on a degree of spatial separation between chaotic and regular orbits. This will lead us to suggest the mixture hypothesis that views the joint information dimension as simply a result of an appropriate ratio of the information dimensions of components. In terms of convergence the large K regimes do well, with Γ appearing to level at some small finite τ value that decreases with increasing K.



(a) $K = 2$

(b) $K = 2$

(c) $K = 4$

(d) $K = 4$

Figure 4.15: Information codimension scaling with sample size N and τ for standard map parameter $K > K_c$. Graphs on the right are rescaled with τ^3/N.

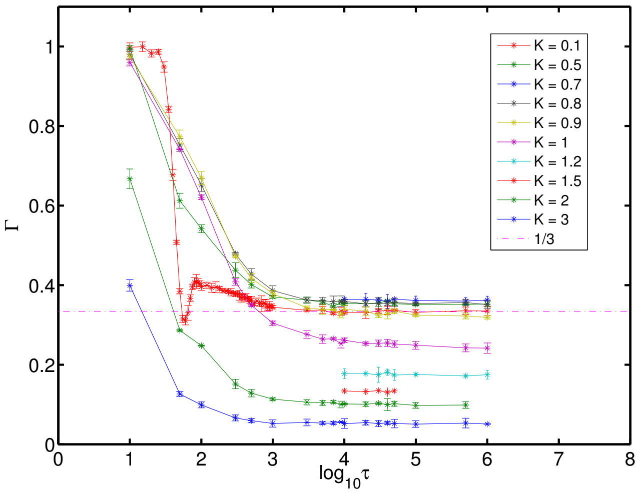


Figure 4.16: Limit of large τ for $N = 50000$. The number of runs Γ is estimated from varies from point to point, but is of order 3. The extent to which plots are populated for small τ varies, and so the dip is only seen on the densely covered $K = 0.1$.

Convergence at the high τ limit

We now investigate convergence further, since especially around K_c it was difficult to form any conclusions. The high τ limit,

$$\bar{\Gamma}(K, N) = \lim_{\tau\to\infty} [\Gamma(K, N, \tau)] \,, \tag{4.24}$$

can be motivated by Permanently Persistent Mutual Information, $I(\infty)$, defined as

$$I(\infty) = \lim_{\tau\to\infty} I(\tau).$$

Given a sample size N, $I(\infty)$ is thus

$$I(\infty) = \lim_{\tau\to\infty} I_0(\tau) + \bar{\Gamma}(K, N)\log(N/k). \tag{4.25}$$

Here it entirely possible that $I(\infty)$ is resolution-dependent.

Figure 4.16 shows behaviour of Γ for a typical sample size when the time gap is pushed to a computational limit. From this we gauge that for some K, for large regions of τ, Γ does not vary significantly (see later graphs for close-up versions of those regions). Γ is observed to plateau for both rather low ($K = 0.1, 0.5$) and the “fully” chaotic ($K \geq K_c$)

regions. Convergence of Γ with τ is markedly nonexistent for $K = 0.9$ and $K = 1$, with what looks like a linear character in the latter case. However, for $K = 1.2$ and $K = 1.5$, though the figures do not have enough data to be shown, the graphs look flat; so does $K = 2$ and $K = 4$.

For $K = 0$ and large K we therefore associate values of Γ at, say, the largest τ considered, with $\bar{\Gamma}(K, N)$. Whilst there is little doubt that this can be done for $K = 0$ and $K = 2\pi$ the intermediate cases are less obvious. The error that would appear if we were to do the same for all values of K is directly related to the rate *and* qualitative manner in which the Γ plots flatten as the time gap grows large.

At $K = 0$, Γ took its large τ limit value at a finite τ value. For small K, what for $K = 0$ was a straight line starting from $\tau^3 \approx N$, now becomes a curve. We infer that even if it has the appearance of a straight line, as in figure 4.11(b), it will after some τ begin to level off, since Γ cannot decrease below zero. The upper limit of Γ stays 1, but the lower limit seems to be almost beyond the visibility in this τ range. It points to the fact that Γ converges to some limiting value at rates dependent on K (so for example it would do so faster at $K = 0.1$ than at $K = 0.5$). In fact as K increases beyond some point (not necessarily K_c) the speed of convergence begins to once again increase, as even in these ranges for large K, Γ appears to have reached some limiting τ value.

Let $r(N, K, \tau) = \frac{d\Gamma(N,K,\tau)}{d\tau}$. Based on figure 4.17 that represents Γ at some N we conjecture that

$$r(N, K, \tau) = r\left(N, K_c^*, \tau + f(|K - K_c^*|)\right) \tag{4.26}$$

In the next chapter we will find that for a particular N, the linear approximation to the gradient of I,

$$\frac{dI(N, K, \tau)}{d\tau} \approx c_1 \frac{dI(N, K_c, \tau)}{d\tau} + c_2 |K - K_c|^a, \tag{4.27}$$

where for $K_c \leq K < 4$, $a \approx 0.8$, and for a region on the other side of K_c, $a \approx 0.65$. We note that for $K > K_c$ the change is abrupt, and after roughly $K = 4$ the slope of PMI with τ stays zero.

There is still the uncertainty about the asymptotic existence of a peak for the small K range - it was present in fig. 4.6 and was then seen to be brought down if higher τ values were considered by examining Γ v τ behaviour for several K. In all the cases Γ is seen to decrease to 1/3 by the time $K \approx K_c$, suggesting that the apparent elevation peaks at some

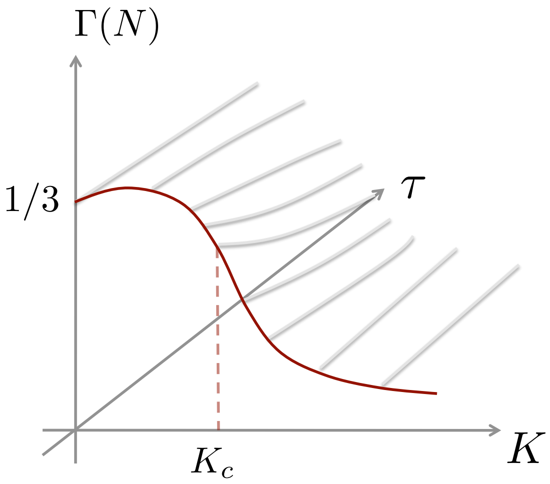


Figure 4.17: Interpretative sketch of $\Gamma(K, \tau)$, showing that for the visible resolution ranges there is a difference between peak location K_p and the regime $\approx K_c$ at which, for that τ range, Γ shows the most rapid decrease. In this picture it is easy to imagine that increasing τ has an effect of making mass flow to both sides away from the peak, so that subsequent rescaling would give the needed value.

K_p that is not the same as K_c. Indeed from the standard map literature we know that nothing special is observed at K_c other than the breaking down of the Golden KAM torus. The peak seems to reflect another phenomenon that is responsible for the increase in the PMI rate with nonlinearity.

Consider again the statistics of nearest neighbour distances. At $K = 0$ all the trajectories are regular and increase at a rate that scales as power law (confirmed below). If at a small and finite K a proportion α of the trajectories has become chaotic, they would still be in regions layered by cantori that are considered 'sticky' in the sense of making trajectories stick by them for a long period of time (which could go up to 10^{10}). As K increases further more regions are freed up, more 'sticky' regions are created, and yet more formerly 'sticky' areas become less restrictive. Indeed all restriction possibly disappear by the time K is comparable to 2π. We infer that if PMI increases faster with K that there exists a level of stickiness such that trajectories diverge slower than in regular quasi-periodic motion. It looks like after the 'peak' parameter value K_p there is simply more free space.

For some N, there seem to be two *separate* K values that characterize $\Gamma(\tau, K)$. $\Gamma(\tau, K)$ reaches its maximum at K_p; but the greatest rate of change with τ is at $K_c^* \approx K_c$. Although these two values may depend on τ, there is at least some τ range for which they are distinct.

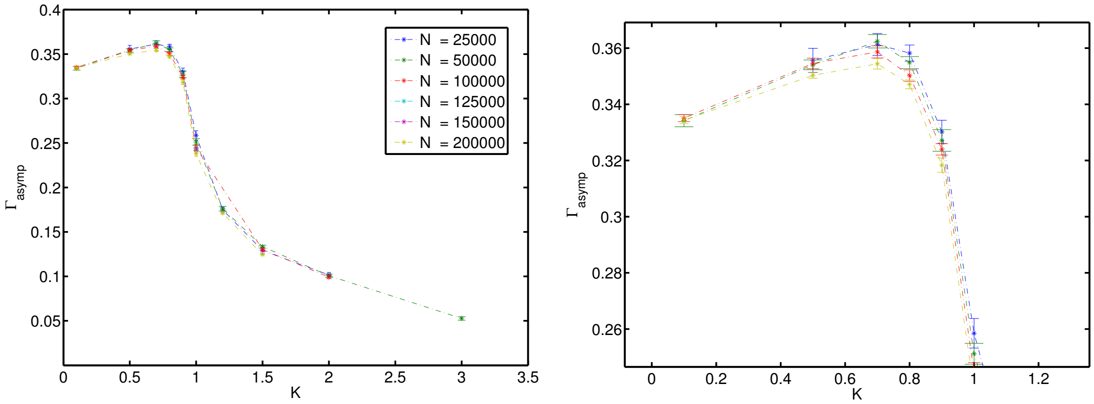
This suggests $\Gamma(\tau, K)$ is a combination of two opposing effects. It is possible that both are related to stickiness; that on the one hand there is the proportion of trajectories affected, and on the other, the effective slowing down that it imposes on them.

We conjecture that the mechanism responsible for the peak is the stickiness of trajectories to the cantori. A higher Γ at those K values means that this behaviour preserves information about the initial condition even better then the periodic and quasi-periodic motion at $K = 0$. After trajectories get 'unstuck' they once again begin to loose information about the past at the rate associated with the chaotic motion in that part of the phase space. The process then repeats. If we therefore assume that this behaviour simply *delays* the destructive effect of chaos on initial correlations, then only out of this analysis in the infinite time limit the peak should not exist, and Γ should decreases monotonically with K.

The reason this might not be the case is the arrangement of the regular/chaotic regions in the phase space. We know that these two are associated with their own specific rates in the limit of infinite τ with which information about the future is destroyed ($\Gamma = 1/3$ and $\Gamma = 0$). However, it is possible that simply *where* the trajectories are - on the scale where only the regions are seen, and not particular trajectories - also contributes towards what the past knows about the infinitely remote future. This structure is not related to level of stickiness but rather to the arrangement of these regions in phase space. Because of this there might be a valid peak, possibly even dependent on the resolution with which we resolve the phase space.

How does the graph of Γ v K look at the largest τ possible? To minimise error we find, for each K, the average Γ over some τ range defined as the largest set of τ values so that the gradient of the line of best fit through $\Gamma(\tau)$ is within some small error of unity, and where we start by considering the largest τ available for that N and move backward. Thus if there Γ still decreases this set would most likely consist of one point. Figure 4.18 shows the result.

From what is observed here K_p is the same for a range of N, but that may of course be simply due to the slow lowering. Another point to make relates to the final asymptotic shape of the (Γ v K) plot. For $K \geq 0.9$, $\bar{\Gamma}(K, N)$ is less than $1/3$. Hence if Γ in K is a stepping down function, the step occurs at $K < K_c$.



(a) Global Trends

(b) Subcritical K

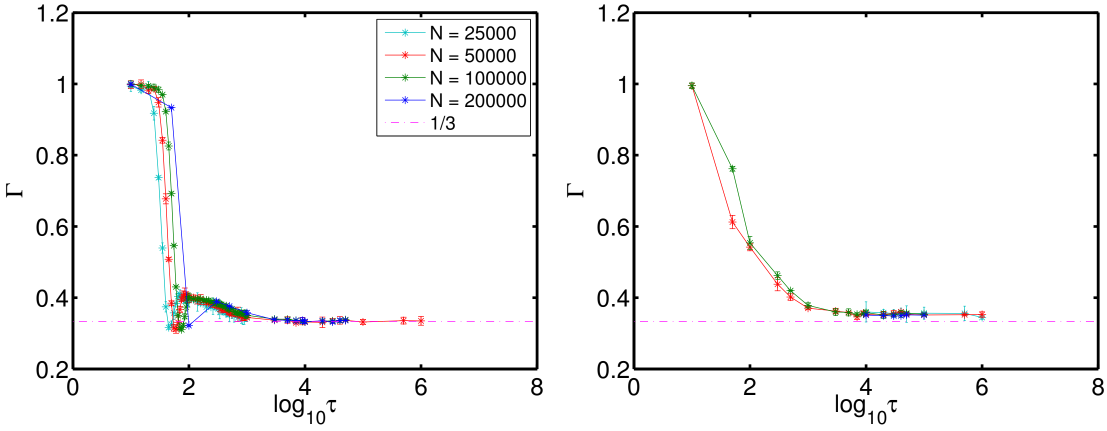
Figure 4.18: Γ averaged over a variable range of relatively large τ (starting backwards from $\tau = 10^6$, see text for details). Errors are standard deviations from the mean over the above range. Colour key the same in both figures.

Variation at finite N and τ

From figure 4.18 we see that dependency on τ changes with N. The final asymptotic values may be independent of sample size, but we have already seen that the (τ, N) interdependency indicate existence of the specific trajectory types. In this section we investigate how Γ changes with N when τ is pushed further towards the asymptotic limit, to see whether any new scaling emerges in these regions. We do not anticipate anything other than a growing influence from sticky trajectories, and hence the only differences we will see will be at intermediate K.

From the high τ figures, if we do see what appears as $\bar{\Gamma}(K, N)$, then it is independent of N, i.e. $\lim_{N\to\infty} \lim_{\tau\to\infty} \Gamma(K, N, \tau) = \lim_{\tau\to\infty} \lim_{N\to\infty} \Gamma(K, N, \tau)$. We associate this to $\bar{\Gamma} = \bar{\Gamma}(K)$, the infinite resolution PPMI. Since therefore in effect $\bar{\Gamma}(K, N)$ does not change with N, it is also the PPMI scaling that is independent of resolution. We can hence conjecture that PPMI is associated with necessarily linear resolution scaling (for this K range). In the fully-integrable case of $K = 0$, Γ asymptotes to 1/3, independent of N. This is implied in the conclusion that the plots collapse. As K increases the lines corresponding to sample sizes separate, Γ decreasing with increased N. After $K > K_c$ the lines begin once again to merge together.

At this point we conjecture that $\bar{\Gamma}(K)$ exists either for all sample sizes, or for none. We also note that the extent to which Γ tends to a final value seems to correlate with how independent of N it is. For example for $K = K_c$ the three plots are quite distinctly sepa-



(a) $K = 0.1$

(b) $K = 0.5$

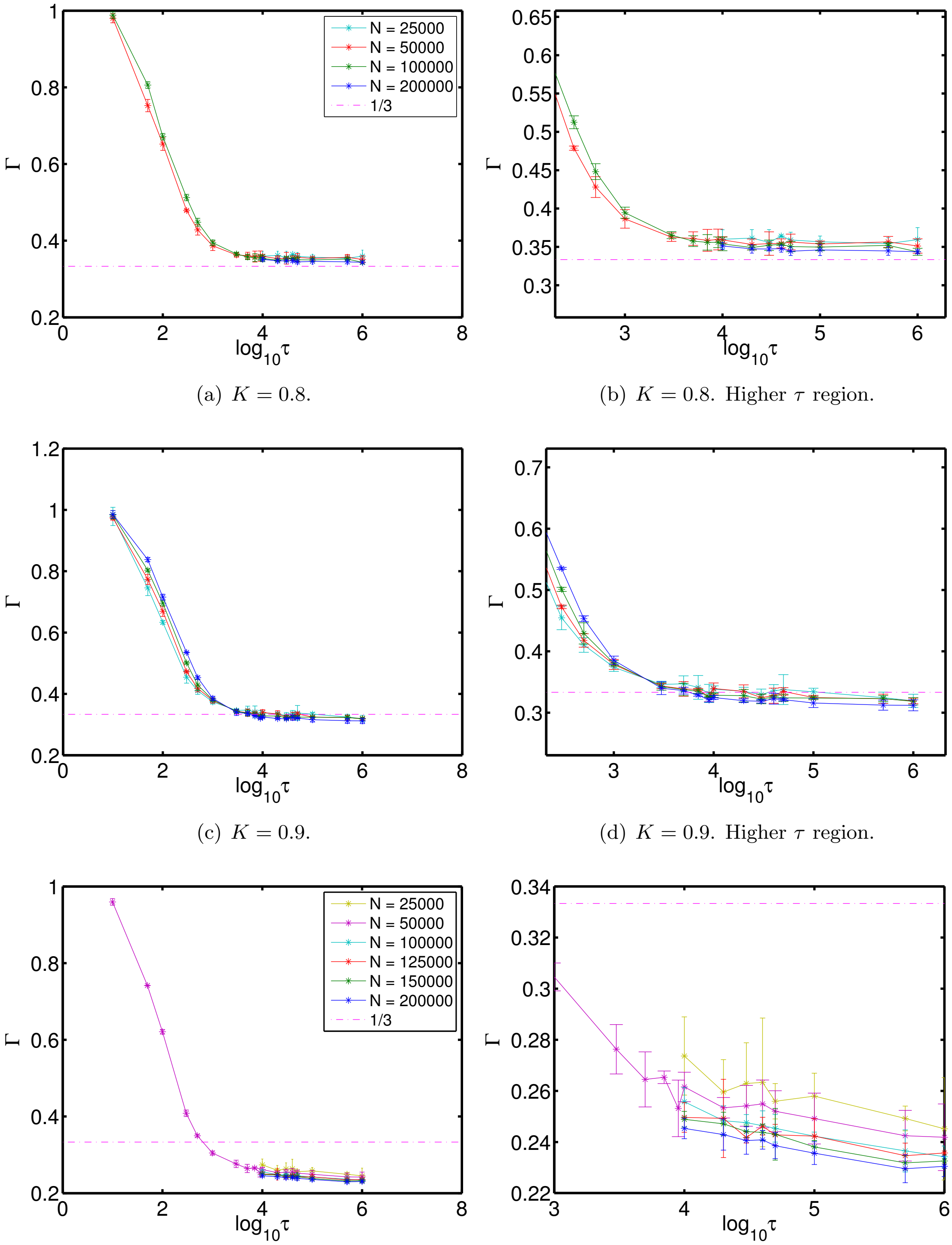
Figure 4.19: Γ v τ for small K. Method of calculation is the same as in figure 4.9.

rated (naturally this depends on the sample sizes themselves since the error bar sizes are correlated). Thus we postulate that a well-defined $\bar{\Gamma}$ does not depend on N.

The Crossover Consider variation of Γ with N at some $\log\tau < 3$ in figure 4.20(c). Higher N results in larger Γ values. This corresponds to the PMI *continuously increasing* with resolution until the slope, after some *finite* value of N (justification for existence of this limit is shown in the next chapter) becomes equal to unity. For smaller N in that region the slope is technically nonlinear, which is directly equivalent to plots of different $\Gamma(N)$ appearing disjoint. We also conjecture that because there *are* correlations at any finite τ, that the small N limit $\Gamma(N_{\text{small}}, K, \tau)$ could exist.

Now consider PMI at $\tau = \tau_c(N, K)$, when all $\Gamma(N)$ plots meet. At that particular τ, PMI scales linearly with N with the gradient given by Γ^*. The question is whether this is indeed true for all N and not just the ones visible in the plots. Since we postulate that it is reasonable to assume that given any τ, an N exists such that all the causal relations are preserved, and also if quite reasonably we then do not expect Γ to jump from unity to Γ^* in no time at all, we must then conclude that for that τ this region of linear scaling of PMI with resolution is *of finite length*, and that at some point, however abruptly, the gradient of PMI will change and tend to unity.

The questions are then whether τ_c is characterised by an actually linear slope of PMI with resolution, or whether the slope is just changing very slowly. Another consideration is whether, for a different τ, a region of resolution exists that appears to give a linear scaling



(a) $K = 0.8$.

(b) $K = 0.8$. Higher τ region.

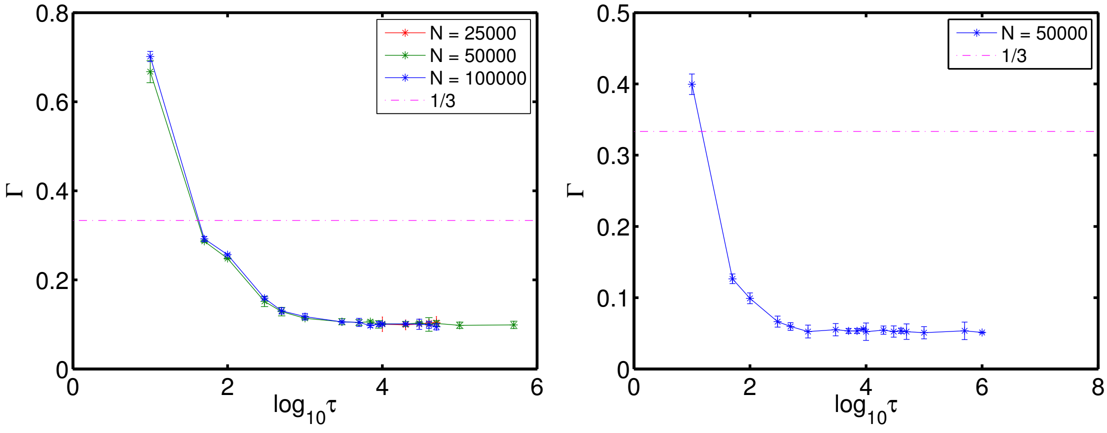
(c) $K = 0.9$.

(d) $K = 0.9$. Higher τ region.

(e) $K = 1$.

(f) $K = 1$. Higher τ region.

Figure 4.20: Γ v τ for intermediate K. Method of calculation is the same as in figure 4.9, colour codes for subfigures 4.20(a) through 4.20(d) same as in figure 4.19.



(a) $K = 2$

(b) $K = 3$

Figure 4.21: Γ v τ for large K. Method of calculation is the same as in figure 4.9.

(inverting to obtain $N(\tau_c)$).

Let

$$\tau_c(N,K) := \left\{ \tau : \left.\frac{d\Gamma}{dN}\right|_{\tau} = 0 \right\}, \tag{4.28}$$

where we identify N with resolution - unless the maximum metric is playing up, PMI at K, τ *is* uniquely defined by $\log(N/k)$. Since in this resolution region Γ is equal for all N, this is actually a contour line (see contour plots later on).

At $\tau > \tau_c(N,K)$ raising N lowers Γ. This visible trend coupled with the already stated assumption that $\Gamma = 1$ should be $\bar{\Gamma}(K,\tau)$ for *any* τ, including $\tau > \tau_c(N,K)$, confirms the requirement that at large N the $\tau_c(N,K)$ line curves.

It might not be obvious how this $\Gamma = 1$ could be achieved - especially when looking at 4.20(d). What one should imagine is the meeting point of the $\Gamma(N)$ lines *moving to the right*, collecting the plots around it. Thus, for a particular τ, increasing N first lowers Γ, but then, after the increase in N made τ smaller than τ_c, Γ begins to rise.

In terms of visualisation, on the landscape of $\Gamma(N,\tau)$, increasing N involves going down, towards the τ_c line, after crossing which an increased N also increases Γ.

What happens in the opposing limit of small N? Since Γ cannot increase with τ (the dip is an anomaly resulting from a relatively non-smooth metric), $\Gamma(N, \tau > \tau_c(N,K)) \leq \Gamma(N, \tau_c(N,K))$. For smaller τ, $\Gamma(N_{\text{small}}, K, \tau) \geq \Gamma(N, \tau_c(N,K))$. It is likely that there is a sample size such that even for τ order of units Γ is very far from unity.

For the visible resolutions τ_c coincides with a constant Γ, but at other N values this need not be so.

Before the crossover, to measure the same information dimension of the joint one has to wait for longer the more trajectories there are to start with. This happens at small timescales, and can be thought of as more trajectories needing more time to spread sufficiently away from each other - the fewer points there are, the sooner this will happen, because the initial distance is then correspondingly larger. After the crossover the opposite happens. Thus, for large time scales, starting with fewer trajectories means having to wait for longer to observe the same information dimension. This is compounded by the fact that chaotic trajectories spend at least some of their time being stuck near cantori, and not exploring the space at all. Thus for large τ at smaller resolutions we see a space of a lower dimension (higher Γ), and at small τ (and therefore temporarily) at smaller resolutions the joint has a higher dimension.

4.3 Summary: Scaling of PMI and Γ through Contour Plots

In this section we summarize the three issues raised earlier: existence of PMI scaling, interrelation of (τ, N) in their effect on Γ, and asymptotic Γ values. Γ to some extent tries to quantify the effect produced by two opposite variables: the growing τ that destroys correlation, and the increasing N that relates to the amount of information available to start with. The last two questions are addressed together by looking at surfaces in the (N, τ) space. Although these contour plots hide the absolute height of the surface they are still useful in assessing the way N and τ are interrelated, and also through the realisation that regions flat in the N direction betray a well-defined PMI scaling.

PMI scaling For every K, for every τ, there exists a finite N beyond which $\Gamma(N, \tau, K) = 1$. Hence $\bar{\Gamma}(K, \tau) = 1$. However, not every N can achieve the $\Gamma = 1$ limit. N can be so small that points wrap around within the first few iterations (and hence at small τ on the typical Γ graphs the lower N limit that does not admit $\Gamma = 1$). Hence at all times PMI has at least one well-defined linear scaling.

At $K = 0$ it is also possible to see another linear scaling regime, associated with the $\Gamma = 1/3$ limit of regular motion. Hence for some τ, PMI will have two *coexisting* linear regimes, the higher one occurring at a higher range of N. The transition between the two will be during a finite N range. If, however, τ is small enough, it is possible to not see the smaller gradient at all.

At all values of K taking a small τ will limit the possible range of PMI behaviour. Thus for example for subcritical K at a relatively small τ (say, before the crossover) PMI would consist of at least one scaling and a long region of N where it is convex. We do not know if Γ associated with the small N scaling is of a finite length, but we conjecture that the second scaling does exist as long as a large enough τ is taken.

Consider a subcritical K, or a K around K_c, for a large enough τ. We postulate that there will be three linear PMI scaling regimes, two of which are of finite length. At large N this is the usual $\Gamma = 1$ limit. Then after a convex region there is an intermediate scaling associated with the cross-over. It is followed by a *concave* region that will be of finite length if τ is large enough.

At high K we should have at least two linear slopes. The possible third slope is an intermediate one, related to the time of regular-chaotic scaling switch. It is visible for only a short

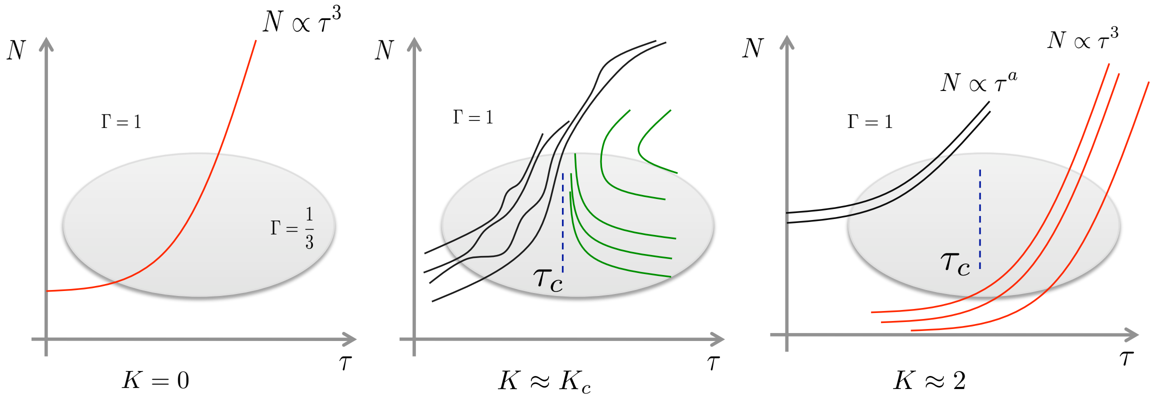


Figure 4.22: Contour plots of Γ, the linear approximation to the gradient of PMI with resolution. Observed range of values is shadowed. The vertical τ_c line is here also a contour line. $a < 3$.

range of N. Its potential existence motivates the mixture hypothesis that we introduce in the next section, and that based on the perceived linearity of PMI tests whether there exists a clear separation between the regular and chaotic components. Coming back, at this high K there might be remnants of the crossover which would add a barely perceptible concave nature to the way PMI converges to a linear slope (but for it it would have done so from a convex region that would give the 'regular' Γ scaling).

Our understanding of the underlying process can be expressed as contour plots of Γ in (τ, N). It is possible to draw these plots automatically using the data behind the various Γ figures, but the result would want clarity in terms of presentation, and we use interpretative sketches instead.

In all the contour plots the top half of the plane will have as a limit the plateau corresponding to $\Gamma = 1$ that begins at a finite N that grows with τ. Additionally, the speed with which the landscape of Γ changes will, for small K, depend on the metric used (which will change the PMI) and the choice of approximation method to the gradient.

Figure 4.22 (a) shows the Γ contour plot for $K = 0$. Since for any τ there will be a $\Gamma = 1$ scaling, that region stretches away to infinite τ as well. The line should be interpreted as a dividing point: no $\Gamma = 1$ scaling to its right, no $\Gamma = 1/3$ scaling to its left. Exactly how close to the line these two can come to depends on both the metric and whether Γ is defined through varying k or N. From the section above we saw that in the usual Γ v τ graphs the change is slow, so the contour plots can have many lines. If, however, the gradient is one-sided and defined by varying N at $k = 1$, the change will be much faster if not instantaneous. Therefore there is a haziness about the neighbourhood of the line.

The blue (dashed) line in figure 4.22(b) defines τ_c, the function of N at which the derivative of Γ with respect to resolution is zero. Here it is shown as a finite interval, indicating the range of resolutions for which it was observed. It will always be locally a contour line, but globally many contour lines may pass through it, since the values of Γ at τ_c may slowly change. Therefore at higher N it will likely curve to the right along with the black contour lines.

These contour lines are curved in unpredictable ways to reflect the lack of clear scaling that happens around K_c. One could imagine them to be made up of lines with regular scaling, chaotic scaling, and scaling that somehow reflects the sticky behaviour. This is a sort of 'crunch zone' when as K increases further the chaotic orbits go from being ones that relax slower than the regular ones to ones that do so faster.

As K increases up to K_c 'regular' scaling shown in red on the subgraph on the left breaks up and the deviations become more pronounced (note that on the first subgraph only one line is drawn; but if the Γ is considered through variation in k we will have a family of lines just like in the other two subgraphs). On the other hand, as K increases beyond K_c, the wavy lines split into two distinct classes, corresponding to the black and red plots in the final subgraph.

The reverse trend with N that happens after the crossover is shown in green. The fact that the three lowest green curves appear equidistant reflects the appearance of scaling present in the plots at high τ, for both N and τ (though we do not have sufficient data to make strong conclusions about the nature of this scaling).

