Some Foundational Tools and Concepts for Complex Systems:

Entropy, Information, Computation, and Complexity.

2008 Complex Systems Summer School

Santa Fe Institute

Institute of Theoretical Physics, Chinese Academy of Science

David P. Feldman

College of the Atlantic

and

Santa Fe Institute

dave@hornacek.coa.edu

http://hornacek.coa.edu/dave/

List of references: `http://www.citeulike.org/user/dpf`

Contents

Part I

Introduction and Motivation

Introduction to the CSSS

- This is the 5$^{\mathrm{th}}$ CSSS in China.
- The Santa Fe Institute has held CSSSs in Santa Fe since 1988.
- Three main goals for the Beijing CSSS:
  1. Give students an introduction to some of the tools and methods of Complex Systems
  2. Give students experience working in interdisciplinary, collaborative teams.
  3. Give students experience working in international collaborations.
- The first goal is met by **lectures**.
- The second and third goals are met through **student projects**.

Thoughts on the Lectures

- There will be times when the lectures seem too slow, and other times when they seem too fast.
- This is the nature of interdisciplinary work. We have many different academic backgrounds, so not every lecture can appeal to everyone.
- Please, ask questions during and after the lectures.
- Please feel free to set up meetings with faculty.
- At the end of the CSSS we will provide you with a CD containing the slides of all the lectures.
- We will try to provide you with hard copies of slides before each lecture.

Projects

- You will do a group project as part of the CSSS. As part of this your group will:

  1. Give a 20 minute presentation
  2. Prepare a scientific poster for the final banquet
  3. Write a short paper or technical report

- Teams must be between three and six people, should be interdisciplinary, and, if possible, should contain Chinese and foreigners.
- The *process* of doing the project is at least as important as the *product*.
- Later today I will give you a handout with more information about projects.
- This afternoon all students will give a short introduction to help begin the process of dividing into groups

Other CSSS Activities

- There will be a few activities designed to help us get to know each other better, learn about China, and learn about each other's cultures.
- There will be several round table discussions:
  - Advice for interdisciplinary research
  - Advice for Chinese students wishing to study in Europe and North America
  - Other topics depending on student and faculty interest
- During weeks two and three there will be some student-led workshops on topics of your choosing.

Other CSSS Thoughts and Advice

- All students at the CSSS are teachers as well as learners.
- The most interesting and valuable parts of the summer school will probably be the discussions you have with other students and the work you do in groups. The lectures are just the starting point.
- The CSSS is an amazing opportunity to interact with people you might not normally interact with. Take advantage of this.
- Don’t spend too much energy worrying about the definition of complexity or complex systems. (You wouldn’t go to a history conference and spend a month debating what history is.)
- Working across disciplines can be challenging. Different disciplines have different vocabularies and underlying assumptions. Be patient—it’s worth it.
- If you’re from away, be sure to spend some time exploring Beijing.
- Pace yourself. It can be a long month. Don’t have too much fun all at once.

A Very Little Bit about the Santa Fe Institute

- Founded in 1984, located in Santa Fe, New Mexico, USA.
- Research is collaborative and interdisciplinary.
- Emphasis is on complex systems in the physical, natural, computational, and social sciences.
- Interdisciplinary research and education center. No academic departments.
- It’s normal to be “weird” and interdisciplinary. You won’t be asked “But is that physics?” (or biology, economics, etc.)
- For much more, see `http://www.santafe.edu` and ask SFI postdocs, faculty, and external faculty who are speaking at the CSSS.

Introduction to my Series of Lectures

- I will present a number of different techniques for measuring and quantifying different sorts of randomness, unpredictability, structure, organization, complexity, and so on.
- I have two broad goals for my lectures:
  1. (Re)introduce you to some tools and methods that you might be able to adapt for use in your research.
  2. Critically explore conceptual questions about unpredictability, order, complexity, etc.
- These notes are based on lectures I have given at the CSSS since 2004 and at the First Annual French Complex Systems Summer School, Institut des Systèmes Complex de Paris Ile-de-France, August 2007.