At some point as K increases from 0 the green lines will appear. At the apparent peak of Γ we still see what looks like scaling, but a much slower one. So a contour plot at those values of K will have fewer green lines that are also more widely spaced - both reflecting the higher values of estimated Γ. Thus with K the green plots move in from the left and crowd the black curves, resulting in the region defined as τ_c. The squashed green curves can then be considered as a single curve - with only a few, if any, green curves left on the right side of the plane. If there exists a unique, N-independent value of PPMI for high K then this would correspond to a finite number of green curves that should have infinite for the upper and zero for the lower N limits.

Another interesting issue is the value of Γ at lower N, the lower portions of the plots. The only thing we know for certain is that it will be finite and decreasing with τ, but whether as a step function, or in a continuous manner, is unknown.

Now consider a larger K at which there are two types of scaling. Again, it is entirely possible that if the resolution is large enough the time will wash away correlations due to chaos almost instantaneously, so the shallower curves meet the ordinate. Here the main questions are: how do the different curve types meet, and what happens as $\tau \to \infty$. In terms of the contour lines, if the limit of the red contour lines is *not* finite then the $\bar{\Gamma}(K, N)$ is independent of N. This is shown in figure 4.22 (c).

At $K = 2\pi$, since there are at least two regimes, we also anticipate scaling, but only the chaotic one shown in the figure 4.22(c) above. We also note that as K approaches K_c, the 'regular' scaling stops working. Thus we anticipate, with increasing K, that one type of lines gets broken up, then two types appear, and in the end only the second type is left.

4.4 The Mixture Hypothesis

The general trend of decreasing Γ with K, at least for $K > K_c$, suggests an intrinsic dependency on a feature of the dynamics that becomes less pronounced as the nonlinearity parameter is increased. A natural guess for what this would be is the proportion of regular, quasi-periodic orbits in our sample. Given this function we formulate what we call the mixture hypothesis, which can be summarized as follows: the information dimension of the joint distribution is a linear combination of the information dimensions of the spaces defined by the regular and chaotic trajectories, in proportion to the weight of such trajectories.

The mixture hypothesis is introduced on the basis that at high K values we see an apparent linear PMI scaling with resolution *at intermediate values of* τ, between the regular and chaotic scalings of Γ with N. This suggests that there is a time when, for a sample size, the chaotic trajectories are sufficiently mixed, and the regular ones will start mixing after that time. In other words, the mixing processes are distinct, and so trajectory types can clearly be segregated into distinct spatial regions that are well-defined on the scale given by N; components do not appear to mix.

Any point x in the standard map state space X will give rise to a trajectory $T(x) = (x, Fx, F^2x, ..)$. Let us postulate existence of certain (finite) trajectory characteristics which allow partitioning of the set of all trajectories into ones that are chaotic and ones that are not - for example existence of a necessarily finite time t, which might be different for each $T(x)$, but for which $T(x)$, as truncated after t elements, definitely falls into one of the two categories. This leads to a corresponding partition of X: define the chaotic component as

$$X_c = \{x \in X : T(x) \text{ is chaotic}\}. \tag{4.29}$$

X_c is, by definition, closed under the action of the map. Thus the regular component is $X_r = X/X_c$, the complement of the chaotic one.

Let μ be a measure over some suitable σ-algebra on X. Define $\alpha = \alpha(K)$ as the weight of the regular component of the standard map at parameter K:

$$\alpha = \mu\left(X_r\right). \tag{4.30}$$

It should be noted that here we are making implicit the K-dependency of F, and hence the $X_{r/c}$ partitioning, and α, just as we are dropping the μ dependency of α. This is because we will always assume μ to be the measure corresponding to the uniform distribution over the X. Thus α depends on both K and μ.

Let $\mu^{r/c}$ be the (normalised) regular/chaotic measures restricted to the power sets of $X_{r/c}$, such that

$$\mu(A) = \alpha\mu^r(A) + (1-\alpha)\mu^c(A)\ \forall A \subseteq X. \tag{4.31}$$

Then if the X_r and X_c are sufficiently disjoint,

$$S[\mu] = \alpha S[\mu^r] + (1-\alpha)S[\mu^c] - \alpha\ln\alpha - (1-\alpha)\ln(1-\alpha). \tag{4.32}$$

From the usual definition of the information dimension this is then

$$S[\mu] = \alpha D_r \ln\epsilon + (1-\alpha)D_c \ln\epsilon, \tag{4.33}$$

and so the information dimension of the joint μ is

$$D_m = \alpha D_r + (1-\alpha)D_c. \tag{4.34}$$

The mixture hypothesis then is that $D(\mu_J) = D_m$. We also for now assume that in the standard map, $D(\mu_J^r) = 3$, and $D(\mu_J^c) = 4$. This gives

$$D_m = 3\alpha + 4\,(1-\alpha) = 4 - \alpha. \tag{4.35}$$

Moreover, if

$$\frac{D(\mu) + D(F^\tau\mu) - D_m}{D_m}$$

defines Γ_m, then

$$\Gamma_m = \frac{\alpha}{4-\alpha}, \tag{4.36}$$

where we have used the fact that the marginals have information dimensions equal to two. Thus Γ is dependent on K and μ through α, but not on τ. This dependency was hidden in the specific choice of information dimensions of the regular/chaotic joint distributions, about which more needs to be said.

Regular and chaotic trajectories are different in character. We used this to assume that the space of all orbits can be partitioned. Attributing a definite information dimension to the joint distribution of a class is trickier, simply because $D(\mu_J^{r/c})$ will be a function of the time gap τ between the initial and final iterates. Here we assume not only that $D(\mu_J^{r/c})$ exists for all τ, but that for all K there exists a limiting information dimension, $\lim_{\tau\to\infty} D(\mu_J^{r/c})$, which will be 3 for the regular and 4 for the chaotic trajectories.

In practice this translates to a statement about the infinite τ limit of the computed Γ (which is also the definition of PPMI). If the mixture hypothesis *and* the linear scaling of PMI with resolution holds, then

$$\bar{\Gamma}(K, N) = \Gamma_m, \tag{4.37}$$

$$\bar{\Gamma}(K, N) = \frac{\alpha}{4-\alpha}. \tag{4.38}$$

This is supported (and partially motivated) by the fully-chaotic and the fully-integrable scenarios. At $K = 0$ all the orbits are regular, so $\alpha = 1$ and $\Gamma_m = 1/3$. This is in agreement with $\bar{\Gamma}(K = 0, N) = 1/3$, after Γ moved down from the fully-causal limit of 1. When $K = 2\pi$ we only resolve chaos, so both α and Γ_m are zero, once again in agreement with the experimental results. So at least for these limits eq. (4.38) holds.

Implementation and Analysis

The aim of this section is to explain the strategy for testing eq. (4.38) for arbitrary K values. We introduce a method to obtain α numerically by considering distributions of evolved distances between trajectory pairs. We then compare Γ_m to our best estimates of $\bar{\Gamma}(K, N)$, shown in figure 4.18.

We do this by using divergence rates as an equivalence relation on the chaotic/regular classes. Consider a pair of trajectories a distance $\epsilon_{t=0}$ apart. If both are chaotic then

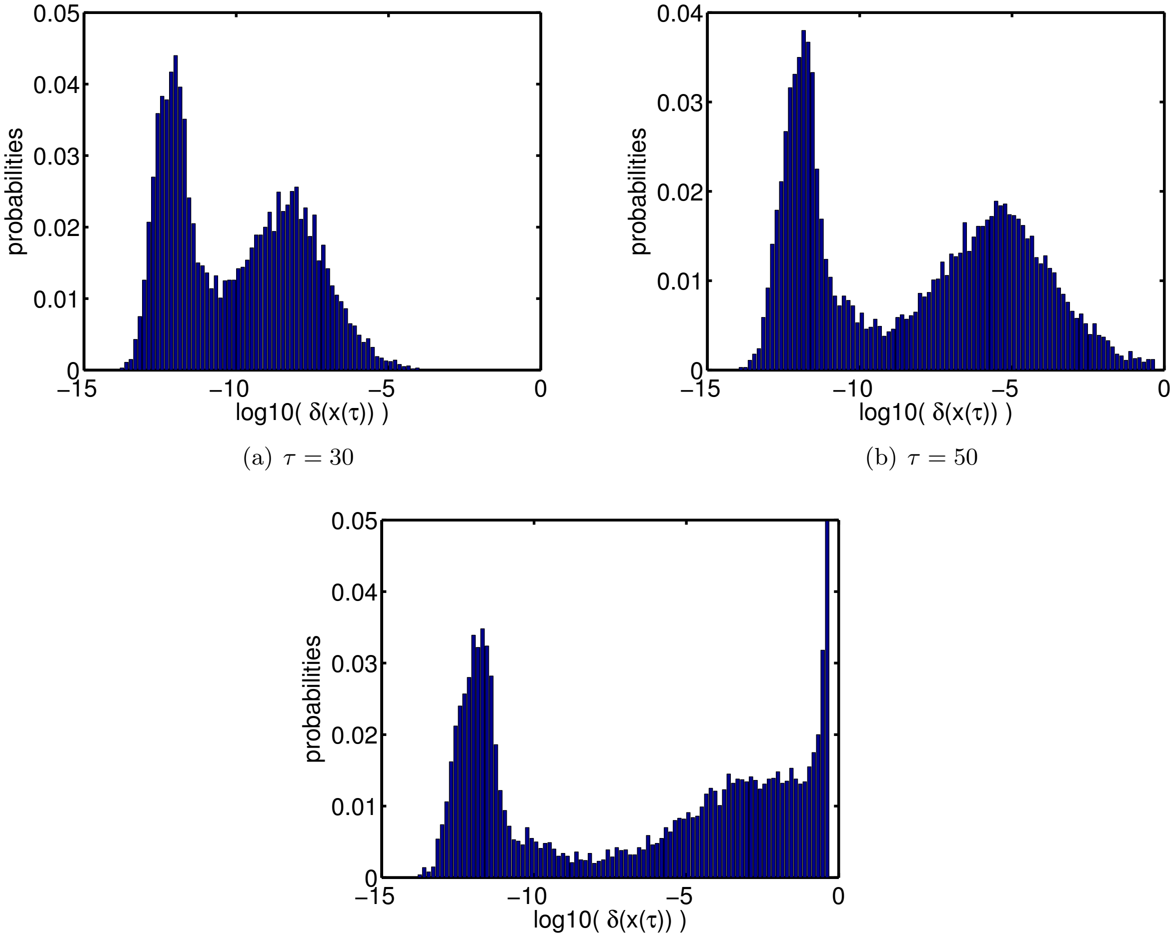
$$\epsilon_t \approx \epsilon_0 \exp^{\lambda t}, \tag{4.39}$$

where λ is the Lyapunov exponent, whereas for regular ones

$$\epsilon_t \approx C t^{\nu}. \tag{4.40}$$

Hence tracing the evolution of ϵ_t allows us to classify the pairs as belonging to either of the classes. Note that if the pair has one of both kinds, then the separation is unlikely to be increasing at a regular rate, and hence we assume that the exponential divergence can also be indicative of a regular-chaotic pair.

The next step is to find the proportion of, for example, the exponentially divergent pairs out of a large sample of pairs, that are characterised by always having one of each pair's initial points being sampled from a flat distribution over the configuration space. Thus we sample the required initial distribution, create a nearby point for each element, and examine the rates of their divergence to classify the element. In other words let the set of sampled pairs be $\{(X = x, Y = y) | X \sim \mu, Y \sim p(y, \epsilon, d)$ s.t. p is flat $, d(x, y) = \epsilon_0\}$. This introduces a distribution ρ_t for ϵ, the distance between orbit pairs. At $t = 0$, ρ_0 is a delta function centered on ϵ_0. At times t, ρ_t is defined through $d\left((F^t(X = x), F^t(Y = y)\right) \sim \rho_t$.



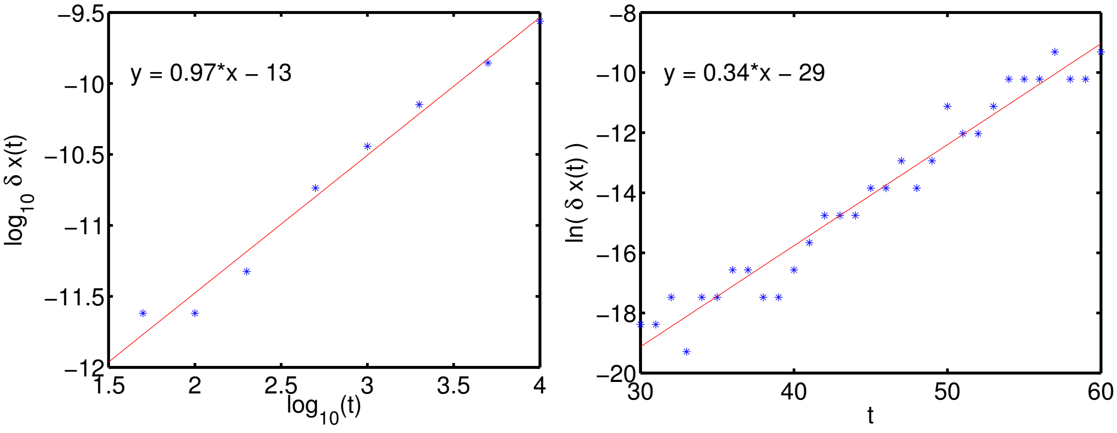
(a) $\tau = 30$

(b) $\tau = 50$

(c) $\tau = 70$

Figure 4.23: Histogram of distance between τ^{th} iterates initially separated by 10^{-12}. $N =$ 10000 points were considered; standard map parameter $K = 1.5$. Trajectories are seen to be split into two types depending on the rates of divergence.

It is the clear bimodal shape of subsequent ρ_t, and the fact that the two peaks evolve at different rates, that makes it possible to classify the underlying distances as either being associated with a chaotic or regular orbit pair. In figure 4.23 we see three instances of histograms corresponding to ρ_t for $K = 1.5$, when the phase portrait of the standard map shows islands of quasi-periodic motion surrounded by the chaotic sea.



(a)

(b)

Figure 4.24: Evolution of peaks corresponding to regular 4.24(a) and chaotic 4.24(b) trajectories. $K = 1.5$, $N = 10000$, initial separation 10^{-12}. Parameters are the same as in 4.23, though the time ranges involved, as well as the binning methods may differ.

The initial peak is seen to split into the slow- and fast-moving regions, which we associated with regular and chaotic pairs by tracing the rate of evolution of regional peaks (fig. 4.24). Here, for example, we find that the Laypunov exponent at $K = 1.5$ is found to be ≈ 0.33. Accurate measurements would attach an error based on the bin width, number of elements in the sample of the distribution, number of time measurements, and possibly the initial separation (see later), but here we are interested in merely in showing that exponential divergence does indeed happen for some trajectories, rather than in exact numerical quantification of its manner.

Hence we see that by introducing a cutoff distance ϵ_c and a time τ_c the following can be assumed: the relative number of trajectory pairs whose separation $\epsilon < \epsilon_c$ for some $\tau > \tau_c$ (or a range of such τ values) corresponds to $\mu(X_c)$, the weight of the chaotic component of the map at some K. Underlying this is the assumption that distance between trajectories is a valid equivalence relation.

There are several sources of error in the estimate of α obtained in this manner. Wrongly classifying trajectories temporarily stuck amongst the cantori debris will tend to overestimate α. The magnitude of this problem will vary with K, since it is safe to assume that some regimes are more likely to result in stuck trajectories than others. This, on the other hand, will also be dependent on the initial separation - the cantori will come with characteristic sticky widths. A relatively large initial separation would increase the likelihood of picking trajectories of different character, and thus overcounting the chaotic ones and underestimating α. Underlying it all is the assumption that unless both trajectories are regular the distance between them increases faster than a power law - which might not be the case if both are stuck - but then larger τ_c might prove of help. Larger τ_c might, however, decrease distances between points, since the exponential divergence is only true for short time scales - which would increase α.

There is another consideration halfway between these conceptual hurdles and the more numerical obstacles en route to sampling a dynamical system. It is that we classify trajectories based on the divergence rates of their arbitrary element with a very specific set of points around that element, as defined by a distance *and* a metric. Now, Lyapunov exponents define the rate of expansion and contraction of subspaces. By picking a distance and a metric we limit ourselves to only a subset of local neighbourhoods, and there is no guarantee that the deformation of that subset will be representative of the subspaces. The problem may be remedied slightly by considering a variety of initial displacements.

The more numerical considerations rest on the tacit understanding that all this is an analysis of the double-precision version of the standard map. Its 'many-to-one' nature may result in effective trajectories that are made up of parts of chaotic and parts periodic sections, since the inevitable approximation to a subset of the rationals intrinsic to every step may move the point to a region with a different character. This, however, is something that we take for granted as not influencing the outcome.

Information Dimensions We first check whether our assumption about the regular/chaotic information dimensions are valid by computing α and then estimating the respective information dimensions using the K-G estimator.

It is tempting to estimate α by inspection alone. At $K = 2$, 50 evenly spaced bins on a logarithmic scale between 10^{-15} and $1/2$ result in histograms suggesting that for $\tau_c = 90$ and $\epsilon_c = 10^{-7}$ the chaotic peak has become sufficiently separated from the regular one, for a particular ϵ_0 (parameters in figure below). We therefore track those trajectories and estimate their information dimension. Once again, we associate the information dimension with the linearized slope of the curve of the estimate of Shannon entropy with resolution in the form of $\Psi(N) - \Psi(k)$, for a particular value of N and the first five values of k, just

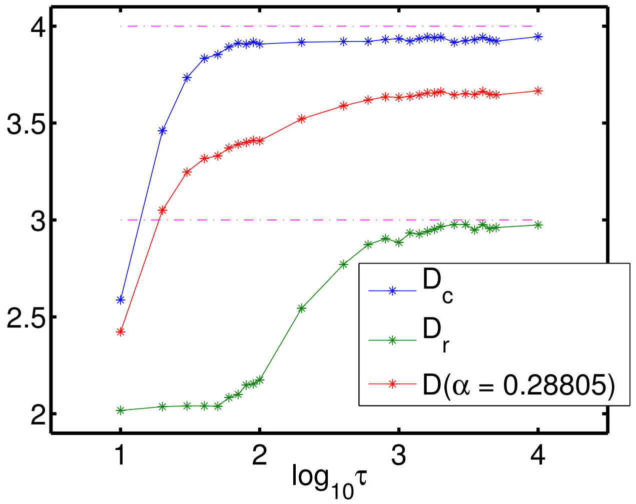


Figure 4.25: $K = 2$. Information Dimensions of the joint distribution (at τ) of trajectories whose separation from their neighbour at $\tau = \tau_c = 90$ was less than $\epsilon_c = 10^{-7}$. Average is over 3 runs. The weight of these ('regular') trajectories is denoted by α, though technically at this point it is $\hat{\alpha}$, an approximation. D stands for the information dimension that is the sum of the regular and chaotic information dimensions weighed by α, and is the same as D_m in the analysis above. Initial separation $\epsilon_0 = 10^{-12}$.

as in the PMI analysis above. Only when the slope is linear with the resolution does the gradient correspond to the information dimension, though we use the name more generally.

Figure 4.25 shows the regular, chaotic and the composite information dimension for $K = 2$ using the cutoff parameters for α found by inspection. From above, we expect the regular information dimension to be 3, and the chaotic one to be 4. The computed information dimensions for both show a small systematic error that lowers the values, more so for the chaotic case. We also see that the it takes time for both trajectory types to 'cover' their respective subsets, and for the measured information dimension to even begin to get closer to the expected value. Naturally enough chaotic trajectories take less time. In fact the almost steady value of the regular information dimension at lower τ can be attributed to motion 'before' the wrapping, where the fully-causal limit is realised.

It is exactly this difference in the manners in which the plots increase that produces the kink at the joint information dimension seen at midrange time scales. This analysis confirms our supposition in the section above that this kind of two-stage behaviour for large K reflects the transition between chaotic and regular scaling. We would therefore see plots of information dimensions for different N scale with those two distinct laws.

It is possible that estimation of α parameters by inspection is the reason behind the slight offsets of information dimensions. We therefore need a more systematic way with which to find the time and distance cutoff parameters.

Another source of error is a computational one. In practice we sample the set X_0 of initial conditions that is a proper subset of X. Yet if some trajectories are unavailable for numerical study it should be reflected in both the PMI and the observed α. Therefore any discrepancy between Γ and Γ_m will not stem from the impossibility of sampling the state space of the standard map. If, however, α were to be obtained analytically from theoretical investigations of the mapping, then an error could potentially arise.

This consideration points to a procedure that could test the extent to which X_0 represents the map in terms of containing different types of trajectories: a theoretical α_t could be compared with a measured one. However the use of a such a comparison is only clear if the MH were to hold: the PMI for the 'true' standard map could be expressed as a function of α_t. Yet this would also greatly reduce the necessity to write down PMI in the first place, at least insofar as its role in understanding the map is concerned: since it would then be α_t, and not the derivative PMI, that would be used to examine the standard map, and which would have been already performed. Hence carrying out the test only makes sense if we want to extend the validity of our results to the 'true' standard map, in which a positive outcome, whilst being a prerequisite, would not be the only requirement.

Testing α

We desire to obtain a best estimate of α, the uniform measure of a set of elements of regular trajectories of a double-precision version of the standard map defined on a subset of the rationals attributed to some standard computer architecture. The method described above has six parameters:

N The number of trajectory pairs. Limited only by the computation speed.

ϵ_0 Initial pair separation. We take the smallest value to be around 10^{-12}, significantly higher than the machine-epsilon for double precision. Its largest value is dependent on both ϵ_c and τ_c, since the three can need to be such so as to allow for sufficient separation between regular and chaotic distances. It is defined by

d the metric, which we keep to be the maximum one, to correspond with one used in the

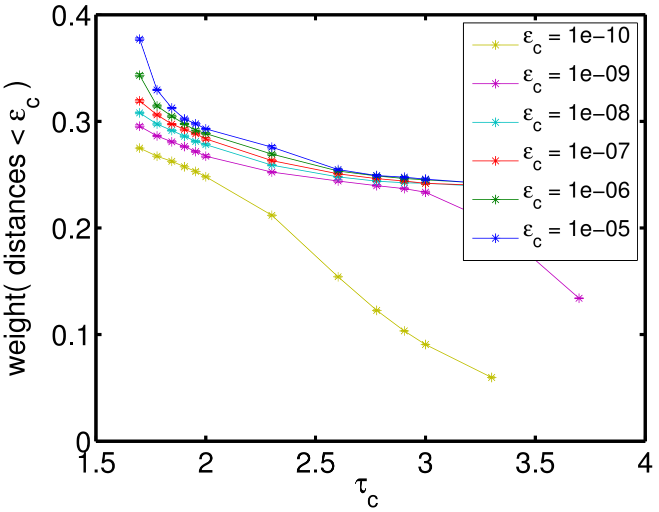


Figure 4.26: Relative number of 50000 trajectory pairs whose separation using the maximum metric at τ_c does not exceed ϵ_c. Initial separation of $\epsilon_0 = 10^{-12}$, for standard map parameter $K = 2$. Each value is an average over 10 runs.

PMI calculations.

M The number of runs, which in some sense overlaps with N, but the presence of which allows calculation of the standard error of the mean. This shall be varied.

ϵ_c The maximal distance two trajectories can be separated by in order to still be classified as *both* being regular, and

τ_c The time at which the distance is calculated.

All these variables are interdependent. In all cases a suitable range of possible ϵ_0, ϵ_c and τ_c is best judged by inspection. As an example we consider $K = 2$, the case discussed above, with $N = 50000$, $\epsilon_0 = 10^{-12}$ (the parameters used), $M = 10$, and a range of ϵ_c and τ_c values to compute the mean fraction of trajectory pairs whose separation at τ_c does not exceed ϵ_c.

Figure 4.26 shows these results for a range of τ_c. The decrease at short timescales corresponds to the chaotic distances leaving the allowed range. Downward slopes at large τ_c are due to regular distances increasing beyond the ϵ_c limit. We conjecture that α is some average over the plateau of $\hat{\alpha}$.

The number of trajectory pairs considered is adequate, giving small enough errors for ($M =$) 10, 5, or even 2 runs (results not shown). This leaves the only untested parameter as the

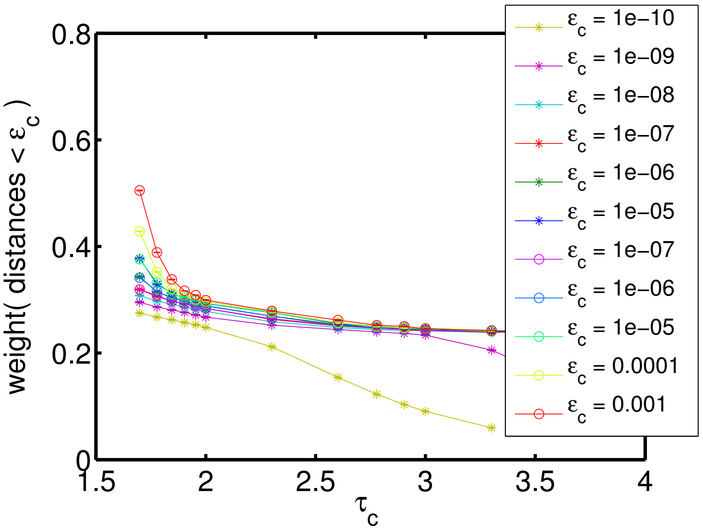


Figure 4.27: Relative number of 50000 trajectory pairs whose separation using the maximum metric at τ_c does not exceed ϵ_c. Initial separation of $\epsilon_0 = 10^{-12}$, averaged over 10 runs, for standard map parameter $K = 2$. The graphs with o show the same but for the initial separation of $\epsilon_0 = 10^{-9}$ (averaged over 5 runs).

initial orbit separation, at least for $K = 2$.

From the analysis above we expect a certain N, τ scaling for regular trajectories. As such the initial orbit separation, which is a function of the sample size N used in PMI calculations, should also have a clear scaling relation to τ (see next section). We therefore anticipate that for each ϵ_0, there exists a combination of ϵ_c, τ_c that cause a graph with a different ϵ_0 to overlay the former. This would mean that the average over the plateau stays the same, and that α does not depend on ϵ_0, which is to be expected. Figure 4.27 shows a graph supporting this notion, where some plots of $\epsilon_0 = 10^{-9}$ lie on top of the previous data.

The lower limit on ϵ_c is given by the largest distance that could, at the given time, separate two regular trajectories. If ϵ_c is below this limit then α would be underestimated. In practical terms a decrease in $\hat{\alpha}$ due to this effect is clearest when the regular and chaotic trajectories are separated in two clear peaks, as is the case for $K = 2$. The regular peak begins to traverse ϵ_c, which explains the drastic drop on the right hand side of fig. 4.26. This limit can be made more precise in anticipation of the case around $K \approx K_c$, when relaxation is slow and the separation distribution may not be bimodal and hence clear. In

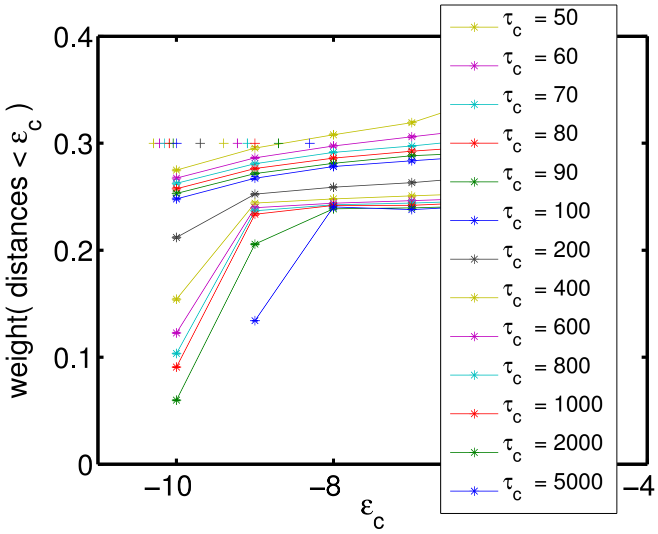


Figure 4.28: Relative number of the 50000 trajectory pairs whose separation using the maximum metric at τ_c does not exceed ϵ_c. Initial separation of $\epsilon_0 = 10^{-12}$, for standard map parameter $K = 2$. Each value is an average over 10 runs. The + symbols are positioned at the analytically-estimated maximal separation of regular trajectories at τ_c.

the next section the maximal separation for regular trajectories is derived to be

$$\epsilon_c^{\min} = \log_{10}(\epsilon_0) + \log_{10}(\tau + 1). \tag{4.41}$$

We check this on the already-familiar data for $K = 2$. Figure 4.28 shows the same information as figure 4.26, but now the plots correspond to different τ_c values. The plot is zoomed in on the + signs. Each is positioned at a $\epsilon_c^{\min}$ value corresponding to the τ_c of the same colour. As such each provides an effective left cutoff, so that points to the left of the + of the same colour are weights of only *part* of the regular trajectories. As expected, we see that the values below the intuitive leveling at around $\hat{\alpha} \approx 2.4$ can thus be disregarded.

Effect of α on regular and chaotic information dimensions We are now in a position to assess the correctness of the plot of information dimensions of regular and chaotic components shown earlier in figure 4.25. The τ_c and ϵ_c values were, respectively, 90 and 10^{-7}. From fig. 4.26 above, which corresponds to the same ϵ_0, this would give $\hat{\alpha} \approx 0.288$, agreeing with the value obtained as an aside during the procedure itself. However, the truer value of α is ≈ 0.24, when the weight plateaus. Since τ_c should be made as low as possible we consider two τ_c and ϵ_c values that would give an α that is reasonably close to the plateau, that of ≈ 0.242, and calculate the information dimensions of the trajectories so defined.

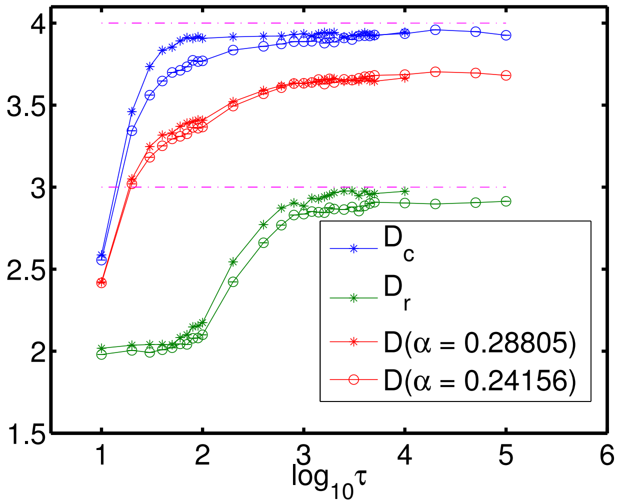


Figure 4.29: Information Dimensions of the regular and chaotic components for $K = 2$. Same as in figure 4.25, but supplemented by data in (o) obtained with a better estimate of α computed at $\tau_c = 1000$ and $\epsilon_c = 10^{-8}$.

The result, superimposed on the earlier data, is shown in figure 4.29. Contrary to moving the information dimensions to their expected values of 3 and 4 the effect is actually the reverse. The reason behind it is the same reason that causes the blue plot at low values of τ to come down: we see that if we allow more pairs to escape, making less mistakes in trajectory categorization, that for a range of τ the information dimension of the joint chaotic component appears smaller. This is because we added trajectories that do not range over the space as fast as the chaotic ones, bringing down the average interpoint distance and hence the information dimension. After some time this is remedied, however, and at large τ, D_c does not change. The same effect brings down the information dimension of the regular component in the joint - by removing pairs that would otherwise result in interpoint distances large enough to raise D_r.

Our expectations of the values of D_c and D_r are based on the assumption that in the marginal the chaotic and regular components have certain integer information dimensions. When the structure of the state space becomes complicated this may not necessarily be true. However the actual values are not computationally obtainable - here we relate D to the slope of measured entropy with resolution; doing the same in the marginal cases does not give a well-defined slope for either of the components, at least at $K = 2$. For the joint the case is clearer, though rather expectedly the chaotic component gives a slope with more

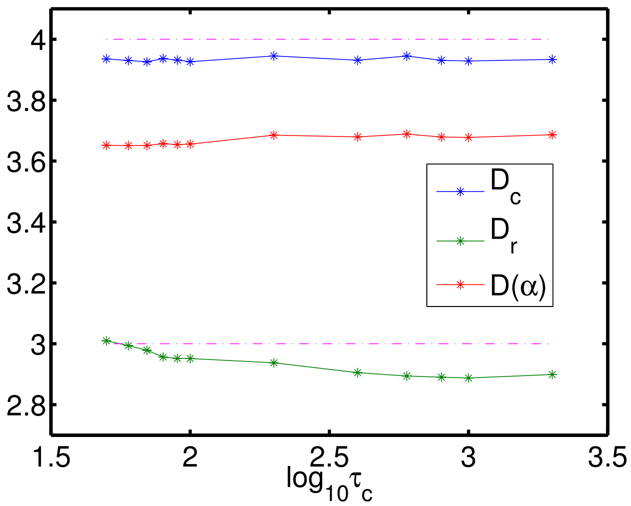


Figure 4.30: $K = 2$. Information Dimensions of the joint distribution at $\tau = 10^4$ of trajectories whose separation from their neighbour at τ_c was less than $\epsilon_c = 10^{-8}$ (denoted by r), and their complement (c), as well as the averaged quantity $D(\alpha) = D_m$, where α is the weight of the former, computed for each τ_c. This curve does not decrease as much as D_r since the weight of the regular contributions decreases with τ_c according to the respective graph in fig. 4.26. Technically α is $\hat{\alpha}$, an approximation.

errors than the regular one.

Consider a relatively high $\tau = 10^4$, for which the dimensions have - roughly - reached equilibrium. For the two different cutoff parameters shown in the graph above the chaotic dimension is seen to stay the same, while the regular one decreases to below 3 as more time is given for the chaotic trajectories to leave. In figure 4.30 we show the dimensions for a range of cutoff values, effectively following the marine curve in figure 4.26. The chaotic dimension is found to stay roughly constant, yet below 4. The regular dimension, however, is seen to continue decreasing below 3 and then level off. This points towards considering a regular dimension of 3 as not the true dimension of regular trajectories, but rather merely a value obtained by including some chaotic trajectories in the sample, which raises the effective dimension.

We therefore treat the slower chaotic trajectories that leave the regular component as having the same dimension as the main chaotic component, since their presence does not change the dimension of the latter, but certainly alters the regular dimension. The fact that $D(\alpha)$ is almost level indicates that the decrease in the regular dimension is in line with the decrease in the weight of those trajectories that make up the regular distribution (and which, before α settles, would include some chaotic ones).

We have seen that without establishing a plateau with τ_c the discretization procedure alone is prone to various errors. However, the difference between regular/chaotic information dimensions due to a different α gets somewhat lessened when the same α changes the proportion of their altered contributions to the D_m. Thus for $K = 2$ the error from using the wrong α is seen to be relatively small.

Figure 4.31 shows the resultant Γ plots. We first note that if we wish to include time-invariant dimensions of the joint components then we would have to talk about the limit of Γ as $\tau \to \infty$. For finite τ the shape of Γ is seen to be the result of the combination of information dimensions of spaces that have so far been explored by the regular and chaotic trajectories. Yet, more importantly, even in this case when we are sure of α and $D_{r/c}$ to relatively small errors, the Γ_m still underestimates the measured Γ (errors are not shown here, but they are smaller than the distance between the plots). Information dimension of the joint is thus, albeit by a small amount, smaller than the proportionate information dimensions of components.

The only way in which it is possible is if in the computation of the joint the neighbour distances were sometimes realised by trajectories of different character. This in turn suggests that the difference between the two values can reflect the extent of 'interlocation' of the regular/chaotic parts. It is interesting that time wise the largest difference happens when the scaling with resolution changes from chaotic to regular.

We now consider a different K, $K = 0.9$. From the standard map theory here we expect α to be larger than at $K = 2$. We also expect Γ_m to be further away from Γ, since we assume that the extent of spatial mixture of regular and chaotic trajectories is greater around K_c.

Figure 4.32 shows the pdf of distances for $K = 0.9$. What is immediately sticking is that here it looses its bimodality, the exact feature that made this framework so amenable to obtaining α. There is no longer a clear peak corresponding to the chaotic trajectories moving to the right. What is shown here is that these no longer have a distinct typical speed of separation. Instead the distances between chaotic pairs leave the peak gradually and at varying times (though whether it is a combination of these effects is an open question).

In terms of estimating α this means looking for a combination of cutoff points that is both

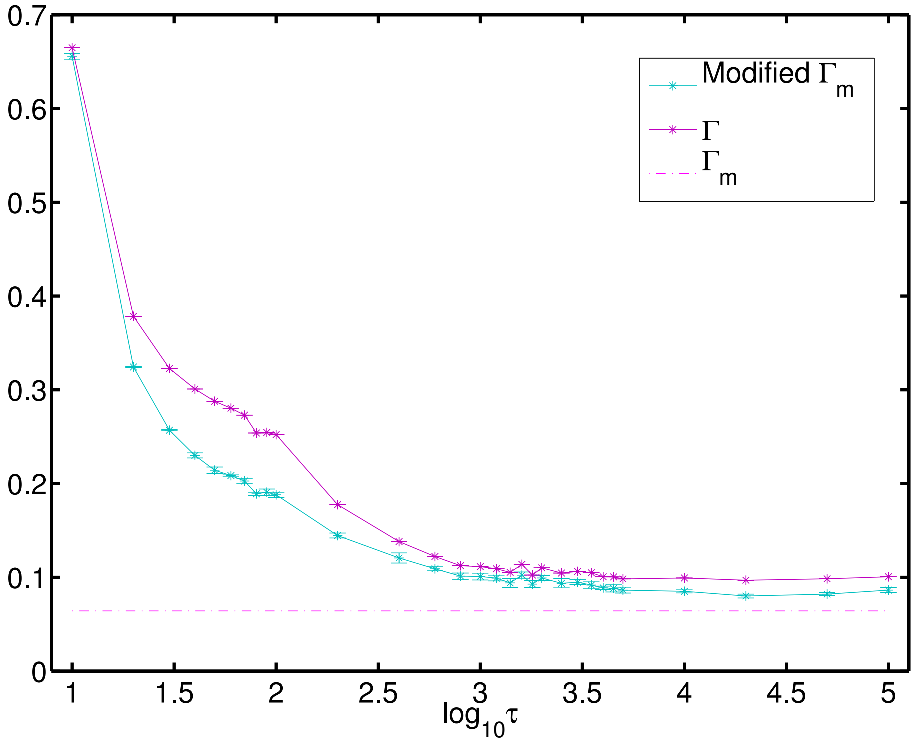
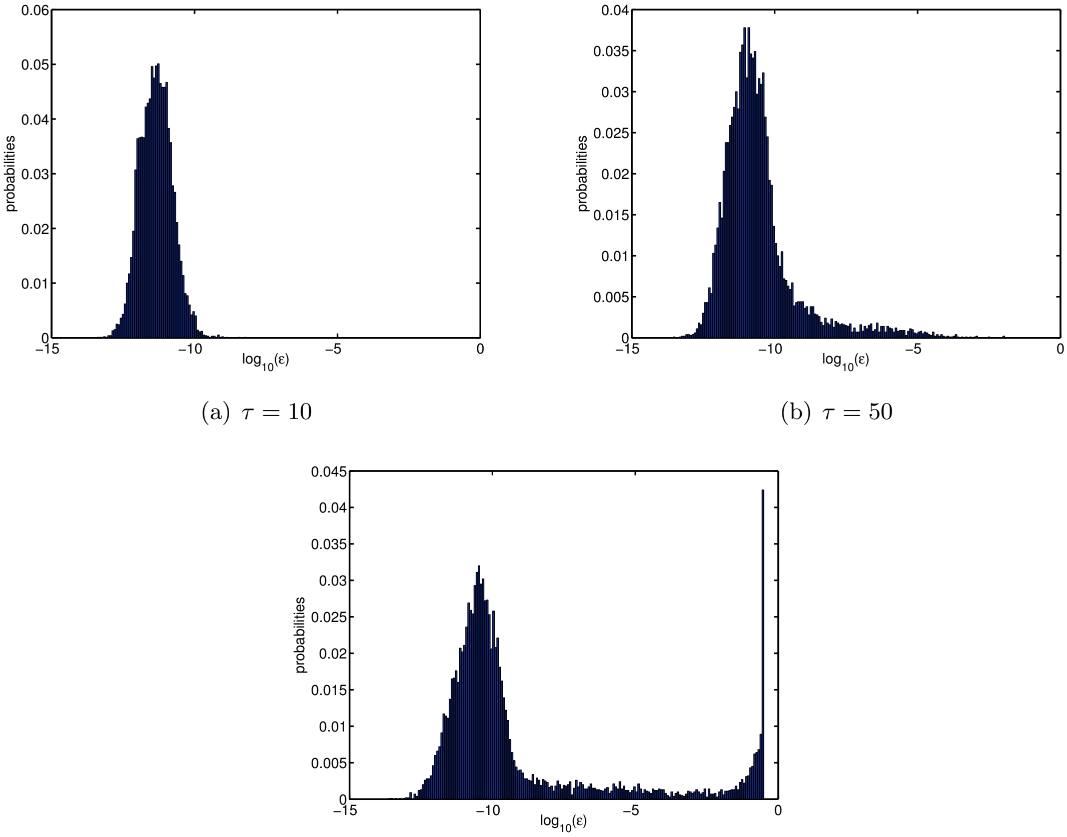


Figure 4.31: Testing the mixture hypothesis at $K = 2$, for $N = 10^4$. α computed with $\tau_c = 1000$ and $\epsilon_c = 10^{-8}$. The straight line is the information codimension under the (initial) assumption of constant information dimensions of the joint components. Modified Γ_m is the outcome of the mixture hypothesis altered to allow varied $D_{c/r}$. Γ is an average taken over three runs, such that each measured value is obtained from sampling the joint as a whole without differentiating between trajectory types.



(a) $\tau = 10$

(b) $\tau = 50$

(c) $\tau = 200$

Figure 4.32: Histogram of distance between τ^{th} iterates initially separated by 10^{-12}. $N = 10000$ points were considered; standard map parameter $K = 0.9$. It is arguable whether trajectories can be split into two types depending on the rates of divergence.

computationally reasonable *and* does not allow any regular trajectories to be mistaken for chaotic ones. Plotting the same graphs of α vs the cutoff parameters, the measured α is seen to plateau, but slowly, without reaching the asymptotic value even in the relatively large τ_c range.

Thus the problem of slow convergence that we saw happen with PMI in the section above translates directly to the problem of computing the weight of chaotic component. In the next chapter we will see that it is exactly the elements making up these marginal pdfs that can be manipulated to give the PMI value. Moreover, there is a possibility that it would not be computationally solvable at all, since typical intermittency times could be larger than the time when trajectories will start to diverge because of error in the finite precision method (although the computational standard map is observed to preserve some features expected of theoretical system, this is not guaranteed to happen for arbitrarily large number of iterations).

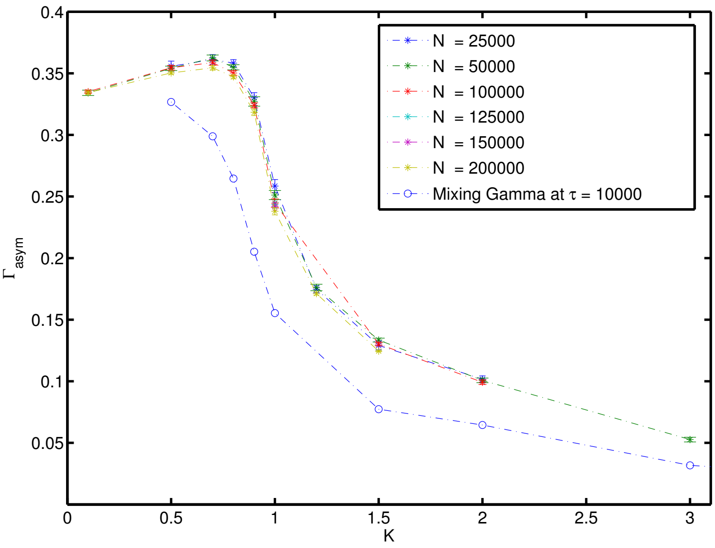


Figure 4.33: The same as figure 4.18(a) but with the mixing Γ obtained with $D_c = 4$ and $D_r = 3$, and the proportions specified by α, where the latter is computed at arbitrary cutoff parameters constant across K.

Thus there are regions, primarily coincidental with regions where convergence of PMI is itself problematic, where the mixture hypothesis is not testable at least by this method and with this categorization of trajectories. We can nevertheless set arbitrary cutoff parameters and compute Γ_m under the assumption that it will be with a large error.

Results are shown in figure 4.33. The first interesting point is that unlike that direct from the PMI, this Γ_m does not peak. The absence of a peak is not merely computation, since we assume that the fraction of chaotic trajectories increases monotonically. Thus the error between the measured Γ and the PMI comprises of a) the over/under-estimation of Γ_m, though the same effects should be responsible for the overestimation of Γ, and b)the difference between considering orbits in isolation from others of different types.

While the first is related to dynamics, the second is to do with the relative spatial arrangements of regions with regular and chaotic motion. The work on the mixture hypothesis suggests that it is the latter that plays an integral part in the change of predictability with resolution. It raises Γ, making the system more predictable for the same price of increased resolution than it would have been had the regular and chaotic regions not been separated. Thus the mixture hypothesis was only qualitatively successful. It showed that for any finite K, there will always be areas where the regular and chaotic trajectories are arbitrarily closer to one another (at least on the range of scales tested by the estimator).

Chapter 5

PMI and Information Codimension from Trajectory Separations Statistics

We have already seen that distributions of evolved distances between pairs of points initially close together, that started out scattered randomly across the state space, can reveal such features of the map as existence of two types of orbits, clearly different in character, which we class as chaotic and regular. In the previous section this setup was used to identify the proportion of regular component as a function of map parameter K. Here we ask the question of whether this framework can be used to predict and/or explain such features of Γ as the possible transience of the peak and its position. The far reaching aim is to find aspects of this picture that directly result in the fractal scaling of PMI, which would then pave the way for forming strong conclusions about existence of such scaling at otherwise computationally-inaccessible limits.

The reason the trajectory separation framework could offer insights into limiting behaviour is twofold. First it enables us to obtain statistical data on much smaller scales than the effective distances one works with when sampling the initial distribution with N points (the average initial separation $\epsilon_0 \approx \frac{1}{2\sqrt{N}}$. So for computationally large sample sizes of $N = 500000$, $\epsilon_0 \approx 10^{-3}$.) Since some of the unresolved PMI issues concern the limit of large resolution this method offers an advantage, given that even allowing for the double-precision version of the map, initial separation could be set as low as machine-epsilon which for doubles is $\approx 10^{-16}$, many orders of magnitude less.

The second advantage is the decoupling of sample size from probability resolution. Measurement of PMI at low resolution could be skewed due to a worse estimator convergence. We first show that tracing trajectory separations in time does yield an algorithm for computing the joint entropy, and hence PMI, for area-preserving systems. We then interpret the joint information dimension using the variables implicit in this framework, and use examples to clarify the interrelation of information dimension with time and sample size.

5.1 Methodology

For area-preserving maps with a normalised state space Persistent Mutual Information at τ is obtained from the K-G entropy estimator:

$$I(\tau) = -\hat{H}_J = -\Psi(N) + \Psi(k) - \frac{4}{N}\sum_{j=1}^{N}\log(2\epsilon_j), \tag{5.1}$$

where ϵ_j is the distance from j^{th} point to its k^{th} nearest neighbour in the joint space. Here N is the number of points considered, or the sample size. The sum can be written as an expectation value w.r.t. some (whose existence can perhaps be only approximated) distribution ρ of the random variable ϵ:

$$I(\tau) = -\Psi(N) + \Psi(k) - 4\mathbb{E}\left[\log(2\epsilon)\right]_{\rho}. \tag{5.2}$$

The maximum metric would pick for ϵ the largest of the initial and final distances. PMI can thus be thought of as a statistical description of the interpoint distances in the joint space.

In this section we review the traditional way of sampling ρ, and introduce a new method. The traditional method iterates the sample itself; part of the computational effort is spent in creating the evolved, future set of distances. The main procedural emphasis is on then combining the initial and evolved samples to create and order the set of distances in the joint.

On the other hand, the method here labelled TS (for trajectory separation) manipulates two sets of marginal distances. Here the emphasis shifts away from ordering the joint points and onto a procedure that combines these sets in a specific manner. The new method compensates for the more complicated procedure by recognising that it requires only a finite number of elements (for each point in the sample) from the second set to complete it. Thus the TS method begins to become advantageous in terms of running time the moment this latter number of elements can be made small enough.

After mentioning the traditional method, we demonstrate that there exists a finite, deterministic procedure for obtaining the joint interneighbour distance from families of initial and evolved distances. We then conjecture that sampling the distance sets and the associated variables *independently* is equivalent to sampling ρ.

Let S be the state space of a continuous, area-preserving dynamical system, and $\mathbb{P}$ a set of probability distributions corresponding to measures defined over some σ-algebra on S. We sample the initial distribution with $X_i \sim \rho_{\text{init}} \in \mathbb{P}$ to produce $X = X_i \in S \ : i = 1..N$.

Traditional Method Consider the calculations involved in finding the mean interpoint distance in the joint as done by the tradition PMI methods in the previous chapter. The set X gets evolved under F for τ times (w.l.o.g. assume F is an iterated map), to Y. To each trajectory indexed by i we associate a family $D_i(X,Y)$ of first, second, .. k^{th} nearest neighbour distances in the joint space referenced by two time elements, 0 and τ: $D_i(X,Y) = (D_i(k,X,Y))_{k=1}^{N-1}$, where

$$D_i(k=1,X,Y) = \min_{j\neq i} \left[d((X_i,Y_i),(X_j,Y_j)) \right]. \tag{5.3}$$

$\epsilon_i(X)$ is then just $D_i(k=1,X,Y)$.

5.1.1 New Method

Consider instead associating a nearest neighbour distances family $D_i(X)$ to each point i in the original sample X. Let $D(X)$ be the set of all $D_i(X)$.

Trajectory separation method traces the evolution of distances between specific points. The outcome is held in the ordered set $D^\tau(X)$ of families $D_i^\tau(X)$, where each k^{th} element is the new distance between the i^{th} trajectory from X and the trajectory that was its k^{th} nearest neighbour in the past,

$$D_i^\tau(k,X) = d(F^\tau X_i, F^\tau X_j) \, s.t. \, d(X_i,X_j) = D_i(k,X) \tag{5.4}$$

We now show that for each i there exists a finite, local algorithm (which we call procedure P) to compute the i^{th} minimum joint interneighbour distance ϵ_i from the families $D_i(X)$ and $D_i^\tau(X)$.

Procedure to compute ϵ_i Consider N trajectories p_i, $i \in I = [1,2,..N]$. Let $d_{ij}^t = d(p_i(t),p_j(t))$, where d is the maximum metric. The joint distance between two trajectories is then $d_{ij}^J = \max(d_{ij}^0, d_{ij}^t)$.

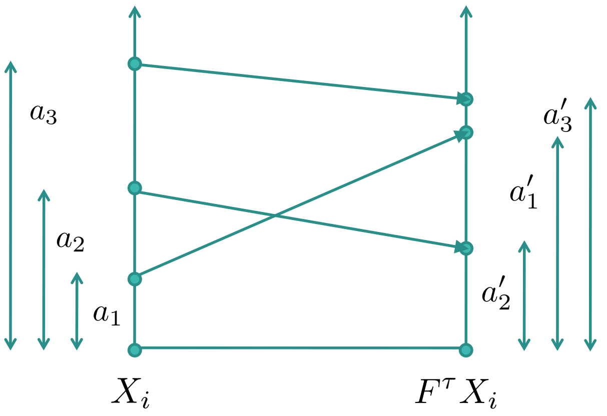


Figure 5.1: Procedure to compute joint interneighbour distance. Here a stand for distances to nearest neighbours from point i.

For trajectory i we label other trajectories based on the proximity to i at time $t = 0$. This creates a set I_i indexing the elements of D_i, a family of nondecreasing distances between trajectory i and others:

$$D_i = \left(d^0_{ik}\right)_{k \in I_i}. \tag{5.5}$$

Let a_k be the k^{th} shortest distance between trajectory i and some other, $a_k = D_i(k)$. Let $a^{'}_k$ be the distance between respective trajectories at a future time τ, $a^{'}_k = d^\tau_{ik}$. We look for the trajectory that is p_i's nearest neighbour in the joint space by considering successively larger neighbourhood in the past. Figure 5.1 demonstrates the general principle: the joint distance between trajectories i and k is

$$d^J_{ik} = \max(a_k, a^{'}_k). \tag{5.6}$$

The distance between trajectory i and its nearest neighbour in the joint space, d^J_i, is realised by such trajectory $j^* \in I_i$ that

$$d^J_{ij^*} \leq a_{j^*+1}, \tag{5.7}$$

with

$$d^J_i = \min_{j^*} d^J_{ij^*}. \tag{5.8}$$

Because candidate distances are bound by interpoint distances in the past (at time τ), it is straightforward to construct a search by considering marginal nearest neighbours of increasing index, computing distances between the respective trajectories in the future time, and each step checking for completion. The algorithm is a simple update of *candidates* (joint distances) and *cutoffs* (marginals distances) :

```
nnmax = trialM; //sets estimated limit within which to look amongst the nearest neighbours in the past for the point forming the closest connection in the joint space

past_nnDistances = getnnDistancesFromKDtree(thispoint, nnmax);

int nncount = 0; \\index of a from the graph

cutoff = past_nnDistances( nncount ); \\a_1
candidate = evolvedistance( thispoint, cutoff, map ); \\a'_1 = max(a_1, a'_1)

while (candidate > cutoff)
  nncount++;

  cutoff = past_nnDistances( nncount );
  newcandidate = evolvedistance( thispoint, cutoff, map );

  candidate = min(candidate, newcandididate);

  if( nncount == nnmax ) .. //reset trialM to a larger value and repeat process

end

epsilon = candidate;
```

5.1.2 Sampling

Practical computation of PMI entails taking averages over samples. This allows us to attribute meaning to ρ.

The K-G estimator is unbiased because rather than being multiplicative it involves a sum of terms. The fact that the terms, which are the joint interneighbour distances, are not independent of each other (since need to have N points distributed uniformly in state space) means the errors will not be independent.

Hence we get an unbiased estimate of $\log \rho$ (for nearest neighbour index k of one) by sampling *local* marginal interpoint densities, localising them randomly and evolving the respective points, and then applying the deterministic argument above to both the initial and the evolved counterparts. Because we use independent sampling, it is possible that singular events will not be seen. In this respect we to some extent assume that here smooth (in the sense of non-singular) marginal interpoint distributions would give a smooth joint interpoint distribution. The marginal interpoint distances for two random points would likely not be independent. Hence the lower the sampling depth the more independent our distances become.

Localisation is an interesting problem. D_i^τ cannot be computed from $D_i(X)$ alone; evolved distances are entirely dependent on the *position* of the initial separation vector, not just its length. To obtain D^τ we also require the *absolute*, not just the relative, position of the set of separation vectors.

Consider the combination procedure. Each evolved separation family $D^\tau(i,k,X)$ is a result of a deterministic function that depends only on the initial location of the k^{th} trajectory pair:

$$D^\tau(i,k,X) = f\left(D(i,k,X), X_i, L_k, F, \tau\right), \tag{5.9}$$

where L_k is the information about the arrangement of the separation vector for point i, and X_i is the position of point i. Consider a random variable $V = (D, L, X_{\text{pos}})$, where the variates inside the brackets stand for $D_i(X)$, L_k and X_i. The above equation states that there is a function of V that results in a value identified with $D^\tau(i,k,X)$. A further procedure (all deterministic) then gives the inter-neighbour ($k=1$) distance associated with that point i. This constitutes the TS method - we sample V and use a specific algorithm to get ϵ. Taking averages with respect to the sampled V should give the correct averaged $\log \epsilon$.

The main issue here is of course that the v values come in specific configurations, whereas we assume independent sampling of V. Moreover, we ignore the specific interdependencies of elements of V. Specifically, we orient ourselves to sample the correct *marginals* of its elements. We propose:

- Sampling M random points. Thus elements of X_{pos} would be distributed with ρ_{init}.
- The marginal of the location of separation vectors is by symmetry an equidistribution. If point pairs are defined by the lower/left-most point, sampling from its marginal involves (recalling that the metric is a maximal one) picking randomly the axis where separation is some given number a, and creating the second point higher/to the right, shifted by $b \sim \mathbb{U}[0, a]$.

- Finally, we sample from the marginal of the 'past distances' vector D by drawing individual D_i from some table (effectively computing them from a kdTree)- that way relative sizes of first, second, .. k^{th} distances are preserved.

A natural extension of the method would be to sample D_i theoretically. The sampling technique itself is not straightforward, since the probability of having the k^{th} nearest neighbour at a certain distance would be dependent on the obtained values for the previous $k-1$ ones. However, in practice retrieval of the 'past' distances, especially if a kdTree method is available, is not a computational burden, especially if M is low. Jumping ahead, for the average (N, τ) used in the section above the highest k index is of order hundreds. Considering that M does not need to be much higher than that to achieve correspondence between TS PMI and the true value, the factor that contributes most towards running time is τ.

Errors would stem from how likely we are to miss something singular in terms of dependencies. However, we claim that in practice the typical dependency is only on close points.

5.2 PMI Scaling: Information Codimension from Trajectory Separations

The key step now is to understand how to use trajectory separation distributions to say something about statistics of nearest neighbour distances in the joint space. Specifically, linear PMI scaling with resolution can be understood from the trajectory separation picture. Given two sample sizes, N and AN, $A > 1$, and some k,

$$\Gamma = \frac{-\Psi(AN) + \Psi(N) - \frac{4}{AN}\sum_{j=1}^{AN}\log(2\epsilon_j') + \frac{4}{N}\sum_{j=1}^{N}\log(2\epsilon_j)}{\Psi(AN) - \Psi(N)}, \tag{5.10}$$

and hence, rewriting the above in terms of the mean values and canceling the log(2), we get

$$\mathbb{E}\left[\log(\epsilon)\right]_{\rho_{AN}(\epsilon)} \approx \mathbb{E}\left[\log(\epsilon)\right]_{\rho_N(\epsilon)} - \frac{\Gamma+1}{4}\log(A), \tag{5.11}$$

5.2.1 Fully-integrable case of $K = 0$, $\Gamma = 1$ limit

From (5.11), when $\Gamma = 1$, for large sample sizes N,

$$\mathbb{E}\left[\log(\epsilon)\right]_{\rho_{AN}(\epsilon)} \approx \mathbb{E}\left[\log(\epsilon)\right]_{\rho_N(\epsilon)} - \frac{1}{2}\log(A), \tag{5.12}$$

i.e. the mean distance to k^{th} nearest neighbour in joint space associated with sample size AN will be smaller by $\frac{1}{2}\log(A)$ than the respective mean distance associated with sample size N. This is reasonable since a larger sample size means smaller initial separation, so we expect smaller interpoint distances in general. We are now in a position to show that decreasing the sample size by a factor of A shifts the expected value of $x = \log\epsilon$ by $\frac{1}{2}\log A$.

We are going to argue that for $K = 0$ and $\Gamma = 1$,

$$\mathbb{E}\left[\log\epsilon\right]_{\rho_N} \approx \mathbb{E}\left[\log\epsilon^0\right]_{\rho_N} + f(\tau), \tag{5.13}$$

where f is manifestly not a function of N. If (5.13) is true, then since marginalising ρ_N gives the mean distance to k^{th} nearest neighbour in the past to be $\frac{1}{2}\sqrt{(k/N)}$, eq. (5.12) reduces to

$$\log\left(1/2\sqrt{\frac{k}{AN}}\right) \approx \log\left(1/2\sqrt{\frac{k}{N}}\right) - \frac{1}{2}\log(A), \tag{5.14}$$

where the LHS is equal to the RHS since the two N-independent f functions cancel from both sides.

Consider computing ϵ from trajectory separation. Using the new terminology this corresponds to

$$\epsilon = f^{\mathrm{det}}\left(D(k, X), X, L, F, \tau\right) \tag{5.15}$$

where f^{det} is some procedure. Implicit in that procedure is the computation, at least partial, of the respective evolved distances $D^{\tau}(k, X)$.