Other Thoughts on my Lectures

- The first few days will be broad introductions to several different fields. The last few days will be more advanced, will explore bigger questions, and will draw upon what we've done in the first few days.
- My aim is to give a somewhat selective and opinionated survey of tools, methods, and ideas from complex systems.
- We can't cover anything, so I've chosen what I think is most important and what you're less likely to get elsewhere.
- I expect to cover around two thirds of the slides I've included in this handout.

Goals

1. Present some tools, models, paradigms that are useful in complex systems.
2. Discuss the applicability and un-applicability of these various tools.
3. Provide references and advice so you can learn more about these topics if you wish.
4. Present some thoughts about what makes the study of complex systems similar to, and different from, other types of science.
5. Provide some background which may help you get more out of other lectures.
6. Have fun.

Other Comments about my Series of Lectures

1. I am trained as a theoretical physicist. So my approach will be somewhat physics-centric.
2. There is a lot of material that I'm trying to cover, so I'll move quickly and at times will skip some topics that I really shouldn't skip.
3. Please, ask questions during and after the lectures.
4. I'd also welcome comments during my lectures; you may be more of an expert than me on some of the topics.
5. I will make these slides available on my website and the CSSS wiki.

Before starting the first lecture on chaos, some big issues to help frame the questions I'm interested in.

What are Complex Systems?

I'm not interested in a strict definition of complex systems. However, it seems to me that most things we'd think of as a complex system share many of the following features:

1. Unpredictability. A perfectly predictive theory is rarely possible.
2. Emergence: Systems generate patterns that are not part of the equations of motion: *emergent phenomena*.
3. Interactions: The interactions between a system's components play an important role.
4. Order/Disorder: Most complex systems are simultaneously ordered and disordered.
5. Heterogeneity: Not all the elements that make up the system are identical.
6. Adaptive or Dynamic: System properties change over time.