We first rewrite the above equation in terms of a shift G applied to an arbitrary function of $D(k, X)$, which we choose to be the initial interpoint distance $D(k = 1)$, which would render G necessarily positive, due to the maximum metric (we also rewrite in log basis):

$$\log \epsilon = \log D(0, X) + G\left(\log D(k, X), \log D^{\tau}(k, X)\right). \tag{5.16}$$

The important thing to notice is that the shift only concerns the relative values, i.e. distances between points in the past and in the future. It does not take into account localisation of the separation vectors (if manipulations of finite separation vectors is how we choose to visualise the process). The four other variables contributed towards creating the respective future separations vector (similar to eq. (5.9) from the previous section)

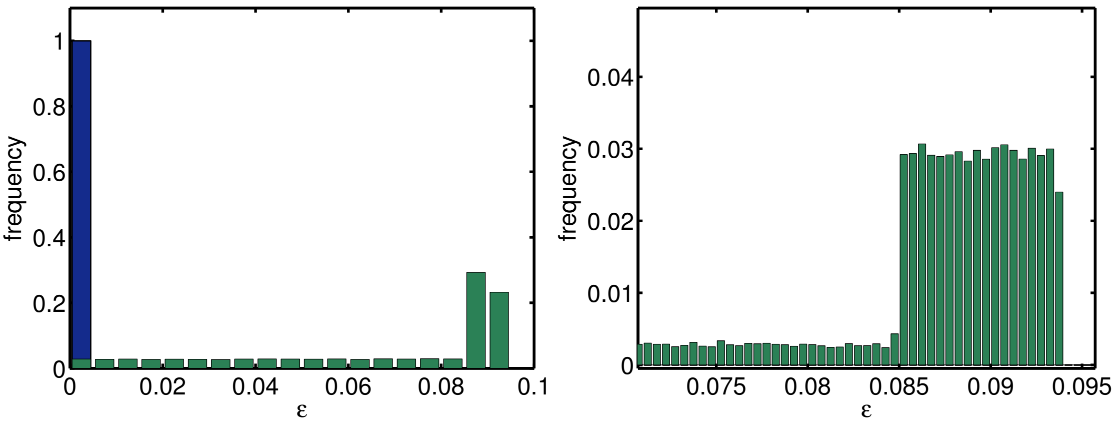
$$D^{\tau}(k, X) = f^{\mathrm{det}}\left(D(k, X), X, L_k, F, \tau\right), \tag{5.17}$$

where f^{det} is some (different) procedure.

In other words, the interneighbour distance (for some point) in the joint space consists of starting with the nearest separation in the past, noting the vector of nearest neighbour distances from that point, evolving it, and manipulating the two resulting vectors to increase the starting distance by a certain amount. We now argue that for $K = 0$ an evolved interpoint distance does not depend on the location of initial points.

Let $K = 0$, $\tau = 20$, and $N = N_1 = 50000$. From fig. 4.9, this sample size gives the required $\Gamma = 1$. We look for the first nearest neighbour, and initialise $M = N_1$ trajectory pairs separated by $\epsilon_1(N_1) = -\frac{1}{2}\log_{10}(50000) \approx 0.004$ and evolve for $t = \tau$.

Figure 5.2(a) shows the distribution of distances between trajectory pairs that were initially ϵ_1 away. Aided by figure 5.2(b) we identify two subsets on which ρ_t is relatively



(a) ρ_0 and ρ_{20}

(b) Zoom in on ρ_{20}

Figure 5.2: Frequency count corresponding to ρ_t, the distribution of trajectory separations for $t = 0$ (blue peak at ϵ_1) and $t = 20$ (in green). Maximum metric is used throughout.

flat, but varies dramatically in magnitude.

This is a consequence of the metric (maximum) used to compute (and define) separation of points in the standard map state space. We initialise a pair of points separated by ϵ_0 through 1) identifying the variable (θ or momentum) that will be $\pm\epsilon_0$ away, and then 2) picking the remaining coordinate of the second point to lie a distance $\pm\delta$, $\delta \geq \epsilon_0$ of the first one (uniform distribution). Since the designation of first and second point is arbitrary we use a positive ϵ_0 and a positive/negative δ. For $K = 0$ the mapping is a translation of θ by the amount corresponding to momentum. So if the initial separation as the maximum of the θ and momentum distances is realised by momentum, then the t^{th} iterate would give the distance between two points as

$$\epsilon_t = \delta + t\epsilon_0, \tag{5.18}$$

and those pairs separated through θ will have

$$\epsilon_t = \epsilon_0 + t\delta. \tag{5.19}$$

The momentum-separated trajectories, which as expected form roughly half of the total, would then give a flat $\rho_t = \rho_1$ centered on $t\epsilon_0$, of width $2\epsilon_0$. This is indeed what we see in figure 5.2, where the middle of the high step happens at $20 * 0.0045 \approx 0.09$, and is 0.009 wide. The wider lower region of ρ_t should be $\rho_2 = 2a$ high between ϵ_0 and $t\epsilon_0$, $\rho_2/2$ high between 0 and ϵ_0 and $\rho_2/2+\rho_1$ high $t\epsilon_0$ and $t\epsilon_0+\epsilon_0$. Since $1/2 = 2\epsilon_0 * \rho_1$, $\rho_1 = 1/(4\epsilon_0)$.

Both ϵ_0 and δ are of similar orders of magnitude. Hence as t increases, $\epsilon_t \rightarrow \propto t\epsilon_0$. Since the implication of the final distance not depending on the location of the initial point is also that any k^{th} neighbour distance evolves independently of the initial location, then this is true for whole family:

$$\log D^{\tau}(k) \approx \log D(k) + \log(\tau). \tag{5.20}$$

Thus eq. (5.16) can be rewritten as some

$$\log \epsilon = \log D(0, X) + G' \left(\log D(k, X), \tau \right), \tag{5.21}$$

The shift $G^{'}$, which is a procedure of combining elements of $\log D$ and $\log D^{\tau}$ to produce ϵ, does not depend on the absolute values of the elements. The key point here is that only the relative configurations of distances determine the location of the value taken to be ϵ. We know that in the limit of large N the logarithmic positions of elements of D indexed by k are, on average, $\frac{1}{2}\log(k) - \frac{1}{2}\log(N)$. Therefore the difference between family elements does not depend on the sample size. There can be errors when small N samples do not follow Poissonian statistics, but it will hold in the limit of large N (see figure 5.4).

The RHS of the equation above becomes split between parts dependent respectively only on N and τ, which allows generalisations to be made. The distribution of ϵ can be sampled with the appropriate distribution of $D(0)$, which *does* depend on N. Since G' is independent of N we associate it with f introduced in eq. (5.13), which thus holds. Thus we see that using the TS framework we can recover the infinite-resolution $\Gamma = 1$.

5.2.2 Transition to $\Gamma = 1/3$ in the Fully-integrable case of $K = 0$

From empirical observations we infer that for any sample size N there exists a time τ_{min} such that for any $\tau > \tau_{min}$ Γ will be $\frac{1}{3}$. Equivalently, given a time t we conjecture that there exists a sample size N_{max} such that Γ will be $\frac{1}{3}$ for any $N < N_{max}$. The intuitive explanation is that this is due to wrapping effects which to some extent destroy initial correlations (but not completely, since the map is not chaotic). This occurs at a time that is dependent on the initial separation.

We also infer that there exists a time below which the information codimension is equal to

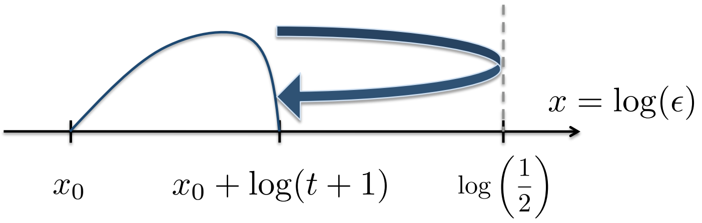


Figure 5.3: Reasoning behind the wrapping time/sample size relation (here $t = \tau$). The curve in blue is the evolved distribution of distances between points that were closest neighbours at $t = 0$, with distance x_0. The combination procedure G' will take into account evolved distances up to, at most, $x_0 + \log(t + 1)$. A reflected evolved distance may begin to participate if it starts 'under' the curve and has time to come back. Maximum metric ensures that G' cannot consider evolved distances that started to the right of the curve.

unity, and that this time also depends on the sample size. Here we attempt to infer the interrelation between the two.

As before, we start with the distribution of initial points, which is associated with a family of $k = 1, 2...$ nearest neighbours distances in the state space. The evolution rule and number of iterates τ gives rise to another family, this time of distances between iterated subjects and evolved points that were k^{th} nearest neighbours in the past. As the underlying distances increase, the support set of the second family moves to the right towards the reflective $\epsilon = 1/2$ boundary.

In the trajectory separation framework wrapping happens when the relevant future separation distributions hit and get reflected off the $\epsilon = 1/2$ boundary. After some settling period the support space becomes $[0, 1/2]$, with the momentum-separated trajectory pairs still giving a peak. That peak, however, is bounded by $[\epsilon_0, 1/2]$. The direction of its travel is a simple function of the evenness of the remainder of $\tau\epsilon_0$ and $1/2$.

Consider an initial position $x_0 = -\frac{1}{2}\log N$. After time τ the peak will be at $x_\tau = x_0 + \log(\tau + 1)$. The support S of the distribution of nearest neighbour distances in the joint space will be a subset of $\left[x_0, \min\left(x_\tau, \log\left(\frac{1}{2}\right)\right)\right]$. When either τ is large enough, or N is small enough, some of the weight resulting from evolved distributions of separations whose initial value was in S would have been in S after having been reflected off the $\log\left(\frac{1}{2}\right)$ boundary. The peaks originating in the upper most limit of S will re-enter S first. Therefore a lower limit of t below which no re-entry is possible will be given by the time for which the final position of the 'most likely' wrapping candidate is outside the right hand limit of S, i.e.

$$x_\tau + \log(\tau+1) < \log\left(\frac{1}{2}\right) + \left(\log\left(\frac{1}{2}\right) - x_\tau\right), \tag{5.22}$$

or

$$2x_0 < \log\left(\frac{1}{2}\right) + 3\log(\tau+1). \tag{5.23}$$

which is equivalent to

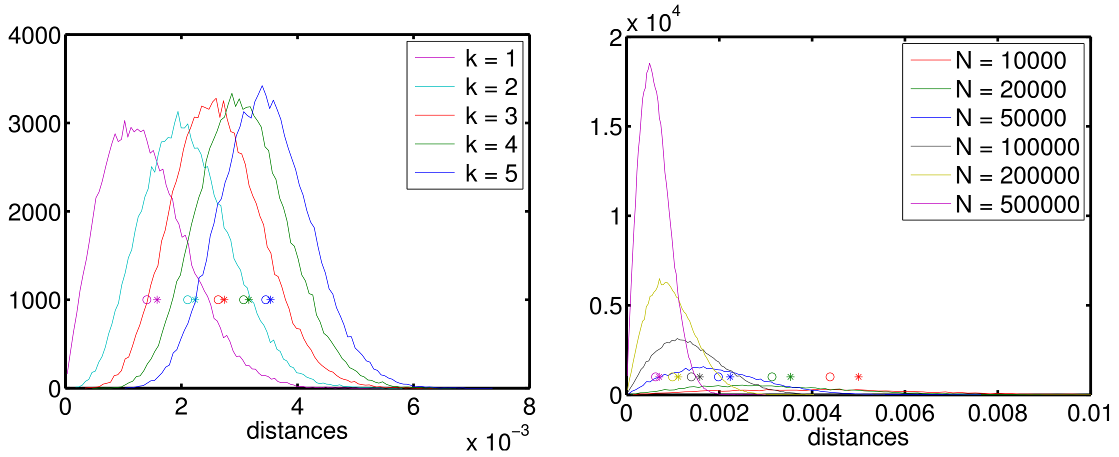
$$\frac{(\tau+1)^3}{N} < \frac{1}{2}. \tag{5.24}$$

In other words, there comes a point (N, τ) when

$$\log D^\tau(k) \approx \log D(k) + \log(\tau) \tag{5.25}$$

no longer holds. When, in addition to that, the element of the $D(k)$ family that does not get evolved according to this rule could potentially be an input into G', the combination procedure, then the $\Gamma = 1$ scaling will break down.

Eq. (5.24) states that periodic boundary conditions could only begin to affect the joint interneighbour distances, and hence the PMI, when $\tau^3 \geq N$. This is indeed the scaling at which plots of Γ for different N collapse, as observed in the previous section. Moreover, it provides a cutoff point that can be confirmed through figure 4.9. There Γ begins to decrease from its plateau of unity when $\tau^3/N \approx 0.3$, close to $\frac{1}{2}$. Correction of τ by one, though stemming from the maximal metric as well as an exercise when initial interneighbour distances are all the same, does bring the cutoff point closer to the one observed. In the same figure 4.9, but taking a particular sample size $N = 150000$ as an example, Γ begins to fall when $\tau^3/N \approx 0.43$, whereas this value is 0.46 for $(\tau+1)^3/N$. Hence we see that TS logic is useful for deriving the mixing properties, at least for $K = 0$.



(a) Varying k, $N = 50000$

(b) Varying N, $k = 1$

Figure 5.4: Unnormalised distribution of distances to $k^{Th.}$ nearest neighbour, given N points sampling a unit square. Horisontal positions of circles correspond to the mean values, whereas stars give $\frac{\sqrt{k}}{2\sqrt{N}}$, the mean w.r.t. the expected Poissonian statistics.

5.3 Implementation

Initial interpoint Distances As N increases both the support, and the shape, of these distributions changes (in a qualitatively different way than would happen by sampling a different set of N points). The deviation from normality is a result of the interdependence of its components. Neither do the means correspond to the expected $\frac{\sqrt{k}}{2\sqrt{N}}$, though the fact that actual values are smaller is at least partially due to the maximum metric used to compute them.

The statistics we require is a set of M interneighbour distances in the joint space. Each such value is obtained from some point $x \in X$, and a set of distances to its k' nearest neighbours. The specific k' depends on the map dynamic. We start with an array of distances to nnmax closest points that we obtain by building a kdTree, picking a random point on it and retrieving the distances. The substance of the trajectory separation method is, however, in evolving these distances themselves, i.e. by assigning them to pairs localised somewhere in the state space. There are several ways this can be done, and in our implementation we distinguish three of these.

- Method 1. Out of the three this method resembles the original, traditional construction most closely. Here the spatial location of the point pairs that represent the distances is such that one the points corresponds to the entry in the kdTree that was used to retrieve the distances. The second of the pair is picked at random, still using maximum metric.

- Method 2. In this method the point on which all the pairwise distances are centered does not correspond to any particular point of a KDtree, but instead is sampled from an flat distribution over the state space.

- Method 3. Here every first point of a pair is picked from a random distribution over the state space. This means that the shortest distance could be between points located in a chaotic region, and the second shortest distance could be between points somewhere else entirely. The value of ϵ obtained this way no longer reflects the action of the map on some neighbourhood; instead it is in some sense already averaged across the state space.

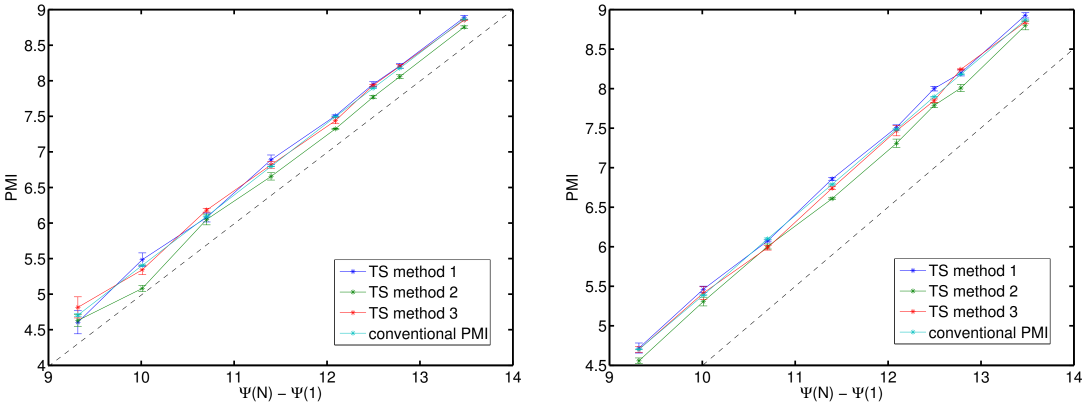
5.3.1 Results

Results could potentially differ based on the method used, the effective map parameters K and τ, the sampling strength M (sometimes expressed here either as percentage of true sample size), the resolution N, and the function in question, since some consistent errors in PMI do not necessarily imply a false Γ.

We first consider the non-chaotic case of $K = 0$, and examine two regimes, first ones where PMI displays well-defined scaling of $\Gamma = 1$ and $\Gamma = 1/3$, and then look at the transition. The far-reaching motivation is to see if there is a link between the validity of the assumptions behind TS PMI (specifically, independence of components of V), and Γ, whether in its existence or in the quantitate sense.

For each of the three TS methods PMI is compared with the value obtained using the traditional (here called “conventional” method), at $k = 1$. This value was chosen since in the section above the least ambiguous Γ was one defined as the gradient with respect to a varying N and $k = 1$.

$K = 0$. In the first graph 5.5(a) the TS PMI for each resolution N is computed with sampling depth $M = 0.01N$; in 5.5(b) this value is kept at $M = 1000$. As a result, the first four values in 5.5(a) for TS PMI are computed with much less data, and are hence



(a) $M = 0.01N$

(b) $M = 1000$

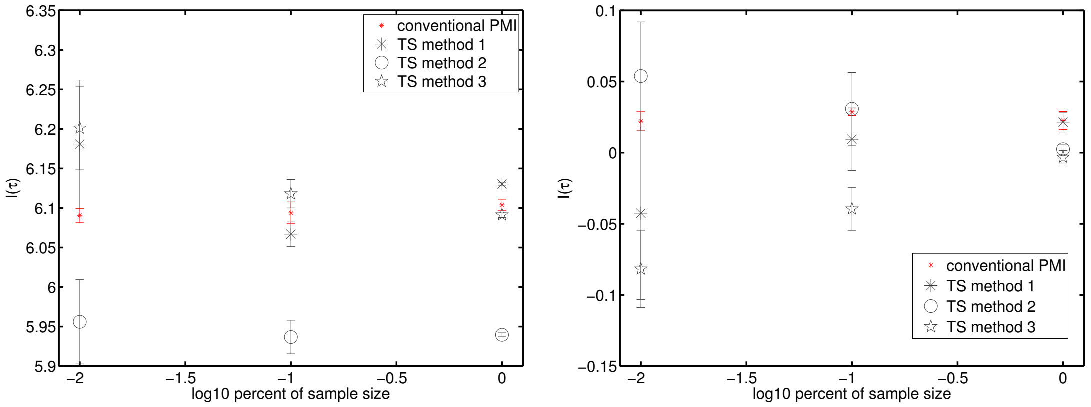
Figure 5.5: $K = 0$, $\Gamma = 1$ regime. PMI v resolution at $\tau = 10$ using three trajectory separation methods (sampling initial distances), and the traditional method ($k = 1$), averaged over three runs. The dotted line indicates the slope of unity.

much more prone to statistical errors, than the respective entries in 5.5(b) (the first value is $M = 63$). Other than that, we see that here TS PMI picks up the correct slope of $\Gamma = 1$, and that the actual PMI values are roughly in line with ones computed by the traditional method.

There are, however, some errors, and a bias towards lowering PMI is apparent in 5.5(b) for method 2. If these are errors stemming not from some inherent bias, they should go away with either more runs, or with higher M. We therefore check whether varying M makes a difference to the PMI values.

Results for the same parameters as above are shown in fig. 5.6(a). First it should be noted that the error bars are computed from three runs, and as such do not give an indication of the actual spread of data. Indeed not only can the relative size of the error bar change if the simulation were run again, the mean values for the trajectory separation methods also display a significant variation. What is evident from several runs is that to each method corresponds some 'true' PMI value that the results fluctuate around to a greater or lesser extent depending on the sampling size (here sampling size M is distinct from sample size N). For the $\Gamma = 1$ the TS PMI for methods 1 and 3 is almost coincident with the true PMI for the traditional method; whereas PMI for method two is lower by a relative error of about 2%. For $\Gamma = 0$ all true values appear to be within an error that in absolute terms is ten times less than the former case.

In other words, results shown in figure 5.5 would not change if a larger M was used; we



(a) No mixing, $K = 0$, $\Gamma = 1$ regime, $\tau = 10$

(b) Fully-developed chaos, $K = 2\pi$, $\Gamma = 0$, $\tau = 100$

Figure 5.6: PMI using three trajectory separation methods (sampling initial distances), and the traditional method, computed at $N = 25000$. Values of the former are plotted as a function of the sampling size used, expressed here as a function of N, and going down to 250.

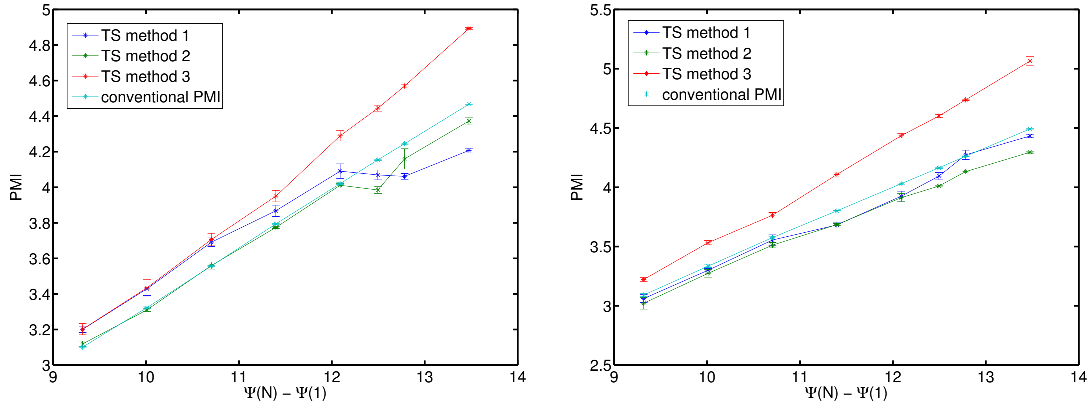
detect a bias in TS PMI when $\Gamma = 1$ (for this $K = 0$ at least). This bias goes away when $\Gamma = 0$, a case of absolute mixing. We thus conclude that for the fully-resolvable case TS PMI does a good job, with a small absolute bias that does not change with N, giving the right Γ.

We conjecture that TS PMI is sufficiently close to true PMI at other (N, τ, K) regimes that correspond to $\Gamma = 1$. We also note that these values are obtainable with sufficiently low errors by a small enough, fixed M.

Now consider another well-defined Γ regime, that of $\Gamma = 1/3$. Figure 5.7 shows PMI v resolution computed at a fixed $M = 1000$, for two different τ values (from the previous section we know that at these parameters $\Gamma = 1/3$, and the plot of Conventional PMI confirms this).

We will see that for small K there is a range of small τ when the methods differ, albeit by a rather small amount, from the traditional PMI. We do not yet have an explanation for these errors. They seem to be smaller when τ is large, perhaps when the system is better settled. Nevertheless in all these cases it would seem that method 1 fares much better than the other two, which is to be expected.

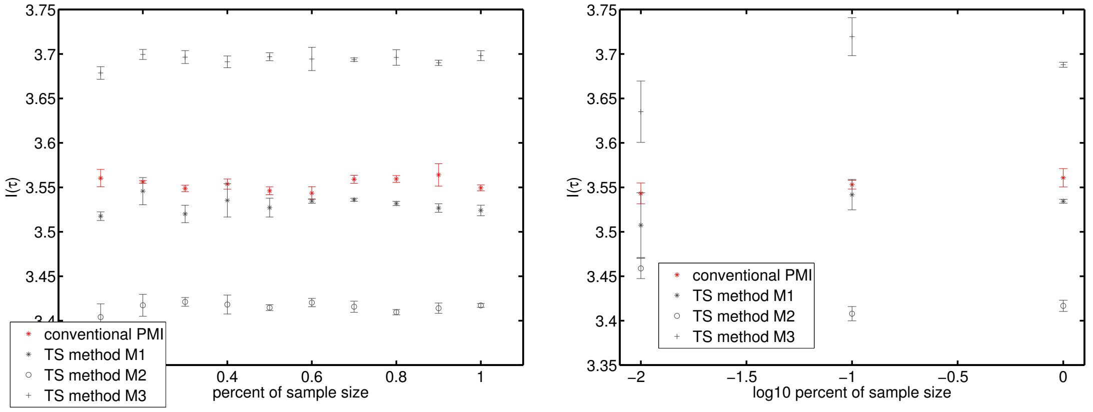
Figure 5.8 shows that the error is indeed a bias, i.e. the error is consistently lower/higher. The smallest bias is in method 1, which by design contained the least uncertainty. We thus conclude that there could be bias that differs at least on Γ; but that for $K = 0$,



(a) $\tau = 200$

(b) $\tau = 1000$

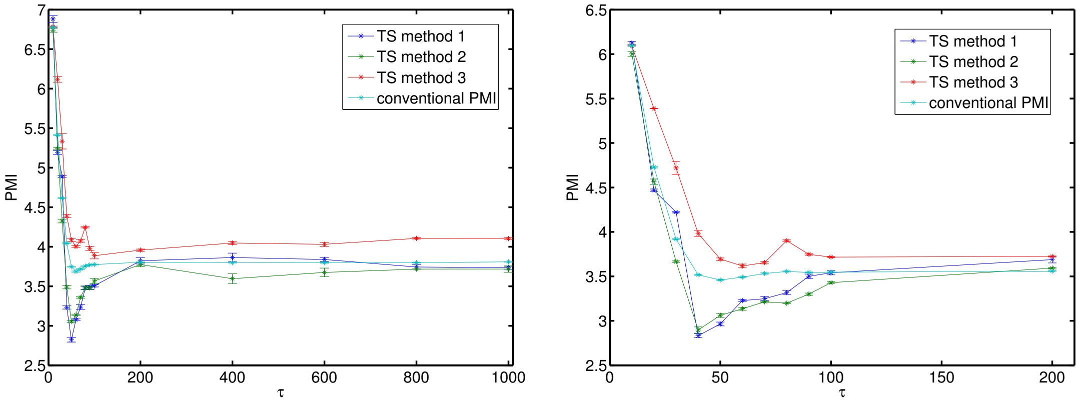
Figure 5.7: $K = 0$, $\Gamma = 1/3$ regime. PMI using three trajectory separation methods with $M = 1000$(sampling initial distances), and the traditional method ($k = 1$), computed v N (averaged over three runs)



(a) Linear Scale

(b) Logarithmic Scale

Figure 5.8: $K = 0$, $\Gamma \approx 1/3$ regime. PMI using three trajectory separation methods (sampling initial distances), and the traditional method, computed at $N = 25000$, $\tau = 100$. Values of the former are plotted as a function of the sampling size used, expressed here as a function of N, and going down to 250.



(a) $N = 50000$, $M = 1000$

(b) $N = 25000$, $M = 0.1N$

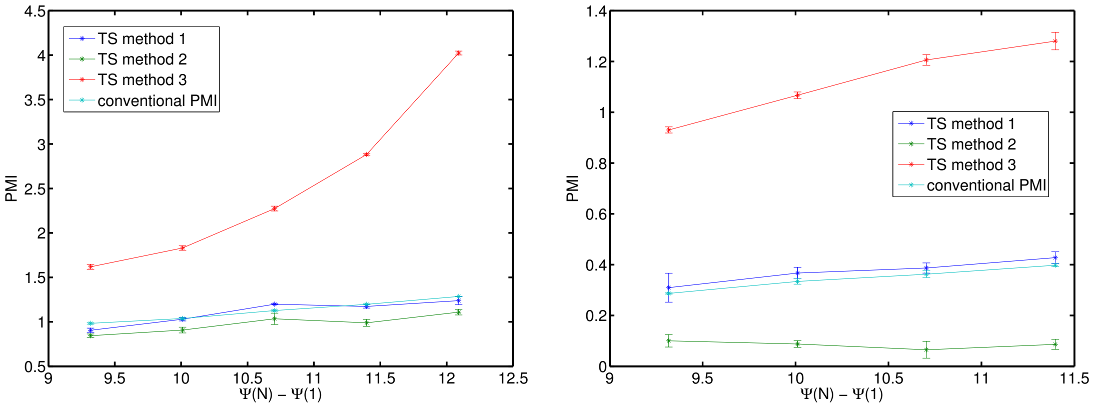
Figure 5.9: $K = 0$, Γ in transit between the two stable limits. PMI using three trajectory separation methods (sampling initial distances), and the traditional method ($k = 1$), computed v τ (averaged over three runs)

$\Gamma = 1$ it is relatively insignificant for all three methods, and that method 1 gives a relatively close PMI. There is more bias for $\Gamma = 1/3$, but it depends on (N, τ) and not on M. We now examine how TS PMI changes as the system makes the transition between the fully-causal and the fully-mixing (no chaos) case.

Figure 5.7 above showed that TS PMI can vary with quite a large bias as τ grows larger, depending on N. These variations correspond to the large τ end of figure 5.9(a), which displays the behaviour of PMI with τ. The former are seen to occur after a significant deviation from the true PMI value that happens when Γ is changing. A zoomed in version on this low τ region is shown in figure 5.9(b). Although N is different there, several sample sizes N and sampling depths M (up to $M = N$) were tested and the qualitative differences are the same, independent of either. Hence for $K = 0$ the TS PMI displays a bias at the Γ transition point. We therefore conclude that from the fully regular case, the assumptions behind TS methods seem to be valid when the system is fully-causal, close enough when the system has settled into the fully-mixing regime, but appear to break down in the state of transition. The next paragraph will test these conclusions across different K regimes.

Large K We now examine two regimes at large $K > K_c$, for which we the asymptotic, N-independent Γ exists. Figure 5.10 shows the variation of PMI with resolution.

It can clearly be seen that TS method 1 performs well, with little to no bias (that does not appear to change with N - neither, from test runs, with M). The performance of



(a) $K = 2$

(b) $K = 4$

Figure 5.10: PMI v resolution for large K. PMI computed using three trajectory separation methods with $M = 1000$ (sampling initial distances), and the traditional method ($k = 1$). $\tau = 1000$, averaged is over three runs.

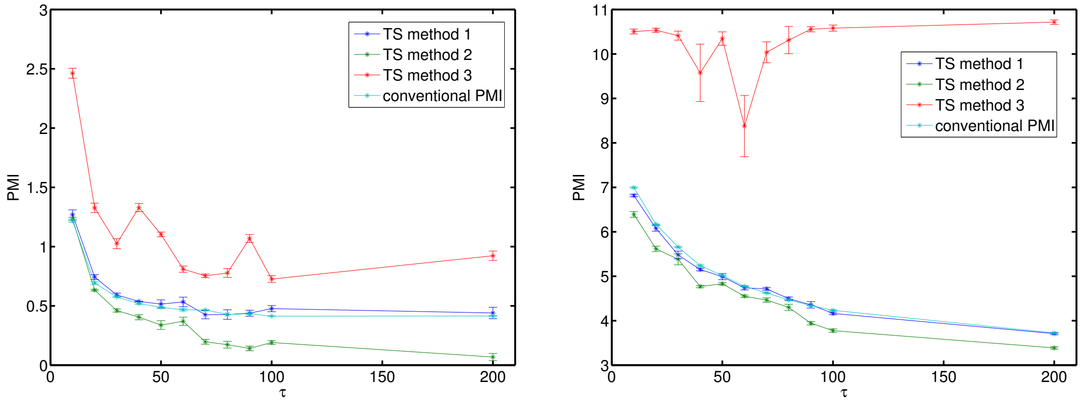
method 2 is variable, and depends on the parameters, whereas method 3 is unreliable. The bais in method 1 does not prevent it from giving the correct slope of Γ.

In figure 5.11 we examine the transition of Γ, and by extension the τ dependency of the TS methods.