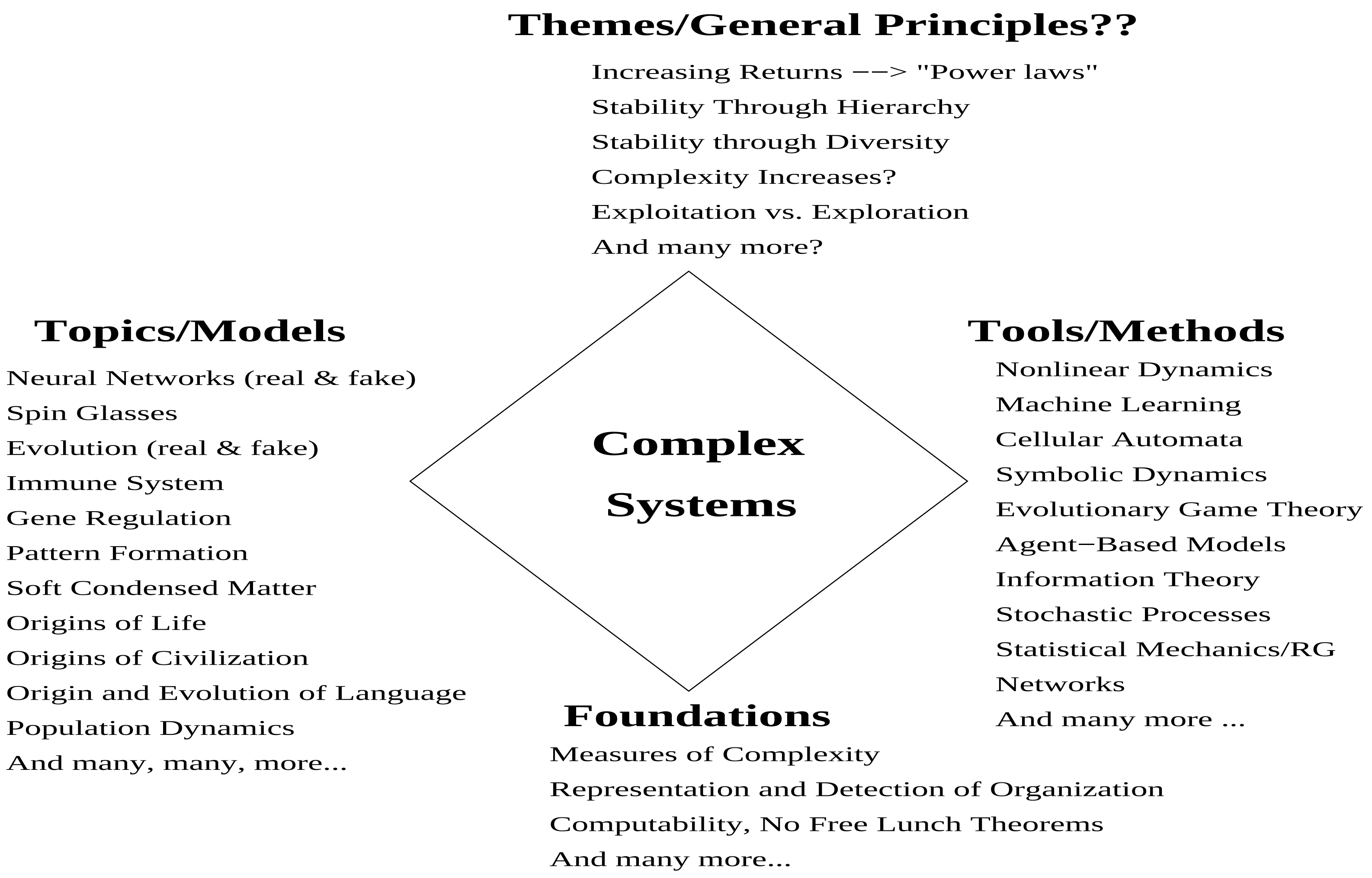
Phenomena and Topics

- Another way to approach a definition of complex systems is to list the things that people think are complex systems:
  - Immune system, ecosystems, economies, auction markets, evolutionary systems, the brain, natural computation,
- Or, one can think about the tools and models that people use to study the things that people think are complex systems:
  - Machine learning, cellular automata, agent-based models, complex networks, critical phenomena/phase transitions, fractals and power laws,...
- This amounts to saying: complex systems are what complex systems people study.
- This does have a nice internal consistency.
- In my opinion, what gets included as part of a discipline is often a frozen accident.

Tools

Many tools and techniques for complex systems will need to:

1. Measure unpredictability, distinguish between different sorts of unpredictability, work with probabilities
2. Be able to measure and discover pattern, complexity, structure, emergence, etc.
3. Be inferential; be inductive as well as deductive. Must infer from the system itself how it should be represented.
4. Be able to handle very large, possibly heterogeneous data sets.



Themes/General Principles??

Increasing Returns --> "Power laws"
Stability Through Hierarchy
Stability through Diversity
Complexity Increases?
Exploitation vs. Exploration
And many more?

Topics/Models

Neural Networks (real & fake)
Spin Glasses
Evolution (real & fake)
Immune System
Gene Regulation
Pattern Formation
Soft Condensed Matter
Origins of Life
Origins of Civilization
Origin and Evolution of Language
Population Dynamics
And many, many, more...

Complex Systems

Tools/Methods

Nonlinear Dynamics
Machine Learning
Cellular Automata
Symbolic Dynamics
Evolutionary Game Theory
Agent–Based Models
Information Theory
Stochastic Processes
Statistical Mechanics/RG
Networks
And many more ...

Foundations

Measures of Complexity
Representation and Detection of Organization
Computability, No Free Lunch Theorems
And many more...

Based on Fig. 1.1 from Shalizi, ”Methods and Techniques in Complex Systems Science: An Overview”, pp. 33-114 in Deisboeck and Kresh (eds.), Complex Systems Science in Biomedicine (New York: Springer-Verlag, 2006); `http://arxiv.org/abs/nlin.AO/0307015`

Comments on the Complex Systems Quadrangle

- The left and right corners of the quadrangle definitely exist.
- It is not clear to what extent the top of the quadrangle exists. Are there unifying principles? Loose similarities among complex systems? Or no relation at all? You should decide for yourself.
- The bottom of the quadrangle exists, but may or may not be useful depending on one's interests.
- Models of a particular topic often become topics themselves. E.g., models spin glasses were developed to study certain magnetic materials, but now some people study spin glasses for the sake of studying spin glasses.
- I'm not sure how valuable this figure is. Don't take it too seriously.
- My talk will focus on some topics from the bottom and the right.
- Toward the end of my talks I will offer some thoughts on what isn't (and what might be) on the top of the quadrangle.