Since in the τ interval considered Γ changes dramatically from some value close to unity to one near the asymptote, the figure above actually shows that TS PMI does not necessarily deviate from the true PMI when Γ is in transition. Methods 2 and 3 display significant variation, *not only in the transitive* Γ *state.* Most importantly, the bias in method 1 that was present at $K = 0$, transitive Γ regimes, appears to be absent at high K values. At large τ values not shown on the graph it was observed that method 1 continues to be in line with the true PMI, while methods 2 and 3 do not consistently converge or diverge; but rather that behaviour depends on K. Since there is nothing special about this N, we conjecture that for $K > K_c$ TS PMI using method 1 is close enough to the true PMI, independent of M.

$K < K_c$ regimes Figure 5.12 shows that for a sample low K TS method 1 is still close enough to the true values, giving the correct Γ; method 3 consistently deviates, and method 2 varies in its bias.

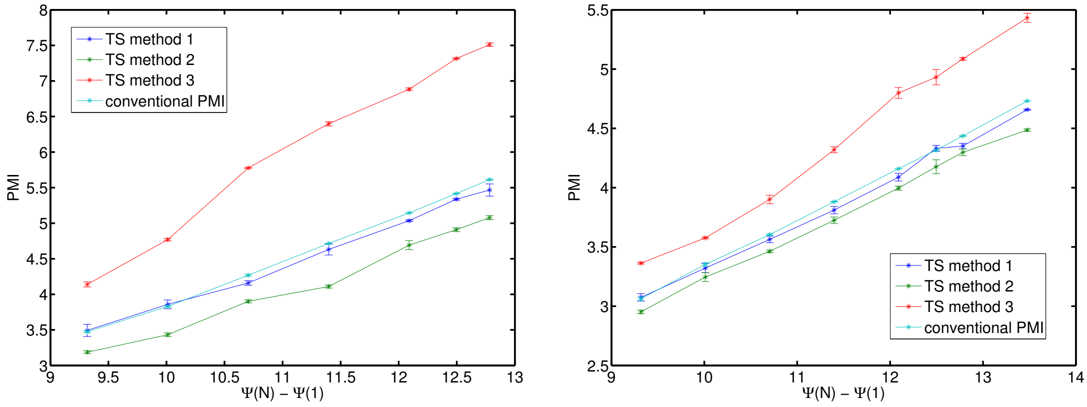
Figure 5.13(a) displays variation of PMI with τ for the same K value. The bias in method 1 does not change with τ, unlike in the $K = 0$ case. Figure 5.13(b) shows that a variable bias does appear as K is lowered, in particular at $K = 0.1$. It is not, however,



(a) $K = 4$

(b) $K = 1$

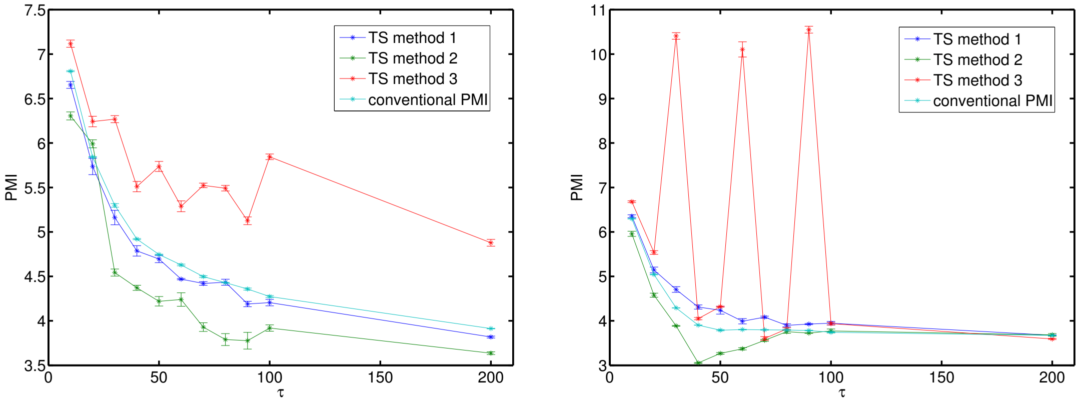
Figure 5.11: PMI v τ for large K. PMI computed using three trajectory separation methods with $M = 1000$ (sampling initial distances), and the traditional method ($k = 1$). $N = 25000$, averaged is over three runs.



(a) $\tau = 100$

(b) $\tau = 1000$

Figure 5.12: $K = 0.7$. PMI using three trajectory separation methods with $M = 1000$ (sampling initial distances), and the traditional method ($k = 1$), computed v N: $\Gamma = 1/3$ regime (averaged over three runs)



(a) $K = 0.7$

(b) $K = 0.1$

Figure 5.13: Low K. PMI using three trajectory separation methods with $M = 1000$ (sampling initial distances), and the traditional method ($k = 1$), computed v τ: $\Gamma = 1/3$ regime, $N = 25K$ (averaged over three runs)

consistent. At $K = 0$, TS PMI using method 1 is lower than the true value; at $K = 0.1$ it is higher; and lower for $K = 0.5$ and $K = 0.7$. However, independent of K, as τ increases to above the range shown here, any bias in method 1 present at low τ disappears.

5.3.2 Running time

The traditional method for calculating PMI involves the following: 1) evolution of N points τ times ($N\tau$ steps) 2) construction of a 4-dimensional kdTree ($4N \log N$ steps), and 3) finding k nearest neighbours for each point ($N \log N$ steps for $k = 1$). The total running time for the traditional method is then

$$T_{\text{trad}} \propto N\tau + 5N \log N. \tag{5.26}$$

It is the first term that causes problems when sampling for the large (N, τ) asymptotic (by typically large values we mean that each of N and τ go up to order of 10^5). It is therefore desirable to find methods that circumvent this dependency on the product.

Consider a variation on the traditional method, one that involves finding the joint nearest neighbour distance by trying to find the nearest neighbour in the joint through first testing whether points in some neighbourhood in the marginal have evolved to stay close enough. In this method all the points are ones from the original sample. The difference is that one does not need to construct a kdTree in the joint space.

Here we would first construct a kdTree in the marginal space to find the interpoint distances associated with each point. Then for each point we evolve it and its nearest neighbour. We accept if the final distance between the two is close enough (see cutoff above). If not, evolve the second nearest point, etc, populating some *evolved* array. This need only be done once, so doing the same process for another initial point may require simply looking up the evolved distance in the array. We thus evolve some $N_e \leq N$ points.

Since the mean joint interpoint distance is an estimate, one could use only $M \leq N$ samples to find it. Accepting the TS conjecture means that this procedure could therefore stop after M steps (providing points are sampled randomly). Let k' be the average number of interpoint distances in the future that one has to check for each initial point before the candidate distance is accepted as the inter-neighbour one in the joint. Then the running time for this method is

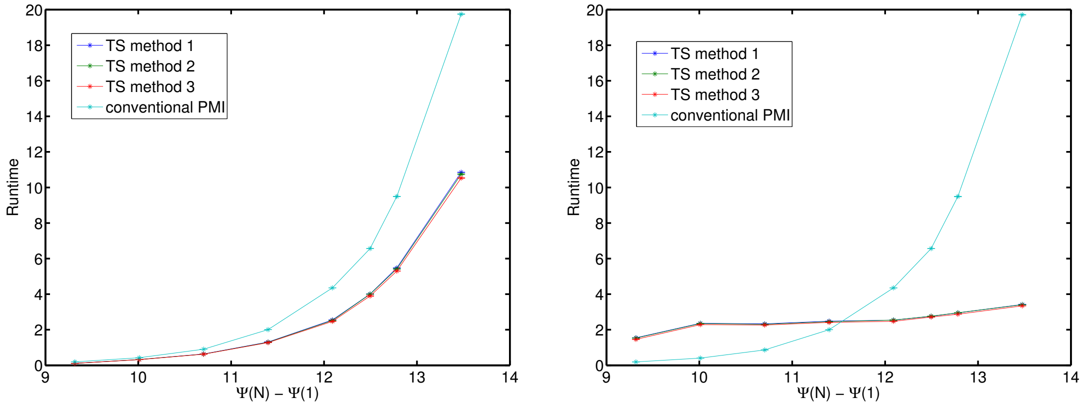
$$T_{\text{new}} \propto 2N \log N + Mk' \log N + N_e \tau, \tag{5.27}$$

where $M \leq N_e \leq N$.

If a smaller sample $M \neq N$ was considered as part of the tradition method, the proportionality of the traditional method running time on $N\tau$ would not change, since to do a selective search on a joint kdTree would still mean a full N-node kdTree has to be constructed first, and all N points have to be evolved for that. Here N_e would depend on M, and of course on the extent of mixing (in a non-technical sense) that τ iterations of F result in. The latter should also have an effect on k': hence for each given N, and a picked M, we have $N_e = N_e(K, \tau, N)$ and $k' = k'(K, \tau, N)$.

So at large (N, τ) we aim to obtain a sample of joint nearest neighbour distances by selectively sampling the marginal and assuming that some small neighbourhood size will be sufficient for determining the mixing behaviour. If, however, $M = N$, the running time with τ for both algorithms is the same, with T_{new} increasing its dependency on $\log N$ by at least an order of magnitude, depending on the map (since it is not reasonable to expect k' to stay of low order, and in our simulations we do see it increase by several orders of magnitude at least). The only reason to use the new method would thus be with a low enough M.

A further variation would start with initial inter-neighbour distances, and evolve pairs of trajectories with the second element drawn from a random distribution, given some set distance to the initial point. The change in the running time is in the number of evolved



(a) $M = 0.01N$

(b) $M = 1000$

Figure 5.14: Running Time for PMI computation in figure 5.5 using three trajectory separation methods (sampling initial distances), and the traditional method ($k = 1$), computed v N: $\Gamma = 1$ regime, $K = 0$ (averaged over three runs).

distances, which in this case is not bound from above by N, since repetition is not possible:

$$T_2 \propto 2N \log N + Mk' \log N + M(k' + 1)\tau. \tag{5.28}$$

This running time does not change depending on whether the initial point is drawn from the original set, or randomly from ρ.

The TS methods in the section above were all variations on the latter procedure, and so we expect the running time to scale as shown in eq. (5.28). We now test the validity of this claim, by first examining variation of running time with M, N, and τ, bringing out the significance of k'.

Runtime v M and N Consider fig. 5.5 that showed that TS PMI for the fully-resolvable case of $\Gamma = 1$, $K = 0$, was relatively close to the true values for both fixed and varying sampling size M. Figure 5.14 displays the respective running times for each of the subfigures in fig. 5.5. From the section before we expect that if $M = N$, the running time for TS methods would be significantly higher, depending on the order of magnitude of k'. Making M to be a small percentage of N lowers the running time, while preserving the overall qualitative dependency on the sample size.

Another way to reduce the running time is to use a fixed M. Figure 5.14(b) shows that when M is greater than roughly 2% it presents a significant decrease in running time. In fact the running time appears to stay almost constant (same order of magnitude). Since

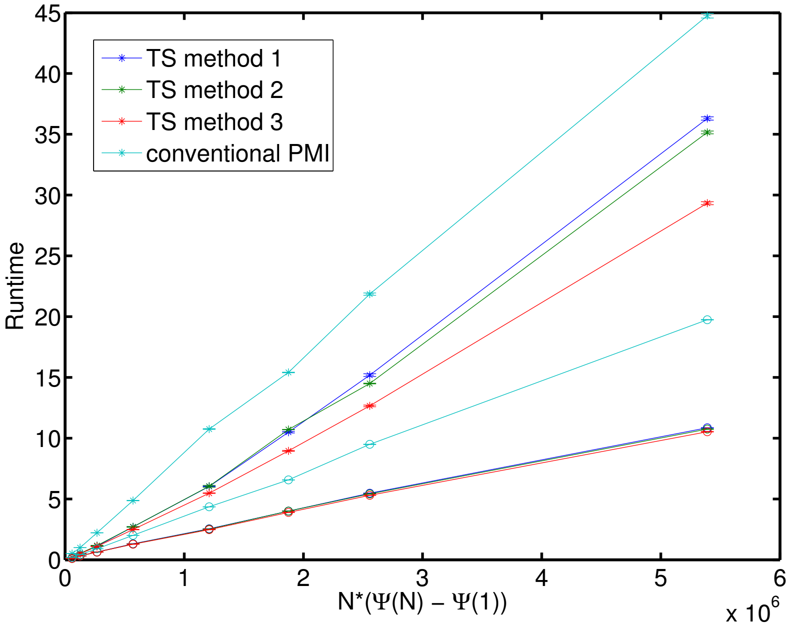


Figure 5.15: Running time comparison for $\tau = 10$ (circles) and $\tau = 200$ for the PMI computation partially in figure 5.14(a) ($M = 0.01N$) using three trajectory separation methods (sampling initial distances), and the traditional method ($k = 1$) (averaged over three runs).

M is kept fixed, this must be due to k'. This is not surprising, since k' is ultimately linked to the mixing properties of the map, and this range of N (given a τ) was chosen for its *fixed* Γ regime, where no qualitative change occurs. We therefore expect running time for TS method 1 to change, even for a fixed M, when the (N, τ) parameters are in one of the intermediary regimes of Γ (see section below on τ dependency).

How would the above figures change with τ? We examine the variation of the first subfigure: the second is looked at in the next section. That is because the best way to examine k' dependency on the parameters of the joint distribution is to use fixed M. For the variable M, on the other hand, we compute a counterpart of figure 5.14(a) but for the τ corresponding to $\Gamma = 1/3$, and plot it on a $N(\log N - \text{const})$ scale. The result for both τ values is shown in figure 5.15.

A higher τ results in lower PMI, and so in higher running times for both methods. The plots also show the expected dependency on sample size N. On the other hand, while the running time for the traditional method will continue in the same manner as N goes up to infinity, the TS methods will show a change if at some point the parameter space (which includes N) results in a transition to a different Γ. Hence these results should be used with care.

We also see that the running times of different TS methods begin to differ at large τ. Given the tentative link between running time and PMI itself, we should expect the PMI values for the methods to differ as well. The respective PMI measurements were shown in fig. 5.7(a), where indeed TS method 3 followed was significantly off from the trends in methods 1 and 2, which deviated from each other only at the larger N.

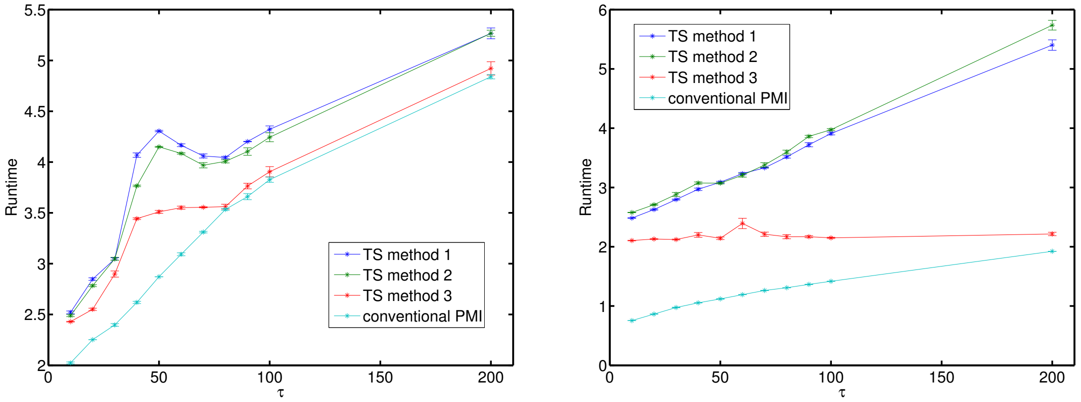
Runtime v τ

This section examines how the running time of TS methods varies with τ. From eq. (5.28), running time is proportional to $M(k'+1)\tau$, where M is the sampling depth of V, and k' is what we will call 'effective neighbourhood', i.e. the size of the neighbourhood out of which all the points have moved out of by the time $t = \tau$. In other words, it is the index of the first nearest neighbour in the past that is further than the nearest neighbour in the joint (using the terminology in fig. 5.1, $k'+1$ is just j^* averaged over the M sampled points). In metric terms the neighbourhood size could be estimated using the average point separation, a function of N.

We have already noted that k' potentially depends on (N, τ, K), and the description above supports the notion that k', while expressing the level of difficulty in computing TS PMI, is by doing so indicative of the level of mixing (in the non-technical sense) of the map. The purpose of this section is hence not so much to find the regimes where TS calculations offer an advantage in terms of running time, but rather to better understand the mechanisms involved.

Figure 5.16 displays the running time behind calculations for $K = 0$, and $K = 1$. Comparing it to the variation of PMI values in time, we a) cannot assume that the running time of the conventional PMI method does not also to some degree reflect the PMI value - since the variation is smooth in both cases, and b) can conclude that the sensitivity of TS to map behaviour translates, to a large extent, to the time it takes to perform these computations (compare the figures displaying values and the corresponding running times for an almost exact mimicking of the trends). A stronger statement would regard the running time itself as being a good indicator of map behaviour (see below).

We now look at the global trends. Figure 5.17 shows the running time for traditional PMI, as well as the three TS methods, for the fully-regular, and primarily chaotic, motion. In



(a) $K = 0$, $N = 50000$ (for fig. 5.9(a))

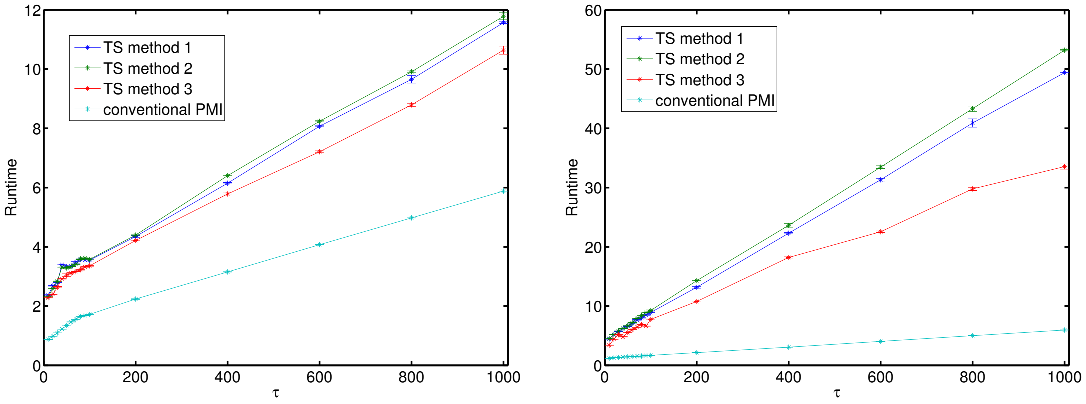
(b) $K = 1$, $N = 25K$ (for fig. 5.11(b))

Figure 5.16: Running time for PMI computation for the three trajectory separation methods (sampling initial distances), and the traditional method ($k = 1$), computed v τ (averaged over three runs). $M = 1000$.

both cases the traditional PMI running time shows, as expected, a linear behaviour with τ. Its slope changes with N and would, in fact, purely in graphical terms be equal to that of TS PMI for method 3, if $N = 50000$ (in the $K = 0$ case). We also see that, especially for large τ, the TS methods (we focus on method 1, which gives the smallest error between TS PMI and traditional PMI) display a constant slope, suggesting a certain constant, or slowly varying, k' (in all these plots we keep $M = 1000$).

Given that the slope should be directly related to k', and we posited that the latter has a direct relation to PMI, it is informative to look at the variation of estimated slope $< k' >$ of method 1 running time with τ. Let $I_\tau = [100, 1000]$. For each sample size N we estimate $< k' >$ in I_τ for various K parameters. The number of points from which the estimation was done is relatively low (below ten) due to the almost exact alignment with a straight line that can be drawn through them. This is true for all the K values tested. The error shown is a 95% confidence interval, which as we see is small enough to give an indication of the general trend.

Results for three sample sizes that are middle range as far as conventional PMI calculations are concerned are shown in figure 5.18. In this picture N, as before, represents the resolution of the distribution; the level of visibility of the effects of the map. $< k' >$ is the factor by which a neighbhourhood, of some size defined by N, expands during an iteration. We see that that this rate is, on average, greater on smaller scales. Following individual N behaviour, we also see that at small K regimes this expansion rate is the same - the



(a) $K = 0$

(b) $K = 4$

Figure 5.17: Running time for PMI computation three trajectory separation methods (sampling initial distances), and the traditional method ($k = 1$), $N = 25K$, computed v τ (averaged over three runs). $M = 1000$.

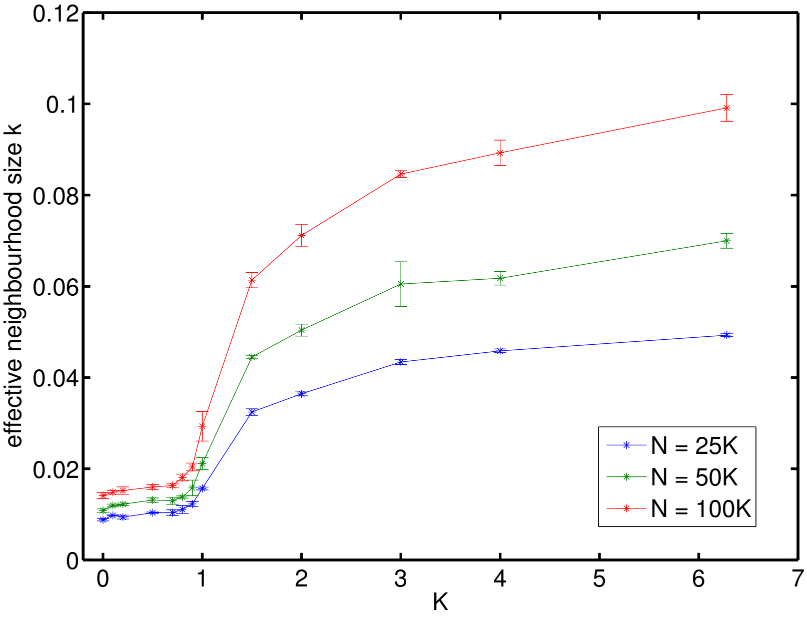
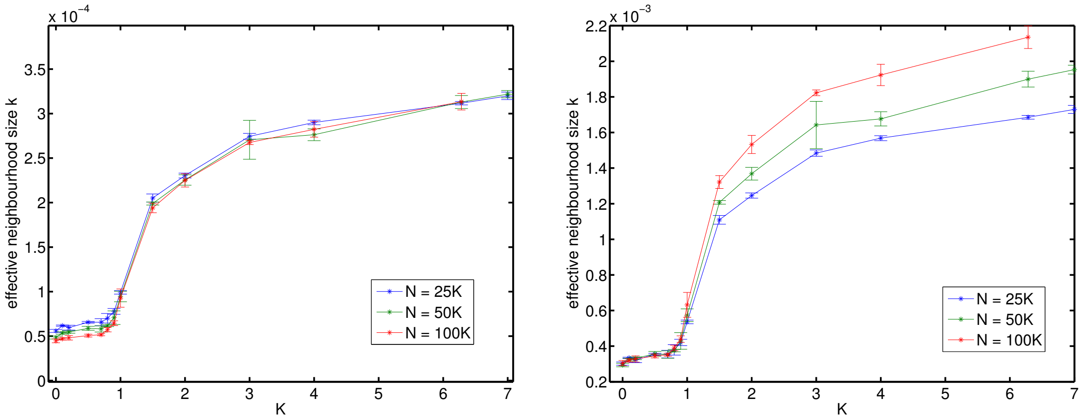


Figure 5.18: Each value represents the slope $< k' >$ of method 1 running time (averaged over 3 runs) v $100 < \tau \leq 1000$, at that particular sample size N ($M = 1000$ throughout).



(a) Scaled by $N^{-1/2}$

(b) Scaled by $N^{-1/3}$

Figure 5.19: Collapsed plots for $< k' >$, the slope of the running time for TS PMI method 1, $N = 25K$, 50000 and 100000, computed v τ (averaged over three runs, $M = 1000$)

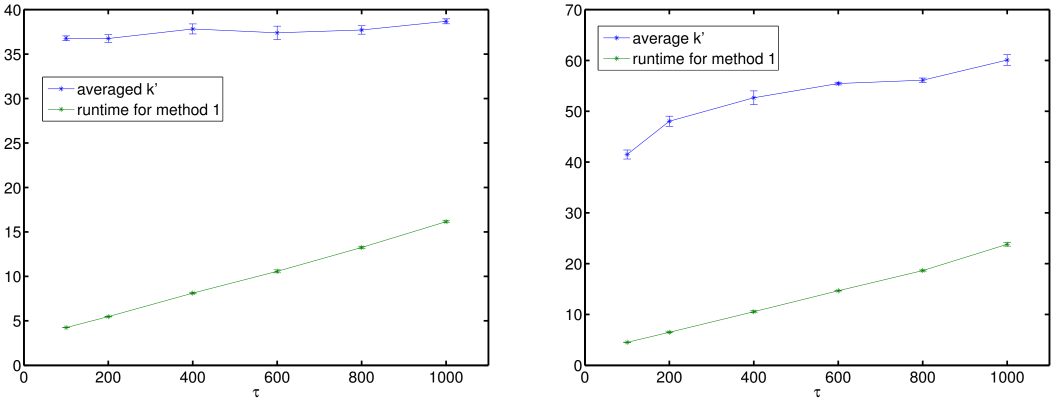
extent of chaos does not change it much. Then after $K \approx 0.6$ the rate begins to increase dramatically. What is interesting is that this change happens on all scales - all the N plots begin to curve upwards roughly at the same K. In this picture K_c does not appear to be at all significant - nothing qualitatively new happens at the break down of the last KAM torus.

The difference between plots of different N also changes with K. Figure 5.19 shows that scaling by $N^{1/2}$ collapses plots at large K, whereas at small K values that factor looks more like $N^{1/3}$. Zooming in on the plots it becomes clear that the cutoff point circa K_c between the two regimes is only an apparent threshold related to the visible scale of the graph below; the real change over begins at smaller values of K.

This suggests that $< k' >$, the average neighbourhood expansion rate (as found through looking at the gradient of running time v τ) scales as $N^{a(K,\tau)}$. We also see that there exists a region of τ where $a(K, \tau) \approx a(K)$. From the graphs, we infer that for large values of K, $a(K) \approx 1/2$, and for small K, $a(K) \approx 1/3$.

We do not yet have an interpretation of this. It would be tempting to see if these results can related to the possible multifractal nature of the joint.

In the calculations above we estimated the slope and assumed a link to the number of nearest neighbours considered. It is also possible to compute the latter directly. We examine the k'_{av}, or k' averaged over its M values. Figure 5.20 below shows both the running time and k'_{av} for two K values featured on the previous graph.



(a) $K = 0.5$

(b) $K = 1$

Figure 5.20: Running time, and k'_{av} for TS PMI method 1, $N = 50000$, computed v τ (averaged over three runs, $M = 1000$)

In both cases running time appears to increase linearly with slope $< k' >$ which we take to be equal to k'_{av}. However, while true for some map values (5.20(a)), it is manifestly not so for others (5.20(b)). We thus understand that $< k' >$ would also vary depending on which τ subinterval is used to compute it.

We first disregard the variation with τ and check that the global behaviour of which both k'_{av} and $< k' >$ are indicative of is the same. Just as $< k' >$ is a function of the τ interval, so we take the second average of k'_{av} with respect to it. We use the same letter under the understanding that a function of K would always encompass averaging over I_τ.

Figure 5.21 shows a transform of $< k' >$, scaled up for better comparison with k'_{av}.

Also shown in the figure is a plot of the linearly transformed PMI. The actual PMI graph is of course inverted, to some extents mimicking the Γ plots with their parameter-dependent peak location. The reverse dependency stems from the fact that a growing k'_{av} is indicative of a wider distribution, which in turns implies higher joint entropy, and hence a lower PMI. We see that a linear transform allow us to align PMI with k'_{av} almost exactly; but that this logic breaks down for small values of K.

The graph above contains no information about whether trends will change with τ, in other words, the PMI and k' are averaged over τ with no regard for whether they are stationary or not. Yet figure 5.20 leads us to expect interesting interdependies on K.

Instead of averaging PMI and k'_{av} with respect to τ, we can examine their stationarity by looking at the slope over I_τ. The values fluctuate, so we increase the number of measurements to bring the error bars lower. Figure 5.22 shows the resulting slopes for a range of

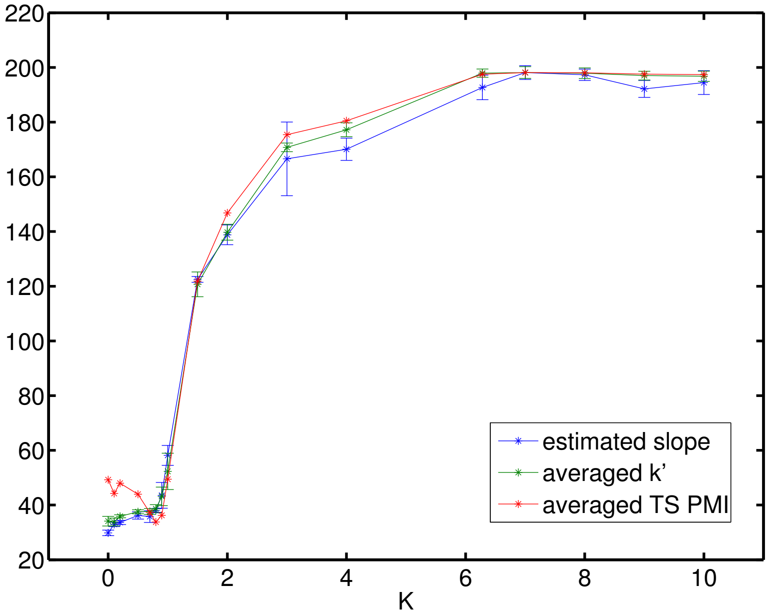


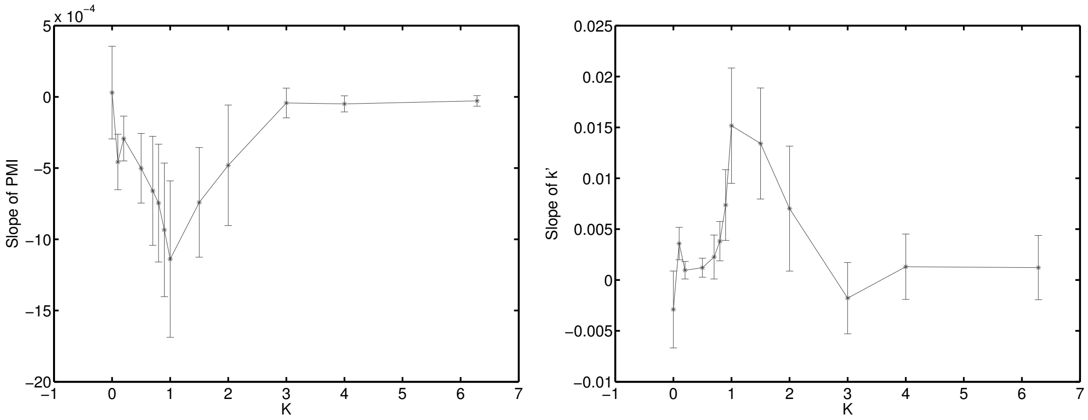
Figure 5.21: k'_{av}, and linear transforms of both TS PMI and averaged slope $< k' >$ of running time v τ, for $N = 50000$ ($M = 1000$ throughout, three runs for each (N, τ, K) value).

K.

Slopes of k'_{av} and PMI behave, as expected, in opposite ways. Maxima of absolute values *for both* are reached at roughly the same point, close to K_c. This is interesting because until now K_c did not herald any qualitative change. Whereas here, for a particular N the linear approximation to the gradient of I is

$$\frac{dI(N,K,\tau)}{d\tau} \approx c_1 \frac{dI(N,K_c,\tau)}{d\tau} + c_2|K-K_c|^a, \tag{5.29}$$

where using the data behind figure 5.22(a) we find that for $K_c \leq K < 4$, $a \approx 0.8$, and for a region on the other side of K_c, $a \approx 0.65$. We note that for $K > K_c$ the change is abrupt, and after roughly $K = 4$ the slope of PMI with τ stays zero. Of course these are linear approximations, whereas for example for $K = 1$, plotting k'_{av} with τ does not give linear behaviour with errors. So rather the above graph should be used as an indication of *nonlinear* behaviour, rather than noisy linear behaviour. This scaling, however suspect, is interesting, because this is the common scaling form in literature for behaviour of various functions off K_c. Future research could test whether PMI captures some already-known numerical scalings in the standard map.



(a) Slope of PMI

(b) Slope of k'

Figure 5.22: Slope of PMI and k'_{av} measurements w.r.t. τ, for a range of K values, with $N = 50000$, $M = 1000$. Each $I(\tau, M, N, K)$ is an average over 6 runs; tau is still in I_τ, but the number of values has increased to bring the 95% confidence intervals down.

Concluding Remarks Our considerations of PMI led us to expect that evolution of future interpoint distances could be indicative of some map features. Here we took this idea further and showed that by combining the past and future interneighbour distances for each trajectory it is possible to derive the joint interneighbour distance ϵ, the key ingredient in the K-G entropy estimate used for computing PMI. The basis of this method lies in a *local* perspective of the effects of the map, where we assume that ϵ is more likely to be realised by an orbit originally in the vicinity of the orbit in question. This forms the basis of the search stategy. The advantage of this method is that it allows sampling of the joint distribution formed by N points without actually having to compute *all* the interpoint distances in the joint space - rather only the needed ones. We conjectured that sampling these initial distances would produce the correct PMI.

We have seen that the TS framework, through treating time and sample size as variables that influence the separation of localised marginal interpoint distributions, can accurately predict dependencies which lead to a qualitative change in Γ. This was done for the fully-regular $K = 0$ regime when Γ reduces from unity. An interesting analytical extension would be to derive the large τ limit of $\Gamma = 1/3$. In that case the shape of the final distribution is known, and if a variable initial separation is assumed it would lead to a more smooth final curve, with less travelling peaks. An even further extension is to see how this picture leads to the chaotic scaling observed for the two-staged Γ plots for the larger K. A naive separation of linearly and logarithmically moving regular and chaotic distance peaks does not yield the correct scaling. We conejcture that at least part of the reason lies in the fact that at nonzero K the initial distance does not simply translate, but is varied, especially since the maximum metric, as we have seen, tends to obfuscate behaviour between points that are both on a curved trajectory. This would alter the relative positions of distances in the evolved D^τ family, to an extent defined by K. And in fact when we get closer to K_c we do observe that the original τ^3/N scaling breaks down. As a result successfull analytical treatment of the TS method would more explicitly relate these properties of the map to the change in the relative positions of D^τ with τ, and potentially use the former to derive Γ for other regimes.

For practical PMI computation using the TS method we successfully decoupled the depth of sampling of the state space, N, from M, the strength of sampling of the joint distribution for nearest neighbour distance. We investigated three possible variations of the TS method, and found that the one that keeps track of both the positions of the initial separation family and the actual initial interneighbour distances, which we sampled from a marginal kdTree, works best (though we also saw that there are regimes where the methods coincide). This optimal TS PMI was found to coincide, with relatively small errors, to the value found using the traditional method where the nearest neighbour parameter k was set to one. We found that certain regimes introduce a small consistent bias that does not change with N, but that is within reasonable limit of the true PMI. We also found that that bias is not a result of sampling strength M and thus could not be decreased by considering a larger amount of neighbourhoods. On the positive side this means that for practical purposes M does not have to be proportional to N but can be chosen to be a constant. Enough information about the joint distribution is obtained even with M decreasing to less than one percent. The only regimes where TS PMI deviates from the true values are in the small K regions, during the transition from $\Gamma = 1$ to $\Gamma = 1/3$.

Having shown that, particularly for large τ, TS PMI is a successfull candidate for the true PMI, we looked at whether this method could present an advantage in terms of practical calculations. We find that there are regimes for which running time is *significantly* decreased. If the sampling depth (M) is kept fixed, which the previous section demostrated to be a feasable solution, with small enough errors, then we have shown that the running time can be kept under control as the resolution of the map increases. The main problem with the method is that decoupling the strength of sampling from the state space resolution comes at the expense of having the execution time become dependent on the details and mixing properties of the map. We showed that there is a variation of the method where the execution time is limited above by a factor proportional to the resolution N; though this was not utilised since the aim was to test if the method works with known map parameters. We also saw that for some parameters the length of the running time comes to reflect directly (through linear transfroms) the features of the map picked up by PMI. It is an interesting question of why small K values force a deviation between the two notions. Thus the TS method can be used with a significant running time advantage when the map regimes are more or less predictable, which of course is the case with slow relaxation near the critical K value. PMI values for higher and higher N could thus be found for higher τ, while keeping M at a low level, without expecting a qualitatively different rate of running time increase. Due to the nature of the method, however, the problem of finding the regimes when TS is advantageous is tied in with the problem of knowledge of mixing properties of the map itsef.

We can see by the TS construction that k', the effective neighbourhood weight that has escaped by τ, while directly responsible for the running time, is also an indicator of map behaviour. Its variation with time mimics that of PMI, and is most nonlinear near the critical K value. Preliminary results for pdfs (not displayed) show that their shape undergoes a qualitative change as K passes through K_c. Looking at this could provide an understanding of averaged behaviour, and an interesting test would yield comparisons to variation of averages with time between pair distributions in the previous section.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this work we introduced Persistent Mutual Information $I(\tau)$, a probabilistic measure of nonlinear inter-relation between the past and future ensemble distributions separated by a time gap τ. The initial motivation was the possibility that a quantifier of this type could potentially detect dependencies persisting over time, and that there is thus a sense in which it could be said to have detected strong emergence.
We used data generated by attractors of several dynamical systems in an attempt to understand whether there are specific features of these systems that make them qualify to be strongly emergent. The conceptual idea behind using deterministic dynamical systems to generate ensembles is the uncertainly in the specification of the initial condition.
For the simple archetypal examples of one-dimensional chaotic systems, the logistic and tent maps, PMI picked out the existence of a global clock. That is in line with the expectation that the only persistent features possessed by the attractors are the different phases, given that no initial uncertainty can withstand the 'mixing' effects of chaos. By conjecturing that it is the limit of $I(\infty)$, the Permanently Persistent Mutual Information, that can be considered a signature of the strong emergence, we find that for systems with global periodicity T, $I(\infty) = \log(T)$, independent of the chaotic overlayer. This holds for both maps.

PMI does not require the arbitrariness of a finite partition, which is exactly where a number of symbolic dynamics measures with similar functional form fail to give a universal answer potentially applicable to systems with differing state spaces. This also sweeps away the major computational difficulty of empirically computing the distribution over block variables. In addition to that, the initial condition is uniquely associated with the distribution over the infinite past, so the main object becomes the joint distribution embedded in the space with box-counting dimension equal to the sum of those of the marginals. In our computations we used the parameteric K-G estimators of entropy and information, where varying the parameter allowed us to change the depth of sampling. For instances of infinite periodicity, and generally those where PMI grows indefinitely with resolution, such as the period-doubling accumulation point of the logistic map, we extended the phenomenology and demonstrated that PMI grows with the logarithm of probability resolution at a rate Γ, dependent on the information dimension of the underlying spaces. This allowed us to make sense of PMI in a variety of systems, including area-preserving ones with no defined attractors.

To the best of our knowledge no measure currently exists that quantifies the amount of emergence in the behaviour of these dynamical systems in a way that corresponds to shared intuition as much as PPMI does (though of course a lot of it is by design). Unlike excess entropy or the entropy rate it picks up intrinsic persistent features of the maps. As such, it is essentially a categorization tool. It has all the potential to become useful, and indeed it should be tested on a wider range of dissipative systems to see what features, other than clocks, it picks up on. These could then be assessed in terms of whether or not we want to count them as strongly emergent.
A possibly interesting extension here is to revise the way uncertainty enters into these dynamical systems. For instance the distribution over the initial conditions could be replaced with the distribution over the past by considering a map with noise. The down point is that calculations would now have to involve distributions over block variables, with the assumption of infinite block lengths. However, the end result would be seeing whether by this measure a noisy map displays the same extent of emergence as the fully deterministic one. It should be practically testable, and results could prove interesting in terms of speculations about how emergence should be defined. Also we notice that in the Logistic map the linear gradient of PMI with τ shows the qualitative features of the Lyapunov exponent. This idea could be put of a firmer foundation. Thus there is in general scope for extending the phenomenology of PMI in terms of stochastic processes, with work in this field already being done by Gmeiner [2012].

We then computed the PMI in the standard map. This allowed us to move away from systems with clocks, and into the territory where there were no obvious, intuitively expected results. The standard map was the natural next choice as an area-preserving map with a different route to chaos, one nonlinearity parameter, and very rich behaviour. Since here PMI was shown to increase indefinitely with resolution our analysis was done in the language of fractal methodology, and results were expressed in terms of Γ. Γ was shown to be useful in describing the extent of causality, where by causality we mean lack of distortion of the initial conditions. As expected by the phenomenology of the standard map we found that in the fully regular case with no chaos Γ saturates to a value that gives the joint information of 3. Moreover, we found that regular trajectories result in a particular scaling of Γ with resolution and τ. This scaling persists at large K values, and is purely the result of the existence of regular trajectories. Likewise, at those K a new, chaotic scaling emerges. It would be the next step to relate that to the Lyapunov exponent.

The manner in which causality in the standard map changes with resolution and time separation can be viewed on the (τ, N) contour plots of Γ. These are interesting in that they unite the resolution and time variables, and hence could also be used for comparisons between different systems. Based on preliminary results for the Double Pendulum shown further on we suggest that that would be a good first comparison. We see some similar Γ behaviour, which implies that a phenomenology of the KAM breakdown route to chaos in terms of causality could possible be developed.

We then investigated whether the joint distribution can be decomposed into distinct subsets stemming from the regular or chaotic evolution. This was formulated in a mixture hypothesis. To test it we developed a method to find the (assumed existing) proportion of regular trajectories by tracing the evolution of interpoint distances. The bimodality of these distributions could clearly be seen to vary across K, so much so that around K_c the chaotic component did not have a 'typical' average divergence rate. This is a clear demonstration of the slow relaxation times around the golden KAM breakdown.

Tracing the separate evolution of regular and chaotic trajectories (where these were defined as equivalence classes based on a neighbourhood expansion rate) we showed that for at least one K value where α appears to be clearly defined, $\Gamma = \Gamma_J > \Gamma_{\text{mixture}}$, with little indication of convergence with τ. A higher Γ means that as one looks at a higher resolution of the map one would obtain more information about the future state than would have been possible if the orbits were clearly disjoint. In the mixture hypothesis the new orbits one would see would be of the same type. We thus conjecture that this is not so in the standard map; that there is at least a substantial subset of the state space where the chaotic orbits and regular orbits come arbitrarily close together. We do not see it as being the effect of stickiness since no change seems to occur with time separation. The multifractal methods shown further could potentially be of help in determining whether and at what K the joint is a true multifractal. We also proposed to use the difference between the 'true' and 'mixture' Γ values to quantify the extent of this spatial interlinking of orbits of different type.

We then used the pair-wise separation distances to express $\Gamma(k = 1)$. In traditional

PMI calculations the resolution N that defines the joint probability distribution was equivalent to the depth of sampling of the joint, $M = N$. We noted that for any point in the sample the joint nearest neighbour distance can be computed from a set of local marginal interpoint distances, and used this to show the correct exit time from the $\Gamma = 1$ regime for the integrable case. We then decoupled N from M, developing a method that samples the joint with any M. It was experimentally shown to converge well in most cases, and showed a qualitative and quantitative improvement on runtimes. This can be particularly useful in testing the (τ, N) asymptotics, since in the standard map literature 'settling' times over which trajectories can be expected to be seen as unstuck can go up to 10^{10}.
The runtime of this method is, however, implicitly dependent on the way (loosely defined by k') in which the map mixes up the neighbourhoods. It could be possible to link this rate to a bound on the metric entropy by considering the distinguishability of trajectories arguments motivating the latter. We also noticed that the K ranges where k' very closely aligns with PMI are primarily for $K \geq K_p$. The fact that there is a qualitative difference at small K values could hint at the way in which map dynamics influences PMI.

Computations were much simplified through the use of the kdTree routine. It has shown itself to be readily adaptable to both a change of metric and a relatively large dimensionality - the largest tested was eight for the joint space of the Double Pendulum. Our preliminary results suggest that the Double Pendulum shares at least some Γ phenomenology with the standard map. Both of these systems could also be tested for multifractality of the joint, which once again was made computationally feasible by the kdTree construct admitting a variety of routines that can find either the distance to k^{th} neighbour, or number of neighbours within a certain distance, a fact that came especially useful when testing reliability of the (q, τ_q) variables in the multifractal analysis.
The next two sections show some preliminary evidence that the joint distribution of the standard map around K_c does appear to be a true multifractal. Around that regime we also see that stickiness, the particular mode of transport introduced by the cantori, results in an apparent peak in Γ, which corresponds to higher causality and consequently better predictive regimes, peaked around some $K = K_p$. Interestingly for the observed range of data $K_p \neq K_c$. At K_c we observe logarithmic decays of Γ with τ, and a breakdown of the regular and chaotic scalings. Such behaviour is qualitatively in line with the slow decay in correlations around K_c.

Preliminary results for the Double Pendulum indicate that a similar peak occurs at $E = 1$. We propose testing the joint distribution of the Double Pendulum for a variety of E values. By comparison with the standard map those 'anomalous' peaks in Γ could be associated with the change in the fractal nature of the joint distribution. Thus our research leads us to suggest there there is a level on which area-preserving maps can be discussed in terms of multifractal phenomenology.

The focus of this work was to test a quantity describing the preservation of correlations in time on data from various types of dynamical systems. The natural extension of this effort is to use PMI on real-world datasets, and see whether it succeeds in accurately identifying the number of choices available to the underlying system, in other words the extent of 'strong' emergence. The main computational hurdle in our methodology is the time to estimate mutual information of the joint, which involves a nearest neighbour search in a $2d$ space, where d is the dimension of the (past and future) data. The runtime in the straightoward kdTree method scales linearly with d, so reasonable computation times can be achieved for much higher dimensional systems albeit at the expense of the number of datapoints. We therefore propose PMI as a good candidate to measure emergence in real-world systems.

6.2 Ideas for Future Work

6.2.1 Multifractal Analysis

Consider area-preserving maps. There Persistent Mutual Information was found to scale with resolution as a function of the information dimension of the joint. Yet if the focus shifts away from prediction and towards understanding the system *through* properties of the joint distribution then one could study D_1 as an element of the spectrum of generalized dimensions D_q. In joints with simple fractal support all these would be equal. We propose asking whether the joint is fundamentally a multifractal, and if so, at what regimes.

Our motivation was the notion behind the mixture hypothesis, the possibility that the observed behaviour is just a result of linear mixing from two competing distributions (the joint of regular and chaotic trajectories). If that is true, nothing fundamentally new or different would be seen by increasing the resolution. The effect would be the same as would be produced by moving from one region to another. In other words, do we explore the whole ensemble by just zooming in?

From Halsey et al. [1986] where $\tau_q(q)$ is defined as the separatrix in the $Z_s(q, \tau_q)$ variable, we have

$$Z_s = \sum_{\text{boxes}} (\delta\mu_i)^q \, (b_i)^{-\tau_q} \,, \tag{6.1}$$

where $\delta\mu_i$ is the integrated measure of i^{th} box, and b_i its linear size. We used a variety of methods to compute the spectrum, since it is notoriously prone to errors. Direct estimation of α (Badii and Broggi [1988]) did not do so well. Fortunately, the kdTree routine can easily be adapted to find the weighted radius given some neighbour index k, to be used in fixed-size procedures. However as noted in Grassberger [1990] in terms of kdTree computing box lengths was naturally the easiest, following the fixed mass approach from for example Grassberger et al. [1988].

Having ready access to k allows us to rewrite $q(\tau_q)$ along similar lines to the correction introduced by PG in Grassberger [1985]. If $Z = \sum_{\text{boxes}} (b_i)^{\tau_q}$, $Z = Z(k)$ since boxes are defined by weight k/N, this argument views the rate of change of $\log Z$ as being the approximation to the real 'difference' function which we use to get $q(\tau_q)$:

$$1 - q = \frac{k\,[Z(k+1) - Z(k)]}{Z(k)}. \tag{6.2}$$

To get the $q(\tau_q)$ value in practice we take the mean over measurements at different k.

For each k and τ_q we use the standard kdTree neighbour-distance finding routine to obtain the weighted box $Z(k, N, \tau_q) = \langle b^{-\tau_q} \rangle$. Figure 6.4 shows the graphs implicit in computing $q(\tau_q)$ in the traditional 'slope' method, and the 'difference' method, the latter being careful to ignore the small $k < q - 1$ range. Although not shown, the resultant (q, τ_q) graphs are almost identical. The multifractal spectrum is shown in figure 6.2, where since the straight lines are to be ignored as being the consequence of the slight concavity in the (q, τ_q) picture and a minimizing procedure, the joint is shown to be monofractal with, as expected for those parameters, a dimension of 3.

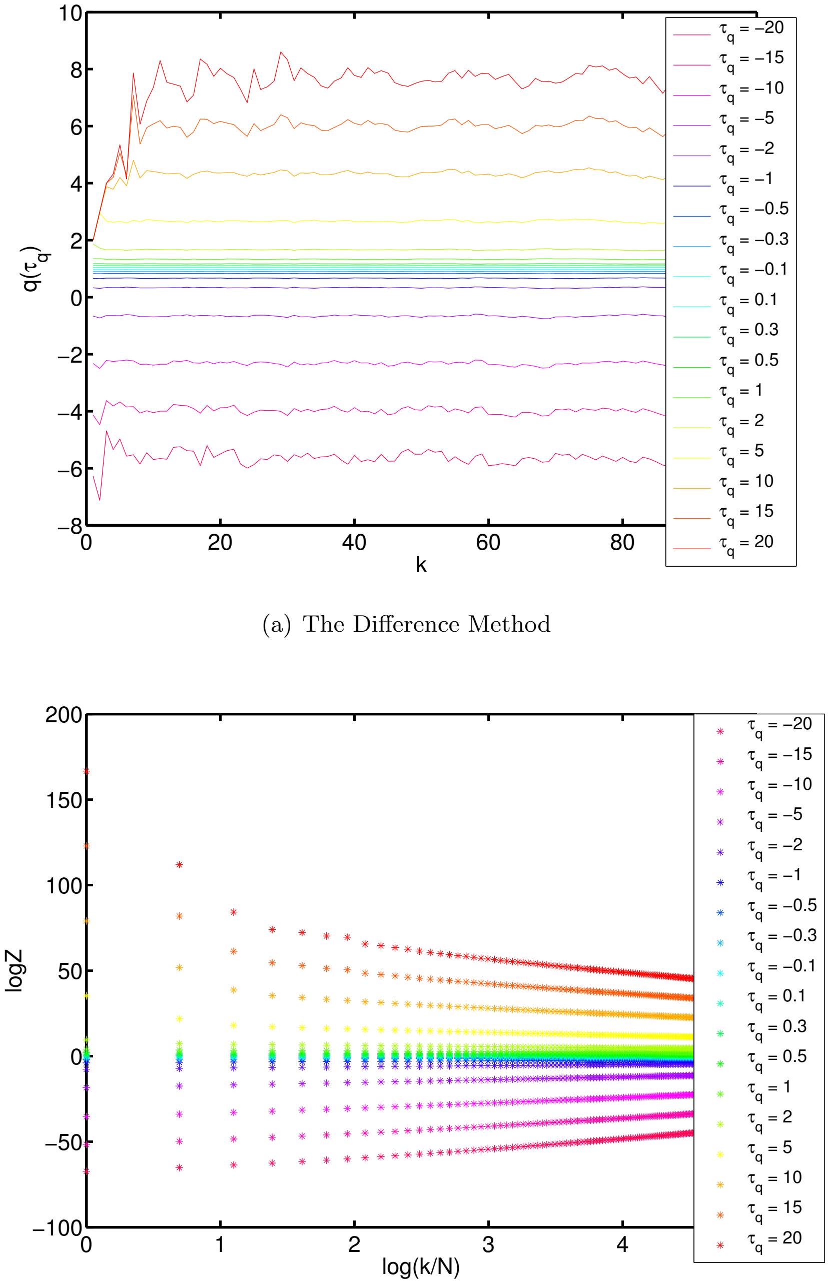
The 'difference' methodology algorithm can be used to capture several $q(\tau_q)$ plots at given ranges of k. This presentation immediately allows us to see differences in the spread of q values, which could bely a varying information dimension. Figure 6.4(a) shows the multifractal equivalent of the Γ dip observed at $K = 0$ (at least we conjecture it is so. The alternative explanation is the commonly given lacunarity, but we do know whether the highest information dimension computed at curved Z regions could even theoretically be higher then the embedding dimension, which is what we see here).

Computing the multifractal spectrum at the low range of k shows that at these high resolutions the joint is monofractal with dimension equal to two (figure 6.4(b)). This agrees with the Γ limit of one.

In order to see the joint dimension of three τ needs to be large enough; for small values the tip of the triangle in figure 6.2 does not reach the line with the slope of unity.

At low enough τ we observe the lack of clear convergence in the 'difference' plots. The same effect can be seen in the more traditional $\log Z$ picture (see figure 6.4 for comparison), where in Theiler [1990] it was mentioned as having come as a result of lacunarity. Figure 6.5 shows that with increased τ the oscillations (whether viewed in on the $\log Z$ plot, or in terms of the 'difference' picture) become more frequent, yet it is possible for the averages not to change. Hence to get to the result of $D = 3$ one then needs to increase either τ or the range of resolution.

As K approaches K_c the joint starts resembling a true multifractal with a range of well-defined point wise dimensions α (figure 6.6). We do not know whether it is the presence of sticky trajectories that turns the distribution into a multifractal (which would mean that it is actually monofractal in the infinite τ limit), or whether it is the arrangement of the spatial locations of the regular/chaotic trajectories. In this respect it does not help that



(a) The Difference Method

(b) The Slope Method

Figure 6.1: $K = 0$. Two methods for computing $q(\tau_q)$, at resolution k. $N = 10000$, $\tau = 100$.

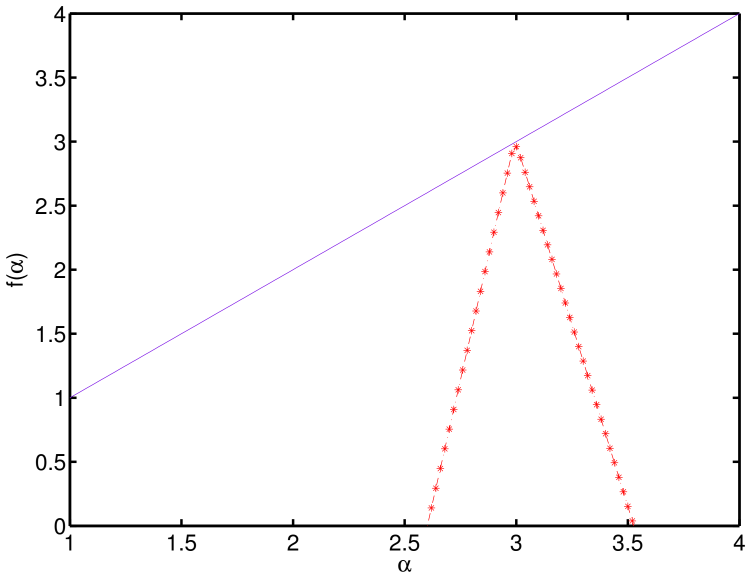
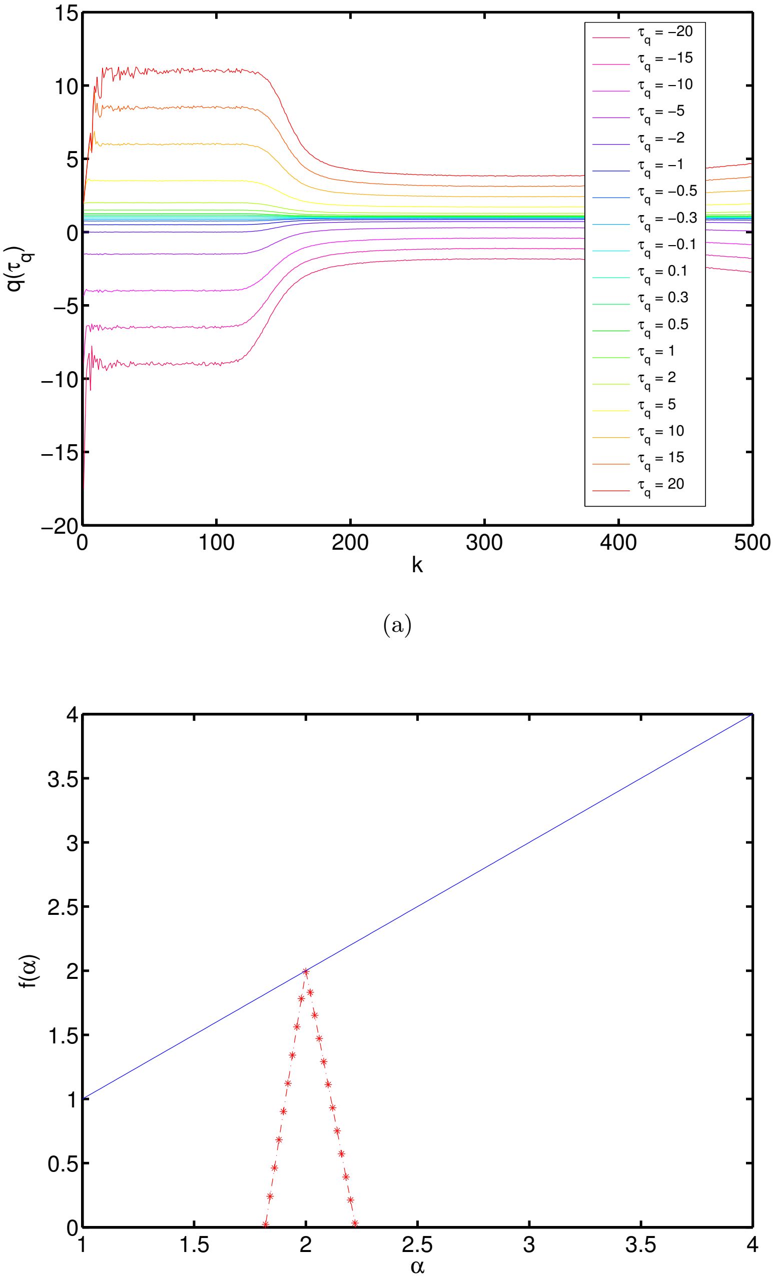


Figure 6.2: The multifractal spectrum at $K = 0$, for $N = 100000$, $\tau = 100$.

the two phenomena are intrinsically linked.

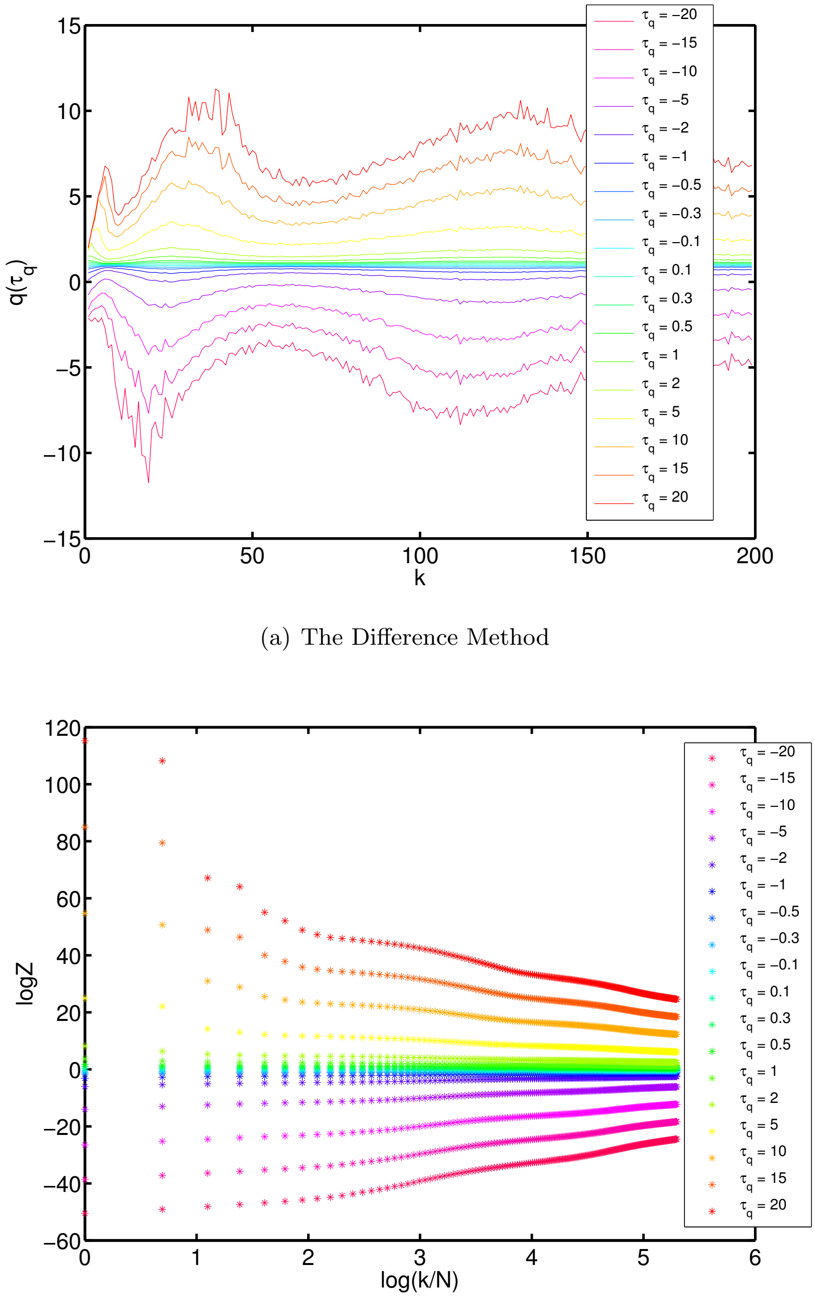
In Chapter IV we proposed using the difference between Γ and mixing Γ to quantify the extent of spatial mixing of various types of trajectories. Multifractal spectra provide a much more multi-faceted description of the structure of the joint. It now seems possible to use for this purpose the quantities typically associated with it, such as D_{∞}, or $D_{-\infty}$. An even simpler quantity is the box-counting dimension of the support of the joint, the peak of the curve. Analogously one could measure the spectrum of the double pendulum at around $E = 1$, where we see a similar peak in Γ. A further point of interest is the long stretch of higher energies where Γ seems stable at a non-integer joint dimension.