Complexity: Initial Thoughts

- The complexity of a phenomena is generally understood to be a measure of how difficult it to describe it.
- But, this clearly depends on the language or representation used for the description.
- It also depends on what features of the thing you're trying to describe.
- There are thus many different ways of measuring complexity. I will aim to discuss a bunch of these in my lectures.
- Some important, recurring questions concerning complexity measures:
  1. What does the measure tell us?
  2. Why might we want to know it?
  3. What representational assumptions are behind it?

Predictability, Unpredictability, and Complexity

- The world is an unpredictable place. And there are different types of unpredictability.
- Nevertheless, there is much predictability.
- But there is more to life than predictability and unpredictability.
- The world is patterned, structured, organized, complex.
- We often have an intuitive sense that some phenomena are more complex than others.
- Where does this complexity come from?
- Is this complexity real, or is it an illusion?
- How is complexity related to unpredictability (entropy)?
- What are patterns? How can they they discovered?

Part II

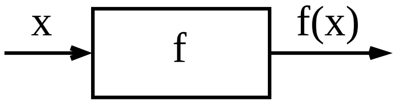
Introduction to Chaos: Basic Definitions, SDIC

A Brief, Introductory Overview of Dynamical Systems and Chaos

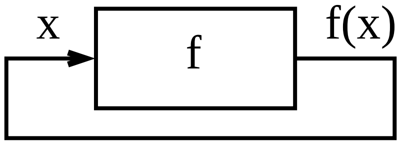
- A **Dynamical System** is any system that changes over time
  - A differential equation
  - A system of differential equations
  - Iterated functions
  - Cellular automata
- The goal of this brief introduction is to define a handful of terms, define chaos and sensitive dependence on initial conditions, and briefly discuss some of its implications.
- I will focus on iterated functions.
- Let’s start with an example.

Example: Iterating the squaring rule, $f(x) = x^2$

- Consider the function $f(x) = x^2$. What happens if we start with a number and repeatedly apply this function to it?
- E.g., $3^2 = 9$, $9^2 = 81$, $81^2 = 6561$, etc.
- The iteration process can also be written $x_{n+1} = x_n^2$.
- In this is example, the initial value 3 is the **seed**, often denoted x_0.
- The sequence $3, 9, 81, 6561, \cdots$ is the **orbit** or the **itinerary** of 3.
- Picture the function as a “box” that takes x as an input and outputs $f(x)$:



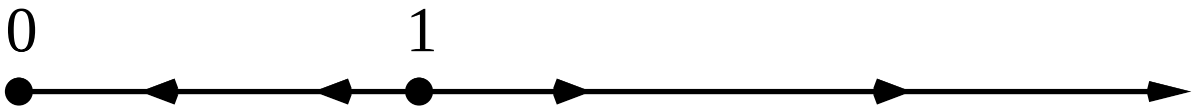
- Iterating the function is then achieved by feeding the output back to the function, making a feedback loop:



The squaring rule, continued

In dynamics, we are usually interested in the long-term behavior of the orbit, not in the particulars of the orbit.

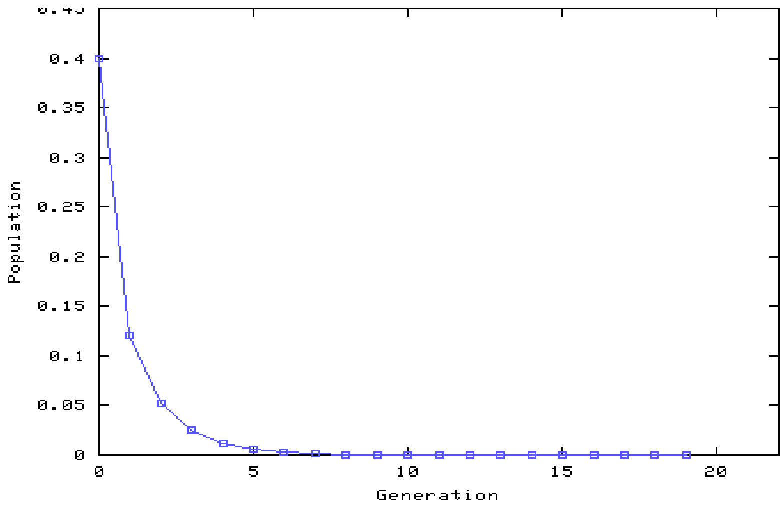
- The seed 3 tends toward infinity—it gets bigger and bigger.
- Any $x_0 > 1$ will tend toward infinity.
- If $x_0 = 1$ or $x_0 = 0$, then the point never changes. These are fixed points.
- If $0 \leq x_0 < 1$, then x_0 approaches 0.
- We can summarize this with the following diagram:



- 0 and 1 are both **fixed points**
- 0 is a **stable** or **attracting** fixed point
- 1 is an **unstable** or **repelling** fixed point

Logistic Equation

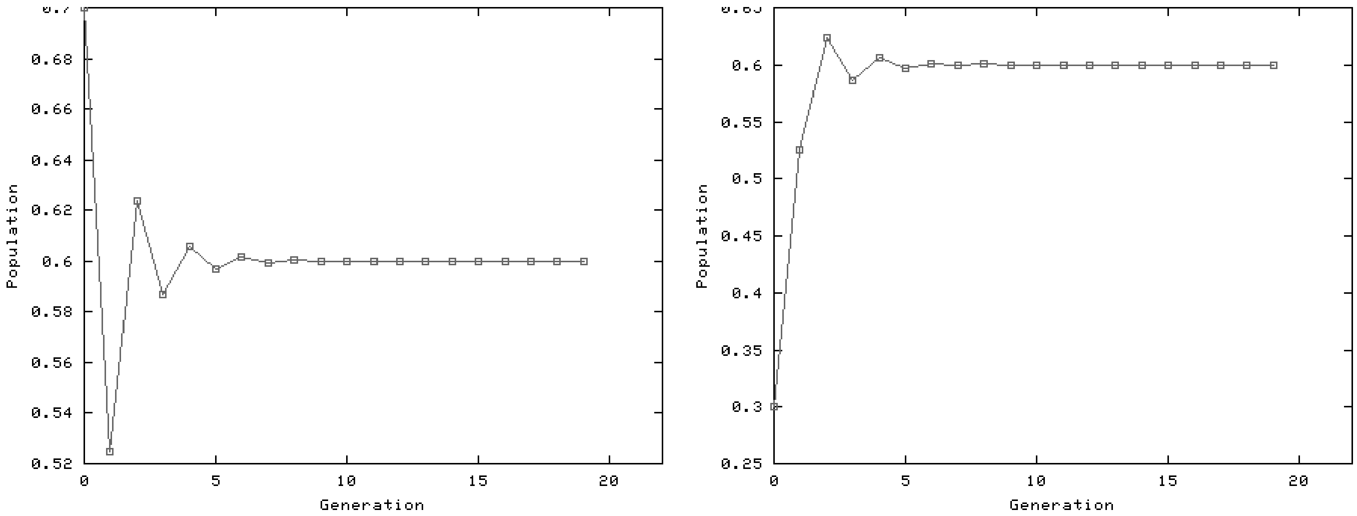
- Logistic equation: $f(x) = rx(1-x)$.
- A simple model of resource-limited population growth.
- The population x is expressed as a fraction of the carrying capacity. $0 \leq x \leq 1$.
- r is a parameter—the growth rate—that we will vary.
- Let's first see what happens if $r = 0.5$.



- This graph is known as a **time series plot**.
- 0 is an attracting fixed point.

Logistic Equation, $r = 2.5$

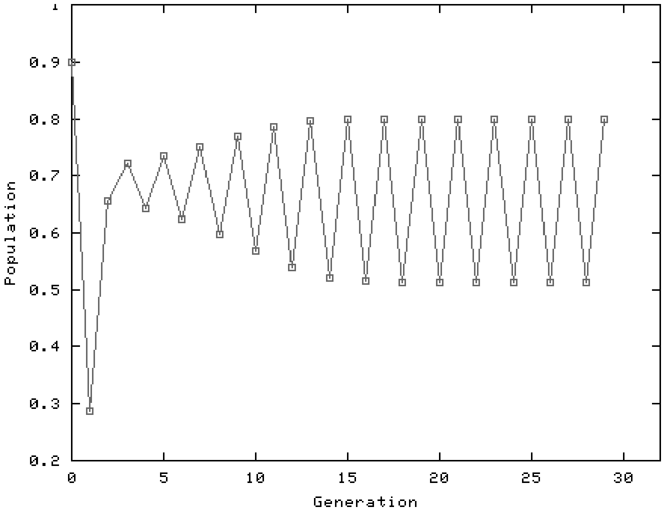
- Logistic equation, $r = 2.5$.