(a)

(b)

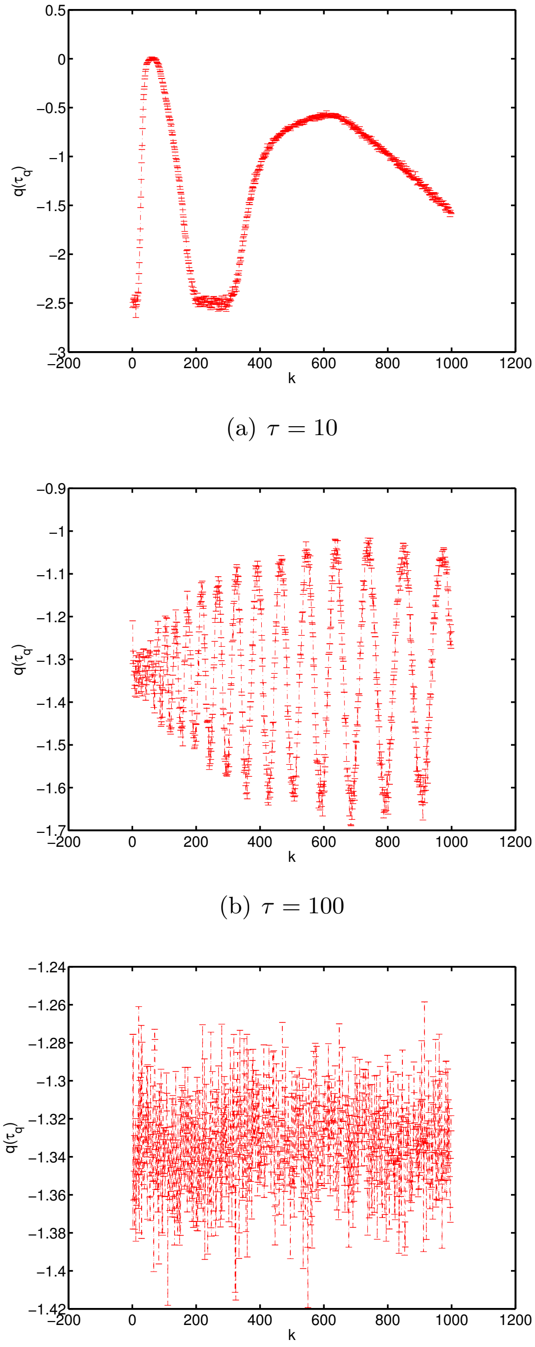
Figure 6.3: $K = 0$. The multifractal spectrum (6.6(b)) and the variation of $q(\tau_q)$ with resolution k (6.6(a)). $N = 50000$, $\tau = 10$.



(a) The Difference Method

(b) The Slope Method

Figure 6.4: $K = 0$. Variations in the $q(\tau_q)$ in the two multifractal methodologies. $N = 1000$, $\tau = 10$.

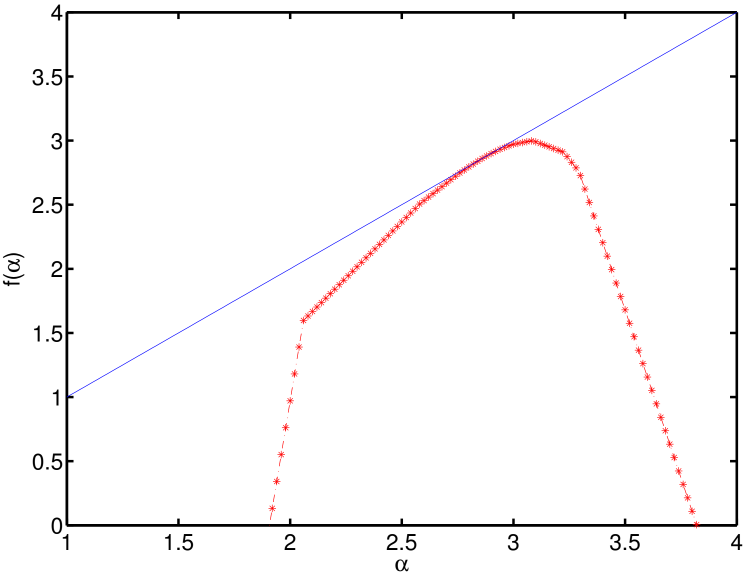


(a) $\tau = 10$

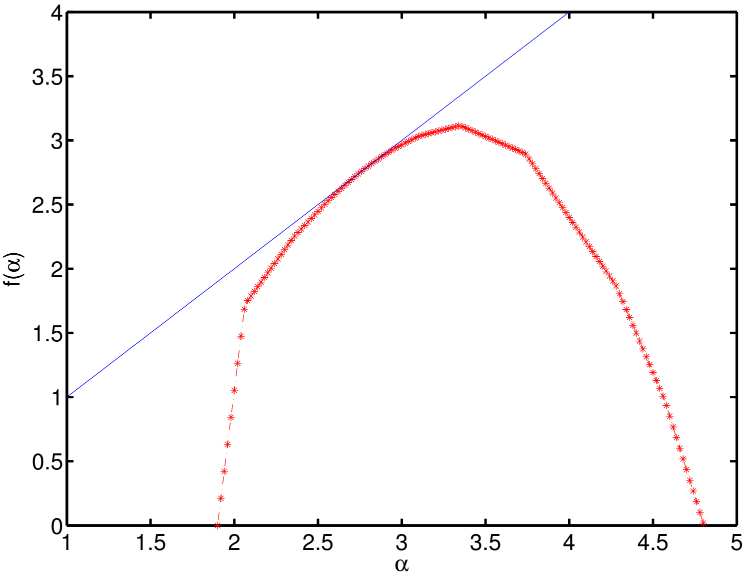
(b) $\tau = 100$

(c) $\tau = 500$

Figure 6.5: $K = 0$. Variation of $q(\tau_q = -7)$ with τ for $N = 10000$, showing the effect of lacunarity or the metric.



(a) $K = 0.5$



(b) $K = 0.97$

Figure 6.6: Multifractal spectra at midrange K, for $N = 25000$, $\tau = 100$.

6.2.2 Persistent Mutual Information in the Double Pendulum

The classic study of a planar double pendulum by Shinbrot et al. [1992] demonstrated through physical experiment that minute variations in initial conditions can lead to exponentially diverging orbits. It can be modelled as a continuous time Hamiltonian system with a four-dimensional configuration space, with a rich spectrum of behaviour. Coupled oscillators are of great interest in applied sciences, and thus any insight PMI can provide about the forecastability of individual trajectories can be a bonus.

It also provides a further test of the PMI formalism. In some ways similar to the standard map, with chaos setting in via the break-up of the KAM curves, it is nevertheless a continuous Hamiltonian system with an eight-dimensional joint state space. These factors challenge both our numerical solvers at large τ, and the algorithms for the nearest neighbour search (both of which proved reliable, particularly the kdTree construction that easily adapts to higher dimensionality and periodicities in various dimensions).

Unlike the standard map here there are many ways of even approaching the problem. The Hamiltonian functional partitions the state space so that dynamics are confined to one subset whose topological nature, interestingly enough, changes with their energy value. Some can be similarly partitioned even further.

There are two integrable limits. High energy nullifies the effects of gravity turning the pendula into coupled rotators, whereas low energy does the same with coupling, resulting in two relatively independent systems (in the first case the other conserved quantity is the total angular momentum L, and in the second these are energies of the separate arms). There are therefore two ranges where increasing/decreasing E effectively drives the system to be more chaotic, so in this sense regions of E can be likened to the nonlinearity parameter K from the standard map. Yet here each E comes with a different phase space, of possibly different sizes. An ideal study would couple results (for example PMI) across the energy spectrum with understanding how variations in initial conditions affect predictability, since after all in practice it might be more natural to express uncertainty in terms of a δ in the initial configuration. An example of this type of study is the recent note by Heyl [2008] where the author investigates the subset of the angle Poincaré section defined by the first time a pendulum arm flips. There is a sense in which this framework can be likened to the Divergence diagrams shown earlier for the Tent map. The configuration of these boundaries, naturally dependent on a variety of arbitrary choices and parameters, is shown to be fractal.

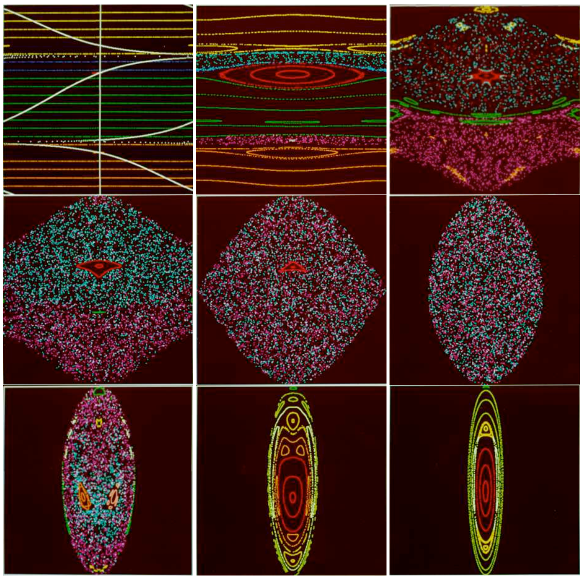


Figure 6.7: The effect of energy on the nature of trajectories. Figure taken from Ohlhoff and Richter [2006]. Poncaré sections of L v the angle of suspension of the first pendulum (see Ohlhoff and Richter [2006] for other specifications). Energy decreasing in the usual order, the top row corresponding to $E = \infty$, $E = 50$ and $E = 10$, the middle to $E = 8$, $E = 6$ and $E = 4$, the bottom row to $E = 2$, $E = 1$ and $E = 0.5$.

Our preliminary work focused on computing PMI for a given energy by having initial data uniformly sampled from the microcanonical ensemble. We used standard parameters that allowed us to reference back to approximate values of E notable for specific behavioural features. This demonstrated the usefulness of PMI in capturing global behaviour. It would make for an interesting project to instead give PMI in terms of some Poincaré variables, obtaining a value that could for example be related to the fractal dimension of the joint and the marginals, since these are implicit in the definition of PMI. Indeed the Matlab code used to evolve the trajectory can readily be adapted to spot crossings of a user-defined plane. Focusing on this would significantly lessen the computational burden on the estimator which currently has to search an eight-dimensional space.

In our work we followed the setup in Ohlhoff and Richter [2006]. The two pendula become unit masses attached to ends of massless rods of unit length, and through further rescaling the original seven parameters become four. Figure 6.7 from Ohlhoff and Richter [2006] shows the effect the decrease in energy has on the trajectories.

As energy is lowered down from ∞ periodic motion ceases to exists and only resonances and quasi-periodic orbits corresponding to irrational winding numbers are left. At about $E = 10$ the last KAM torus breaks down. Chaotic regions spread and at about $E = 4$ the

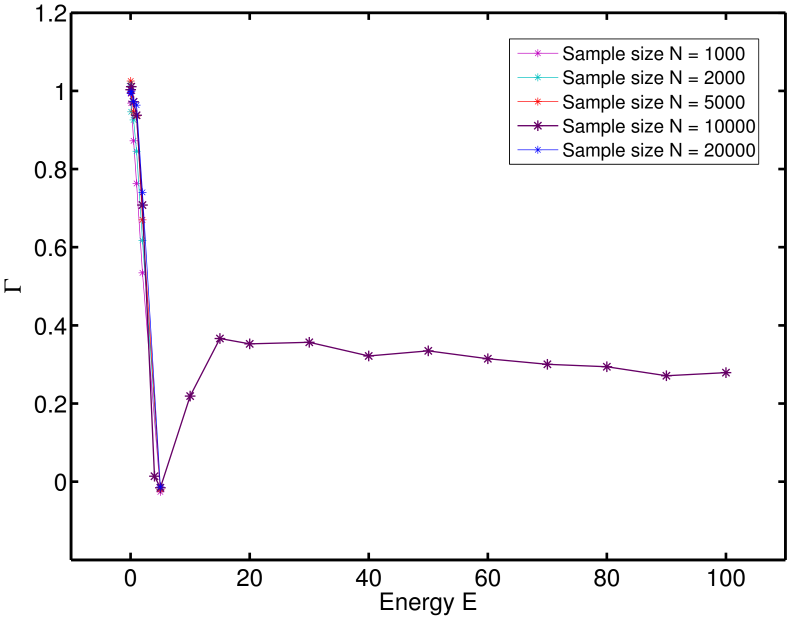


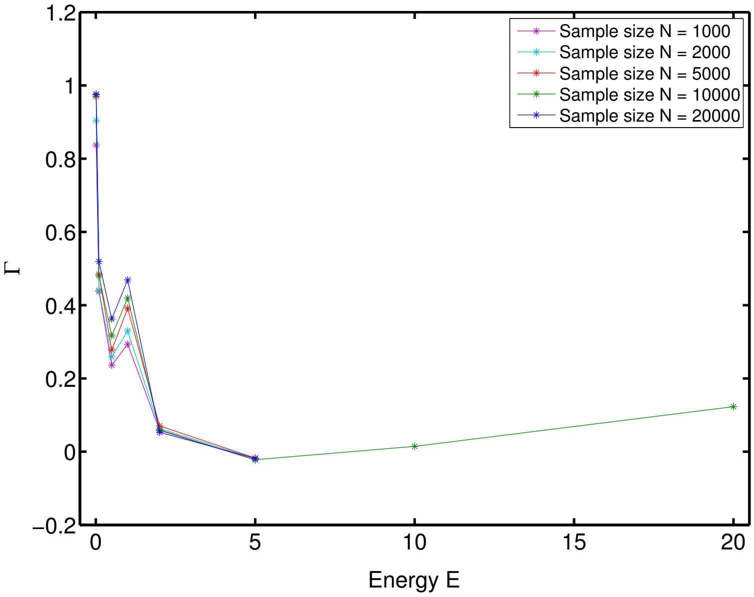
Figure 6.8: Γ v E in the Double Pendulum, for a variety of sample sizes N. Orbits computed using fourth order RungeKutta with both absolute and relative tolerances set to 10^{-6}.

system is fully chaotic and ergodic, at least on resolvable scales. Moving closer to the other integrable limit regular motion begins to once again dominate, especially lower than $E = 1$.

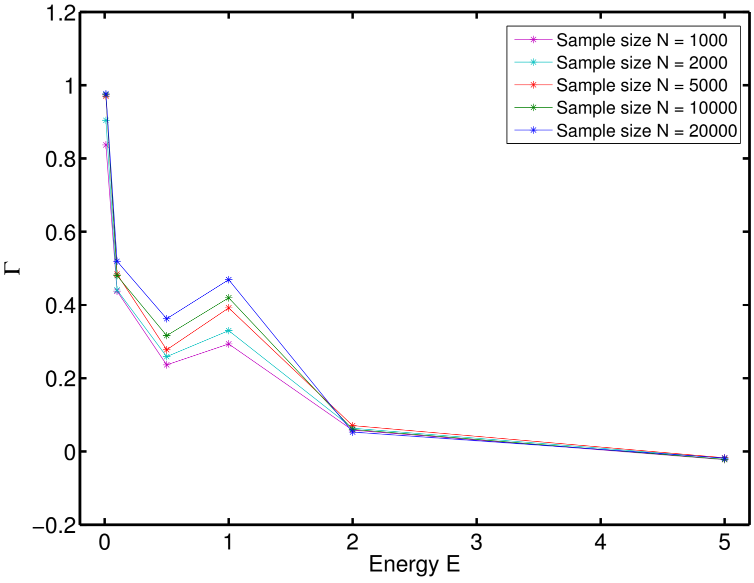
This is exactly what we see for low E when measuring Γ for small τ values. In figure 6.8 Γ is shown to be one for small E, corresponding to the fully-causal system. Plots then decrease to zero at about $E = 4$[1], close to where according to the predictions above the system is ergodic. Then there is an increase and a plateau, which corresponds to some stable rate of change of PMI with resolution for a range of E values. We nevertheless expect Γ to increase back to unity as E goes up. The apparent small decrease here is yet unexplained, and could be either related to the underlying dynamics, or/and to the loss of information as the trajectories drift off the energy shell (the latter can be tested even without doing specialized calculations by simply seeing whether there is an worsening with τ).

Figure 6.9 shows that for large τ, Γ does appear to lower. This process happens both on the high E scale and the low E end. Figure 6.9(b) shows a peak in Γ appearing around $E = 1$. The same might be true for the higher E range, but it was not tested in such detail. Its appearance is certainly reminiscent of a similar feature in the standard map, where we associated the peak with existence of sticky trajectories. Thus PMI seems to suggest that

[1] Γ at $E = 5$ appears to overshoot and give a value slightly smaller than zero. We do not yet know how to interpret this.



(a) $\tau = 800$



(b) $\tau = 800$ zoom in

Figure 6.9: Γ v E in the Double Pendulum, for a variety of sample sizes N.

the same phenomena are present in the double pendulum, at least around $E = 1$.

The variation of Γ with τ is also rich in the variety of behaviour. Figure 6.10 shows results for a variety of energies. The lowest Γ at $E = 5$ is seen to be an overshoot into the negative half-plane, whereas $E = 4$ gives values much closer to the expected zero. Just as in the standard map we see a drastic difference in the speed of convergence. First, energies on both sides of $E = 1$ change their behaviour, whereas at $E = 1$, Γ does not even hint at slowing down. If N is lowered then at $E = 0.5$, Γ appears to briefly plateau at around 0.5, which corresponds to the joint information dimension of 4. It is once again tempting to recall the mixture hypothesis, which conveniently separates trajectories based on the nature of their dynamics, and a marginal D of 2 can then be attributable to chaotic orbits. At larger E the decay of Γ seems to be well described by a power-law with a relatively small exponent.

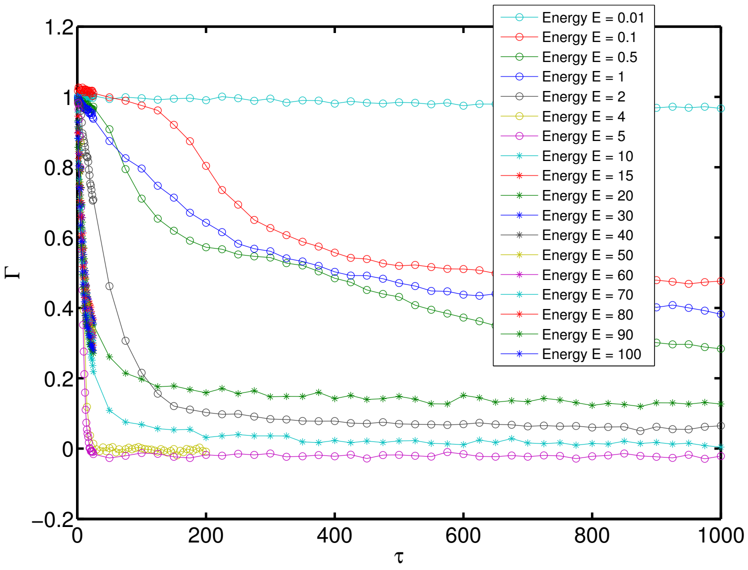


Figure 6.10: Γ v τ in the Double Pendulum, for $N = 10000$. Orbits computed using fourth order Runge-Kutta with both absolute and relative tolerances set to 10^{-6}. Energies below and including $E = 5$ are in circles.

In chapter IV we proposed using the difference between mixing Γ and measured Γ as a measure of entanglement of orbits of different character. Our analysis of the double pendulum, especially the plateaus of Γ, leads us to suppose that by looking at trajectory separations a similar quantity can be found here. If it is made more specific and allowed to focus on particular ranges of the state space it might prove an interesting measure of stability, especially if it is used in the more practical areas like engineering.

Appendix A

Simulating the Double Pendulum

A.1 Setup

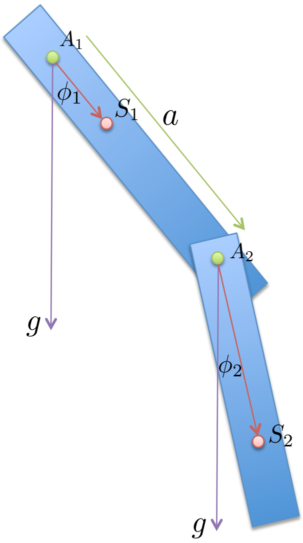


Figure A.1: The inner pendulum with mass m_1 is suspended at point $A1$. $A2$, the point of suspension of the second mass m_2, is on the same plane as the first centre of mass $S1$. This is at angle ϕ_1 with the direction of gravitational pull. The distance between $A1$ and $A2$ is a, and the respective centre of mass $S2$ plane of the outer pendulum is at angle ϕ_2 with the pull.

Following Ohlhoff and Richter [2006] we define a general setup in the following way: There are two conditions attached to making $(\phi_1 = 0, \phi_2 = 0)$ the point of desired stable equilibrium: that s_1, the displacement between $S1$ and $A1$, as well as $m_1 s_1 + m_2 a$ both be positive.

In polar coordinates the Lagrangian reads:

$$
\begin{aligned}
L = \frac{1}{2}\left(\Theta_1 + m_2 a^2\right)\dot{\phi}_1^2 + \frac{1}{2}\Theta_2\dot{\phi}_2^2 + m_2 s_2 a \dot{\phi}_1\dot{\phi}_2 \cos(\phi_2 - \phi 1) \\
-(m_1 s_1 + m_2 a) g (1 - \cos\phi_1) - m_2 s_2 g (1 - \cos\phi_2)
\end{aligned}
\tag{A.1}
$$

Suspension distances are absorbed in the expressions for moments of inertia ($i = 1, 2$):

$$
\Theta_i = \Theta_i^s + m_i s_i^2 \tag{A.2}
$$

At this point we reduce the number of parameters by only focusing on systems where both pendula are suspended by their ends ($\Theta_{1,2}^s = 0$). It is also convenient to stop differentiating between a and the length l_1 of the inner pendulum. The resultant Lagrangian is:

$$
\begin{aligned}
L = \frac{1}{2}\left(\Theta_1 + m_2 a^2\right)\dot{\phi}_1^2 + \frac{1}{2}\Theta_2\dot{\phi}_2^2 + m_2 s_2 a \dot{\phi}_1\dot{\phi}_2 \cos(\phi_2 - \phi_1) \\
-(m_1 s_1 + m_2 a) g (1 - \cos\phi_1) - m_2 s_2 g (1 - \cos\phi_2)
\end{aligned}
\tag{A.3}
$$

Suspension distances are absorbed in the expressions for moments of inertia ($i = 1, 2$):

$$
\Theta_i = \Theta_i^s + m_i s_i^2 \tag{A.4}
$$

Under suitable rescaling the Lagrangian can be expressed in terms of the following dimensionless quantities:

$$
A = \frac{m_1 s_1^2 + m_2 l_1^2}{m_2 s_2^2}, \ \alpha = \frac{l_1}{s_2}, \ \beta = \frac{m_1 s_1 l_1 + m_2 l_1^2}{m_2 s_2^2}. \tag{A.5}
$$

Then the potential energy

$$
V(\phi_1, \phi_2) = \beta(1 - \cos\phi_1) + \alpha(1 - \cos\phi_2) \tag{A.6}
$$

and kinetic energy

$$
T(\phi_1, \phi_2, \dot{\phi}_1, \dot{\phi}_2) = \frac{1}{2}A\dot{\phi}_1^2 + \frac{1}{2}\dot{\phi}_2^2 + \alpha\dot{\phi}_1\dot{\phi}_2\cos(\phi_2 - \phi_1) \tag{A.7}
$$

Standard Scenarios There are two standard scenarios. The first involves solid pendula with an equidistribution of mass along the lengths. Hence $l_i = 2s_i$, and

$$A = \left(\frac{m_1 + 4m_2}{m_2}\right)\frac{l_1^2}{l_2^2},\ \beta = \left(\frac{2m_1 + 4m_2}{m_2}\right)\frac{l_1^2}{l_2^2},\ \alpha = 2\frac{l_1}{l_2}. \tag{A.8}$$

In the case of equal lengths and masses

$$A = 5, \beta = 6, \alpha = 2 \tag{A.9}$$

In the second framework all the mass is concentrated at the end of what are now weightless rods. Here $l_i = s_i$, giving

$$A = \beta = \left(\frac{m_1 + m_2}{m_2}\right)\frac{l_1^2}{l_2^2},\ \alpha = \frac{l_1}{l_2} \tag{A.10}$$

which, in the case of equal masses and lengths, reduces to

$$A = \beta = 2, \alpha = 1 \tag{A.11}$$

We will use the latter scenario throughout.

Hamiltonian of the Double Pendulum

Anticipating our interest in variations of behaviour as a function of total energy we want to write down the Hamiltonian. The Lagrangian is currently in terms of the generalized velocities:

$$L(\phi_1, \phi_2, \dot{\phi}_1, \dot{\phi}_2) = V(\phi_1, \phi_2) + T(\phi_1, \phi_2, \dot{\phi}_1, \dot{\phi}_2). \tag{A.12}$$

These need to be transformed into the associated momenta: $(\dot{\phi}_1, \dot{\phi}_2) \longrightarrow (p_{\phi_1}, p_{\phi_2})$. Since we would need to invert this transformation it is more convenient to work with matrices. In this form the Lagrangian is

$$L = -V(\phi_1, \phi_2) + \frac{1}{2}\dot{\boldsymbol{\phi}} \cdot \mathbf{I}\dot{\boldsymbol{\phi}}, \tag{A.13}$$

where $\dot{\boldsymbol{\phi}} = \begin{pmatrix} \dot{\phi}_1 \\ \dot{\phi}_2 \end{pmatrix}$, and

$$\mathbf{I} = \begin{pmatrix} A & \alpha\cos(\phi_2 - \phi_1) \\ \alpha\cos(\phi_2 - \phi_1) & 1 \end{pmatrix} \tag{A.14}$$

Writing momenta as $\mathbf{p} = \begin{pmatrix} p_{\phi_1} \\ p_{\phi_2} \end{pmatrix}$, the expression

$$p_{\phi_i} = \frac{\partial L}{\partial \dot{\phi}_i} \tag{A.15}$$

becomes

$$\mathbf{p} = \nabla_{\dot{\boldsymbol{\phi}}}\mathbf{L}. \tag{A.16}$$

Differentiating the Lagrangian gives

$$\mathbf{p} = \mathbf{I}\dot{\boldsymbol{\phi}}. \tag{A.17}$$

We can now write the Hamiltonian:

$$\begin{aligned} H = \sum_i \dot{\phi}_i p_{\phi_i} - L &= \dot{\boldsymbol{\phi}} \cdot \mathbf{p} \ + \mathbf{V}(\boldsymbol{\phi_1}, \boldsymbol{\phi_2}) - \frac{\mathbf{1}}{\mathbf{2}}\dot{\boldsymbol{\phi}} \cdot \mathbf{I}\dot{\boldsymbol{\phi}} \\ &= \dot{\boldsymbol{\phi}} \cdot \mathbf{p} + \mathbf{V}(\boldsymbol{\phi_1}, \boldsymbol{\phi_2}) - \frac{\mathbf{1}}{\mathbf{2}}\dot{\boldsymbol{\phi}} \cdot \mathbf{p} \\ &= V(\phi_1, \phi_2) + \frac{1}{2}\dot{\boldsymbol{\phi}} \cdot \mathbf{p}. \end{aligned}$$

The Hamiltonian can now be written in terms of momenta. Using $\dot{\boldsymbol{\phi}} = \mathbf{I^{-1}p}$,

$$H = \beta(1 - \cos\phi_1) + \alpha(1 - \cos\phi_2) + \frac{1}{2}\mathbf{p} \cdot \mathbf{I^{-1}p}, \tag{A.18}$$

where

$$\mathbf{I^{-1}} = \frac{\mathbf{1}}{\mathbf{A} - \alpha^{\mathbf{2}}\cos^{\mathbf{2}}(\boldsymbol{\phi_2} - \boldsymbol{\phi_1})} \begin{pmatrix} 1 & -\alpha\cos(\phi_2 - \phi_1) \\ -\alpha\cos(\phi_2 - \phi_1) & A \end{pmatrix} \tag{A.19}$$

A.2 Mutual Information of the Double Pendulum

Let the state of the system at any given time be described by $x = x(\phi_1, \phi_2, p_{\phi_1}, p_{\phi_2})$. Let $\mathbf{P}_E$ define a set of states with some energy E, $\mathbf{P}_E = \{x : H(x) = E\}$. The evolution of the system under Hamilton's equations of motions is equivalent to movement of x in $\mathbf{P}_E$ under some corresponding evolution operator O^t. Conservation of energy ensures that if at time 0 $x \in \mathbf{P}_{H_0}$, then $O^t x \in \mathbf{P}_{H_0} \forall t > 0$. Thus motion is confined to an 'energy shell' , a 3-dimensional surface in the a 4-dimensional space.

Let thus some energy E define such a surface. Denote by x_{t1} and x_{t2} the states of the system at those time. Then the mutual information between the system at some time 0 in the past and at time t in the future is

$$I[x_{t1}, x_{t2}] = H[x_{t1}] + H[x_{t2}] - H[x_{t1}, x_{t2}]. \tag{A.20}$$

These quantities are functions of measures over the phase spaces. Consider some initial distribution ρ_{t1} on $\mathbf{P}_E$. In some time $t2 - t1$ the evolution operator will evolve this distribution into $\rho_{t2} = O^{t2-t1}\rho_{t1}$. It will also generate a joint distribution $\rho_{t1,t2}$ on $\mathbf{P}_E \times \mathbf{P}_E$. So the mutual information will be change based on the choice of the initial distribution ρ_{t1}. We will be considering the case where ρ_{t1} is flat by sampling the microcanonical ensemble. As such it will be preserved, and the marginal entropies will stay constant. In fact checking the conservation of energy will be a good test of the performance of the particular evolution strategy. Although across varying energy these terms will not be same, since the underlying state space will have different size and other characteristics, this will not matter when looking at Γ since the terms cancel. Therefore there is no need to search marginal spaces (we do it to test entropies are the same, and get positive answers to within the expected errors in all cases). We will therefore look at the PMI by measuring the entropy of the joint using the K-G estimator, the same as was done for the PMI.

A.2.1 Generating Data

Sampling from Energy Shell - the Algorithm

Given an energy E, we want to sample from a flat distribution on the energy shell. This means that the probability of obtaining an x from some subset s of $\mathbf{P}_E$ should be proportional to some natural measure on s. Another way of phrasing this is that given we want N samples from a flat distribution ρ_{flat} on $\mathbf{P}_E$, the latter needs to be split into N subsets of equal weight (which would hence be $\frac{1}{N}$), and that as $N \to \infty$ the counting measure on the sample we obtain should converge to the above.

From the form of eq.(A.18) it is clear the uniform sampling of any three variables will not lead to a distribution that is uniform on the energy shell. We use the following algorithm:

1. Pick ϕ_1 and ϕ_2 randomly, so that $\phi_1 = \alpha_1, \phi_2 = \alpha_2$. Call this point ϕ_α. Accept if $V(\phi_\alpha) \leq E$. Though anticipating the next step this should be strict inequality.
2. Accept further with probability proportional to some natural measure of the set $s_\alpha = \{x \in \mathbf{P}_E : x(1) = \alpha_1, x(2) = \alpha_2\})$
3. if accepted, obtain p_1 and p_2 by uniform sampling of the set s_α

Thus the allowed angles are picked first, and accepted based on the relative weight of the subset of states which have that potential energy (a function of those angles), and subsequently a certain kinetic energy. Note that this is not the same as uniform sampling of potential energy with subsequent acceptance/rejection. Doing the latter would misrepresent the distribution of the underlying arguments (the angles).

The second step in the sampling algorithm involves a measure of a set of states with a given ϕ_α. Call this number $W(\phi_1, \phi_2)$. α also defines a kinetic energy T. It can be written as

$$W(\phi_1, \phi_2) = \int dp_1 dp_2 \, \delta\left(\frac{1}{2}\mathbf{p} \cdot \mathbf{I^{-1}p} - \mathbf{T}\right) \tag{A.21}$$

Let $\mathbf{q} = \mathbf{I}^{-\frac{1}{2}}\mathbf{p}$. Then

$$\mathbf{p} \cdot \mathbf{I^{-1}p} = \mathbf{p} \cdot \mathbf{I}^{-\frac{1}{2}}\mathbf{q} = \mathbf{I}^{-\frac{1}{2}}\mathbf{p} \cdot \mathbf{q} = \mathbf{q^2}, \tag{A.22}$$

where the second to last equality follows from the fact that $\mathbf{I^{-1}}$ is symmetric, hence $\left(\mathbf{I}^{-\frac{1}{2}}\right)^{\mathrm{T}} = \mathbf{I}^{-\frac{1}{2}}$. Since the transformation is linear the Jacobian matrix is just $\mathbf{I}^{-\frac{1}{2}}$, and we can use the properties of determinants to see that

$$W(\phi_1, \phi_2) = \int dq_1 dq_2 \, |\mathbf{I}|^{-\frac{1}{2}} \; \delta\left(\frac{1}{2}\mathbf{q}^2 - T\right) \tag{A.23}$$

This has the form of an integral over a circle of radius $r = \sqrt{2T}$. Changing to polar coordinates the integral becomes

$$W(\phi_1, \phi_2) = \int r \; dr \; d\theta \, |\mathbf{I}|^{-\frac{1}{2}} \; \delta\left(\frac{1}{2}r^2 - T\right) \tag{A.24}$$

Let $q = \frac{1}{2}r^2$. Then $dq = r\ dr$, and

$$W(\phi_1, \phi_2) = \int dq\ d\theta \left|\mathbf{I}\right|^{-\frac{1}{2}}\ \delta\left(q - T\right) \tag{A.25}$$

The weight of this ellipse is thus $2\pi \left|\mathbf{I}\right|^{-\frac{1}{2}}$, and it is defined by having a radius that of $\sqrt{2T}$. The first fact will be used in computing the probability of acceptance of a given set of angles. The second will help sample uniformly from this set.

Optimal sampling is achieved if the set of allowed angles is accepted with a probability correctly normalised its maximum:

$$p_{acpt} = \frac{W(\phi_1, \phi_2)}{W_{\max}}. \tag{A.26}$$

From (A.14), $\left|\mathbf{I}\right|^{-\frac{1}{2}} = \left(A - \alpha^2 \cos^2(\phi_2 - \phi_1)\right)^{-\frac{1}{2}}$ is maximal when $\phi_1 = \phi_2$, giving $W_{\max} = 2\ \pi(A - \alpha^2)^{-\frac{1}{2}}$. Hence,

$$p_{acpt} = \sqrt{\frac{A - \alpha^2}{A - \alpha^2 \cos^2(\phi_2 - \phi_1)}}. \tag{A.27}$$

The last stage of the algorithm requires p_1 and p_2 be sampled uniformly from the manifold characterised by $\frac{1}{2}\mathbf{q^2} = \mathbf{T}$. As mentioned above, in (q_1, q_2) coordinates this is a circle of radius $\sqrt{2T}$. Uniform sampling on a circle can be achieved by picking $\theta \sim U(0, 2\pi]$, giving $q_1 = \sqrt{2T}\cos(\theta)$ and $q_2 = \sqrt{2T}\sin(\theta)$. Momenta is obtained by $\mathbf{p} = \mathbf{I^{\frac{1}{2}}q}$.

Simulating the Motion

If D is the determinant of the matrix $\mathbf{I}$, then the Hamiltonian is

$$H = \frac{1}{2\ D}(p_\theta^2 - 2\alpha p_\theta p_\phi \cos(\phi - \theta) + A\ p_\phi^2) + \beta(1 - \cos\theta) + \alpha(1 - \cos\phi). \tag{A.28}$$

The momenta evolve according to

$$\dot{\mathbf{p}}_{\mathbf{i}} = -\frac{\partial H}{\partial \phi_i} = -\frac{\partial V}{\partial \phi_i} - \frac{\partial T}{\partial \phi_i} \tag{A.29}$$

Partial derivatives of V are straightforward, and for contribution coming from the kinetic energy we can write

$$\frac{\partial T}{\partial \phi_i} = \frac{1}{2}\mathbf{p}\cdot\frac{\partial \mathbf{I^{-1}}}{\partial \boldsymbol{\phi_i}}\mathbf{p} = \frac{\mathbf{1}}{\mathbf{2D^2}}\mathbf{p}\cdot\left(\mathbf{D}\frac{\partial \mathbf{I^{adj}}}{\partial \boldsymbol{\phi_i}} - \mathbf{I^{adj}}\frac{\partial \mathbf{D}}{\partial \boldsymbol{\phi_i}}\right)\mathbf{p} \tag{A.30}$$

Hence, if we use θ and ϕ to denote the respective components of $\boldsymbol{\phi}$, and write Δ_ϕ for $\phi - \theta$, then Hamilton's equations of motion are:

$$\dot{\theta} = \frac{\partial H}{\partial p_\theta} = \frac{1}{D}\left[p_\theta - \alpha p_\phi \cos\Delta_\phi\right] \tag{A.31}$$

$$\dot{\phi} = \frac{\partial H}{\partial p_\phi} = \frac{1}{D}\left[A p_\phi - \alpha p_\theta \cos\Delta_\phi\right] \tag{A.32}$$

$$\dot{p}_\theta = -\frac{\partial H}{\partial \theta} = -\beta\sin\theta - \frac{1}{D^2}\left[\alpha^2 \sin\Delta_\phi \cos\Delta_\phi (p_\theta^2 + A p_\phi^2) - \alpha p_\theta p_\phi \sin\Delta_\phi (A + \alpha^2\cos^2\Delta_\phi)\right] \tag{A.33}$$

$$\dot{p}_\phi = -\frac{\partial H}{\partial \phi} = -\alpha\sin\phi + \frac{1}{D^2}\left[\alpha^2 \sin\Delta_\phi \cos\Delta_\phi (p_\theta^2 + A p_\phi^2) - \alpha p_\theta p_\phi \sin\Delta_\phi (A + \alpha^2\cos^2\Delta_\phi)\right]. \tag{A.34}$$

Using the Matlab platform

In Press et al. [2002] there is a readily available code that allows for different ODE solving schemes. Although C++ is undoubtedly faster, there are several reasons why we start with Matlab:

a) we can use a pre-built ODE solver (in this case ode45, a fourth-order Runge-Kutta method), which is faster to implement. b) it is easier to visualise the motion (no need to transfer data between programs), and this can be useful to gauge the 'shape' of distributions and the underlying phase space. We can also get estimates of the timescales of motion (i.e. by plotting variations of angles with time), which are needed to understand reasonable timescale to model large τ for persistent mutual information.

As with any scheme, iterations will lose accuracy, and motion may drift off the energy shell. We can observe this on an energy v time plot (we expect this to depend on the energy itself but also on the region of the shell, since for most energies motions of different type coexist). Note that being with some error on the energy shell does not imply being *accurate* with the same error. In terms of speed things can be improved by using the Lagrangian framework to evolve the points instead of the Hamiltonian. The current computation time is $\mathbf{O}(\tau N)$.

Errors

Naively the cumulative error is of order (number of steps*error in each step). But this is only true if the distances between nearby points, whether on or off the energy shell, are preserved by the flows. If that doesn't happen, then the true point and the approximated point can be the starting points of two diverging trajectories. The approximated point can flow further away from the energy shell, and also at a different, possibly higher, rate. There are two types of motion to consider here - that of nearby trajectories *on* the energy shell, and *off* it. The character of the divergence of trajectories on the energy shell will go hand in hand with whether the system (and the region of the energy shell) is in the chaotic phase or not. The difference in the motion between the different energy shells is partially to do with the way the shells are layered in space.

Without finding one or another signature of chaos we cannot speak of divergence of trajectories on the energy shell. But we can calculate the Hamiltonian function of the estimated coordinates, which will correspond to the total energy of a system that contains the point $\hat{x}_T$.

Hence the accuracy of the method can be tested by seeing how far solutions move off the energy shell. Let the difference between the starting energy (energy of the system) and the value of the Hamiltonian function of a point x_t obtained by our numerical solver, i.e. $x_t = \Psi^t_H(x_0)$ as before, be

$$\eta^t = H\left(\Psi^t_H(x_0)\right) - E, \tag{A.35}$$

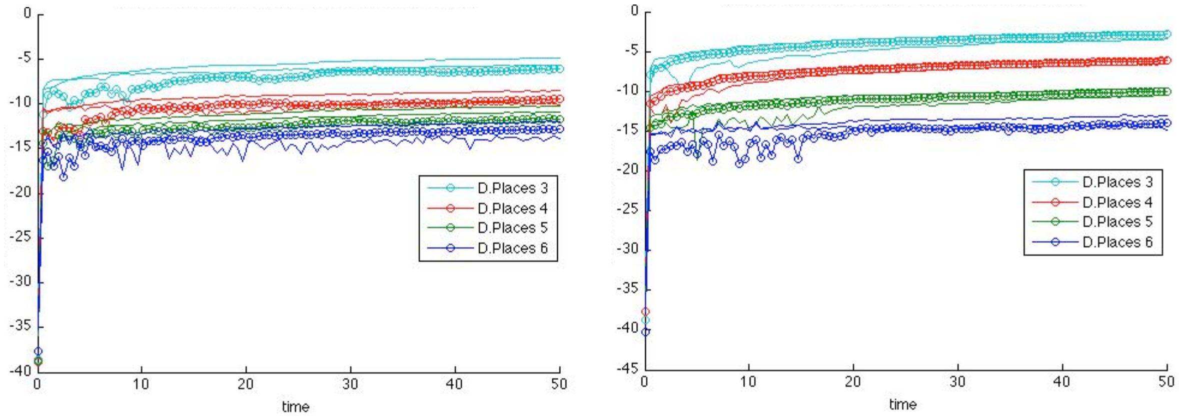
and the relative error

$$\eta_{t_{rel}} = \frac{H\left(\Psi^t_H(x_0)\right) - E}{E}. \tag{A.36}$$

This will depend on:

1. The starting point of the flow, which we take to signal time 0: x_0. This dependency reflects the possibility that the flow is not homogenous, so that the number of discretization steps needed to approximate different trajectories may vary.

2. The energy shell, as labelled by the value (E) of the Hamiltonian for points on that shell. Trajectories will drift off the shell and their behaviour will be partially determined by the structure of the shells in the phase space.



(a) $E = 1$

(b) $E = 5$ zoom in

Figure A.2: Logarithm base 10 of the average relative error in energy, when both tolerance levels in the fourth order Runge-Kutta method are kept at some number of decimal places. Average is over 100 runs whose initial conditions are distributed according to the microcanonical ensemble.

3. The two Runge-Kutta parameters controlling the accuracy of the solver, the absolute and relative tolerance levels, a and r.

4. The solver algorithm, i.e. Ψ. In this section this is assumed to be the fourth order Runge-Kutta.

So the error can be written as

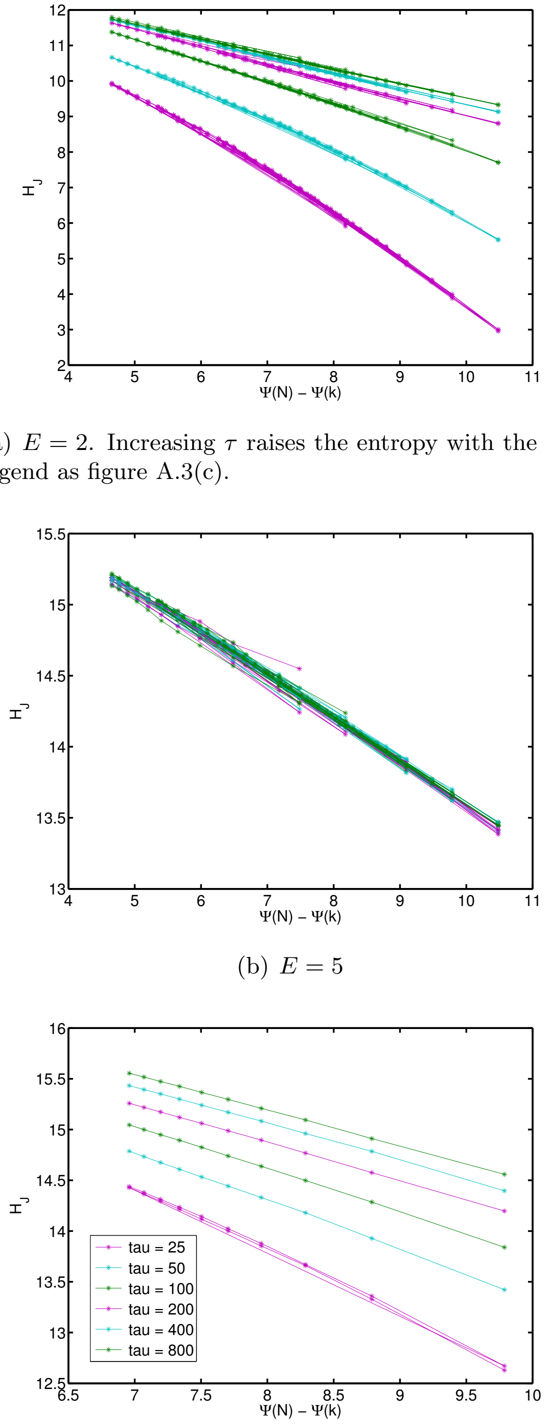
$$\eta^t = \eta^t(x_0, a, r). \tag{A.37}$$

It was found that error levels are tolerable. Figure A.2 shows that they can be minimized at the expense of the running time.

Other graphs for different E value display the same broad character, though the relative positions of the plots may differ.

We used the K-G algorithm to find the entropy of the joint distribution. The kdTree class had to be adapted to admit a metric on eight-dimensional cylindrical spaces.

Figure A.3 shows that joint entropy first of all still scales with $\delta\Psi$, and second that are regions where that scaling is linear. Notice in figure A.3(b) how all the plots have converged to $\Gamma = 0$ even for the minimum τ considered. Therefore there is some justification in trusting the iterative process.



(a) $E = 2$. Increasing τ raises the entropy with the same legend as figure A.3(c).

(b) $E = 5$

(c) $E = 10$

Figure A.3: Entropy of the Joint as a function of resolution for several energies of the double pendulum. Note that to get from the rate of change to Γ one needs to normalize, since the marginals distributions do not fully occupy the four-dimensional marginal spaces.

Bibliography

Roy L. Adler, A. G. Konheim, and M. H. McAndrew. Topological entropy. *Transactions of the American Mathematical Society*, 114(2):309–319, 1965.

Garin F. J. Añaños, Fulvio Baldovin, and Constantino Tsallis. Anomalous sensitivity to initial conditions and entropy production in standard maps: nonextensive approach. *The European Physical Journal B*, 46(3):409–417, 2005.

Paul W. Anderson. More is different: broken symmetry and the nature of the hierarchical structure of science. *Science*, 177:393–396, 1972.

Arthur Koestler and John R. Smythies (editors). *Beyond reductionism: New perspectives in the life sciences: proceedings of the Alpbach Symposium 1968*, volume 1. Hutchinson, 1968.

Henri Atlan. *Entre le Cristal et la Fumée: Essai sur l'Organisation du Vivant.* Points Sciences. Éditions du Seuil, 1986.

Remo Badii and G. Broggi. Measurement of the dimension spectrum $f(\alpha)$: fixed-mass approach. *Physics Letters A*, 131(6):339–343, 1988.

Fulvio Baldovin, Constantino Tsallis, and B. Schulze. Nonstandard entropy production in the standard map. *Physica A*, 320:184–192, 2003.

Robin C. Ball, Marina Diakonova, and Robert S. MacKay. Quantifying emergence in terms of persistent mutual information. *Advances in Complex Systems*, 13(3):327–338, 2010.

Vincent Bauchau. Emergence and reductionism: from the game of life to science of life. In *Self-Organisation and Emergence in Life Sciences*, pages 29–40. Springer Netherlands, 2006.

Mark A. Bedau. Weak emergence. *Noûs*, 31:375–399, 1997.

Charles H. Bennett. Logical depth and physical complexity. In *The Universal Turing Machine: A Half-Century Survey*, pages 227–257. Oxford University Press, Oxford, 1988.

David Bensimon and Leo P. Kadanoff. Extended chaos and disappearance of KAM trajectories. *Physica D*, 13(1–2):82–89, 1984.

Pratip Bhattacharyya and Bikas K. Chakrabarti. The mean distance to the nth neighbour in a uniform distribution of random points: an application of probability theory. *European Journal of Physics*, 29(3):639–645, 2008.

R. C. Black and I. I. Satija. Recurrence of Kolmogorov-Arnold-Moser tori and fractal diagram. *Physical Review A*, 40(5):2864–2867, 1989.

Rufus Bowen. Entropy for group endomorphisms and homogeneous spaces. *Transactions of the American Mathematical Society*, 153:401–414, 1971.

Charlie D. Broad. *The Mind and its Place in Nature.* Routledge, 1925. Volume 3 of the *International Library of Philosophy Series.*

David K. Cambell. Hamiltonian chaos and statistical mechanics. In *Nonlinear science: from paradigms to practicalities*, pages 242–245. Los Alamos Science special issue, 1987.

Gregory J. Chaitin. Randomness and mathematical proof. *Scientific American*, 232(5): 47–52, 1975.

David J. Chalmers. Strong and weak emergence. In *The Re-Emergence of Emergence*, pages 244–256. Oxford University Press, 2006.

Subrahmanyan Chandrasekhar. Stochastic problems in physics and astronomy. *Reviews of Modern Physics*, 15(1):1–89, 1943.

Boris V. Chirikov. Resonance processes in magnetic traps. *Atomic Energy*, 6:464–470, 1960.

Boris V. Chirikov. A universal instability of many-dimensional oscillator systems. *Physics Reports*, 52:263–379, 1979.

Boris V. Chirikov and Dmitry L. Shepelyansky. Asymptotic statistics of Poincaré recurrences in Hamiltonian systems with divided phase space. *Physical Review Letters*, 82: 528–531, 1999.

Philip Clayton. Conceptual foundations of emergence theory. In *The Re-Emergence of Emergence*, pages 1–34. Oxford University Press, 2006.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, 2006.

James P. Crutchfield and Norman H. Packard. Symbolic dynamics of noisy chaos. *Physica D*, 7(1-3):201–223, 1983.

James P. Crutchfield and Karl Young. Inferring statistical complexity. *Physical Review Letters*, 63:105–108, 1989.

Paul Davies. Emergent biological principles and the computational properties of the universe: explaining it or explaining it away. *Complexity*, 10(2):11–15, 2004.

Persi Diaconis, Susan Holmes, and Richard Montgomery. Dynamical bias in the coin toss. *SIAM Review*, 49(2):211–235, 2007.

Marina Diakonova and Robert S. MacKay. Mathematical examples of space-time phases. *International Journal of Bifurcation and Chaos*, 21(8):2297–2304, 2011.

Efim I. Dinaburg. On the relations among various entropy characteristics of dynamical systems. *Mathematics of the USSR-Izvestiya*, 5(2):337–378, 1971.

Ionas Erb and Nihat Ay. Multi-information in the thermodynamic limit. *Journal of Statistical Physics*, 115:949–976, 2004.

J. Doyne Farmer, James P. Crutchfield, Harold Froehling, Norman Packard, and Robert Shaw. Power spectra and mixing properties of strange attractors. *Annals of the New York Academy of Sciences*, 357(1):453–471, 1980.

David P. Feldman and James P. Crutchfield. Discovering noncritical organization: statistical mechanical, information theoretic, and computational views of patterns in one-dimensional spin systems. Technical Report 98-04-026, SFI Working Paper, 1998.

David P. Feldman and James P. Crutchfield. Structural information in two-dimensional patterns: entropy convergence and excess entropy. *Physical Review E*, 67:051104, 2003.

David P. Feldman, Carl S. McTague, and James P. Crutchfield. The organization of intrinsic computation: complexity-entropy diagrams and the diversity of natural information processing. *Chaos*, 18(4):043106, 2008.

Tobias Galla and Otfried Gühne. Complexity measures, emergence, and multiparticle correlations. *Physical Review E*, 85:046209, 2012.

Ronald Getoor. J. L. Doob: foundations of stochastic processes and probabilistic potential theory. *Annals of Probability*, 37(5):1647–1663, 2009.

Peter Gmeiner. *Properties of persistent mutual information and emergence.* PhD thesis, Friedrich-Alexander-Universitat Erlangen-Nurnberg, Cauerstr. 11, 91058 Erlangen, Germany, 2012.

Jeffrey Goldstein. Emergence as a construct: history and issues. *Emergence*, 1(1):49–72, 1999.

Philippe Goujon. From logic to self-organization - learning about complexity. In *Self-Organisation and Emergence in Life Sciences*, pages 187–214. Springer Netherlands, 2006.

Peter Grassberger. Generalizations of the Hausdorff dimension of fractal measures. *Physics Letters A*, 107(3):101–105, 1985.

Peter Grassberger. Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics*, 25:907–938, 1986.

Peter Grassberger. An optimized box-assissted algorithm for fractal dimensions. *Physics Letters A*, 148(1–2), 1990.

Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. *Physica D*, 9(1–2):189–208, 1983a.

Peter Grassberger and Itamar Procaccia. Characterization of strange attractors. *Physical Review Letters*, 50:346–349, 1983b.

Peter Grassberger, Remo Badii, and Antonio Politi. Scaling laws for invariant measures on hyperbolic and nonhyperbolic atractors. *Journal of Statistical Physics*, 51:135–178, 1988.

Niels H. Gregersen. God, matter, and information : towards a stoicizing logos christology. In *Information and the Nature of Reality: From Physics to Metaphysics*, pages 319–348. Cambridge University Press, 2010.

Thomas C. Halsey, Mogens H. Jensen, Leo P. Kadanoff, Itamar Procaccia, and Boris Shraiman. Fractal measures and their singularities: the characterization of strange sets. *Physical Review A*, 33:1141–1151, 1986.

Carl G. Hempel and Paul Oppenheim. Studies in the logic of explanation. *Philosophy of Science*, 15(2):135–175, 1948.

Jeremy S. Heyl. The double pendulum fractal. Unpublished, 2008. URL `http://tabitha.phas.ubc.ca/wiki/images/3/37/Double.pdf`.

Bernardo A. Huberman and Tad Hogg. Complexity and adaptation. *Physica D*, 22:376–384, 1986.

Henrik J. Jensen and Fatimah A. Razal. Private Communication, 2012.

Thomas Kahle, Eckehard Olbrich, Jrgen Jost, and Nihat Ay. Complexity measures from interaction structures. *Physical Review E*, 79(2):026201, 2009.

Jaegwon Kim. Making sense of emergence. *Philosophical Studies*, 95:3–36, 1999.

Jaegwon Kim. Emergence: Core ideas and issues. *Synthese*, 151:547–559, 2006.

Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *Problemu Peredachi Informatsii*, 1(1):3–11, 1965.

Andrei N. Kolmogorov and Vladimir M. Tihomirov. ε-entropy and ε-capacity of sets in function spaces. *Yspehi Matematicheskih Nauk*, 14:3–86, 1959.

L. F. Kozachenko and N. N. Leonenko. On statistical estimation of entropy of random vector. *Problemu Peredachi Informatsii*, 23(2):95–101, 1987.

Alexander Kraskov, Harald Stogbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6, Part 2):066138, 2004.

Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean, and Andrew Y. Ng. Building high-level features using large scale unsupervised learning. In *ICML 2012: 29th International Conference in Machine Learning*, 2012.

W. Li. On the relationship between complexity and entropy for Markov chains and regular languages. *Complex Systems*, 5(4):381–399, 1991.

Chuanwen Luo, Chundi Yi, Gang Wang, Longsuo Li, and Chuncheng Wang. The mathematical description of uniformity and related theorems. *Chaos, Solitons and Fractals*, 42 (5):2748–2753, 2009.

Robert S. MacKay. A renormalization approach to invariant circles in area-preserving maps. *Physica D*, 7(1–3):283–300, 1983.

Robert S. MacKay. Greene's residue criterion. *Nonlinearity*, 5(1):161–187, 1992.

Robert S. MacKay. Robustness of Markov processes on large networks. *Journal of Difference Equations and Applications*, 17(8):1155–1167, 2009.

Robert S. MacKay and I. C. Percival. Converse KAM - theory and practice. *Communications in Mathematical Physics*, 98(4):469–512, 1985.

Robert S. MacKay, James D. Meiss, and I. C. Percival. Stochasticity and transport in Hamiltonian systems. *Physical Review Letters*, 52:697–700, 1984.

Walter T. Marvin. *A first Book in Metaphysics.* New York : Macmillan, 1912.

John N. Mather. Variational construction of orbits of twist diffeomorphisms. *Journal of the American Mathematical Society*, 4(2):207–263, 1991.

Hiroyuki Matsuda, Kiyoshi Kudo, Ryoku Nakamura, Osamu Yamakawa, and Takuo Murata. Mutual information of Ising systems. *International Journal of Theoretical Physics*, 35: 839–845, 1996.

Robert M. May. Simple mathematical models with very complicated dynamics. *Nature*, 261(5560):459–467, 1976.

Paul McGarr. Engels and natural science. *International Socialism Journal*, 65, 1994.

James D. Meiss. The standard map. In *Encyclopedia of Nonlinear Science*, pages 870–873. New York, Routledge, 2005.

A. Ohlhoff and P. H. Richter. Forces in the double pendulum. *Zeitschrift fr Angewandte Mathematik und Mechanik*, 80:517–534, 2006.

Lael Parrott. Measuring ecological complexity. *Ecological Indicators*, 10(6):1069–1076, 2010.

William Parry. Intrinsic Markov chains. *Transactions of the American Mathematical Society*, 112(1):55–66, 1964.

Paul Davies and Niels H. Gregersen (editors). *Information and the Nature of Reality.* Cambridge University Press, 2010.

Arthur Peacocke. Sciences of complexity : a new theological resource? In *Information and the Nature of Reality: From Physics to Metaphysics*, pages 249–281. Cambridge University Press, 2010.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C++: The Art of Scientific Computing.* Cambridge University Press, 2002.

Mikhail Prokopenko, Fabio Boschetti, and Alex J. Ryan. An information-theoretic primer on complexity, self-organization, and emergence. *Complexity*, 15(1):11–28, 2009.

Alfred Renyi. *Letters on Probability.* Akadémiai Kiadó, 1972.

John Rickert and Aaron Klebanoff. Studying the Cantor dust at the edge of Feigenbaum diagrams. *The College Mathematics Journal*, 29(3):189–198, 1999.

Cosma R. Shalizi and Kristina L. Shalizi. Blind construction of optimal nonlinear recursive predictors for discrete sequences. In *Proceedings of the 20th conference on Uncertainty in Artificial Intelligence*, 2004.

Cosma R. Shalizi, Kristina L. Shalizi, and Robert Haslinger. Quantifying self-organization with optimal predictors. *Physical Review Letters*, 93:118701, 2004.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423 and 623–656, 1948.

Troy Shinbrot, Celso Grebogi, Jack Wisdom, and James A. Yorke. Chaos in a double pendulum. *American Journal of Physics*, 60(6):491–499, 1992.

Yakov G. Sinai. Some problems in the theory of dynamical systems mathematical physics. In *Visions in Mathematics*, Modern Birkhuser Classics, pages 425–433. Birkhuser Basel, 2010.

David Singer. Stable orbits and bifurcation of maps of the interval. *SIAM Journal on Applied Mathematics*, 35(2):260–267, 1978.

Milan Studeny and Jirina Vejnarová. The multi-information function as a tool for measuring stochastic dependence. In *Learning in graphical models*, pages 261–298. Kluwer Academic Publishers, 1998.

James Theiler. Estimating fractal dimension. *Journal of the Optical Society of America*, 7 (6):1055–1073, 1990.

Henrik Thorén and Philip Gerlee. Weak emergence and complexity. In *Artificial Life XII: Proceedings of the Twelfth International Conference on the Synthesis and Simulation of Living Systems*, pages 879–886. The MIT Press, 2010.

David K. Umberger and J. Doyne Farmer. Fat fractals on the energy surface. *Physical Review Letters*, 55(7):661–664, 1985.

Alper Üngör. Computational geometry, 2013. URL `http://www.cise.ufl.edu/class/cot5520fa09/CG_RangeKDtrees.pdf`. web tutorial.

Warren Weaver. Science and complexity. *American Scientist*, 35:536–544, 1948.

Roscoe B. White, Charles F. F. Karney, and Alexander B. Rechester. Effect of noise on the standard mapping. In *International Symposium on Physical Design*, 1981.

Stephen Wolfram. Computational theory of cellular automata. *Communications in Mathematical Physics*, 96:15–57, 1984.

Domenico Zambella and Peter Grassberger. Complexity of forecasting in a class of simple models. *Complex Systems*, 2:269–303, 1988.

George M. Zaslavsky. *The Physics of Chaos in Hamiltonian Systems.* Imperial College Press, 2012.