- All initial conditions are pulled toward 0.6.
- (Note that there are different vertical scales on the two plots.)
- 0.6 is an attracting fixed point.

Logistic Equation, $r = 3.2$

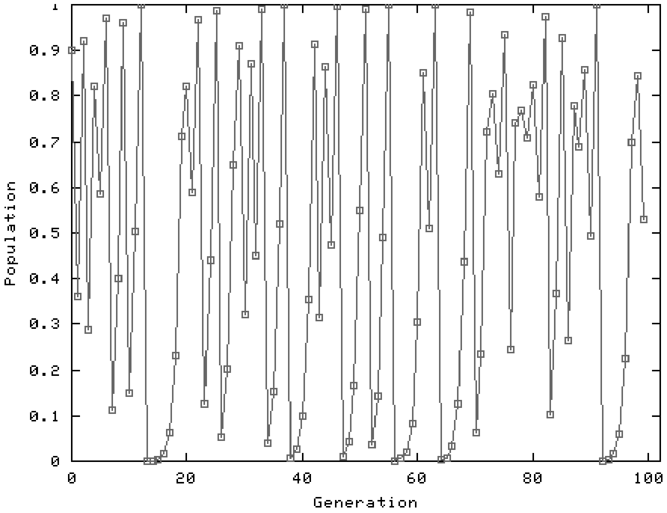
- Logistic equation, $r = 3.2$.



- Initial conditions are pulled toward a **cycle** of period 2.
- The orbit oscillates between 0.513045 and 0.799455.
- This cycle is an attractor. Many different initial conditions get pulled to it.

Logistic Equation, $r = 4.0$

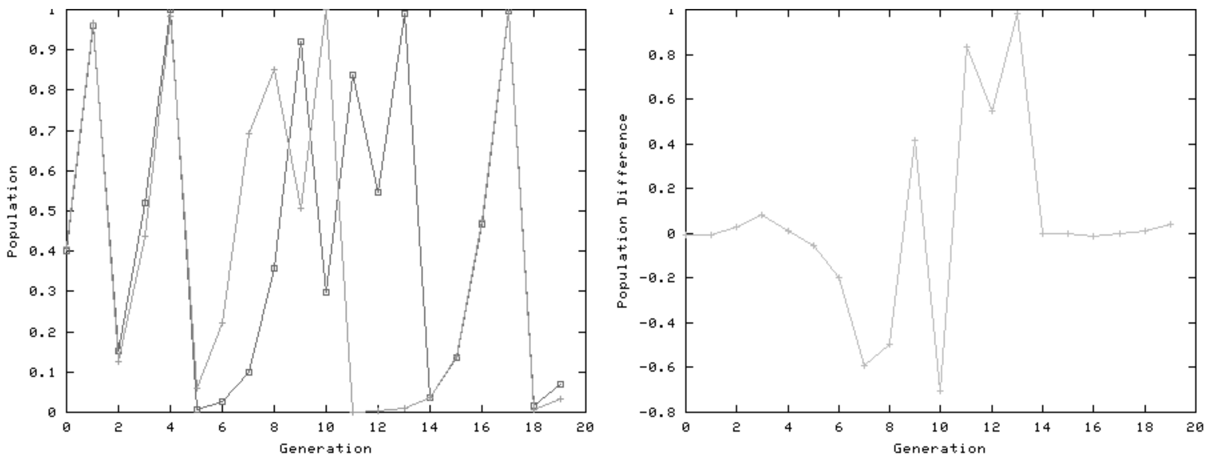
- Logistic equation, $r = 4.0$.



- What's going on here?!
- The orbit is not periodic. In fact, it never repeats.
- This is a rigorous result; it doesn't rely on computers.
- What happens if we try different initial conditions?

Different Initial conditions

- Logistic equation, $r = 4.0$. Two different initial conditions, $x_0 = 0.4$ and $x_0 = 0.41$.

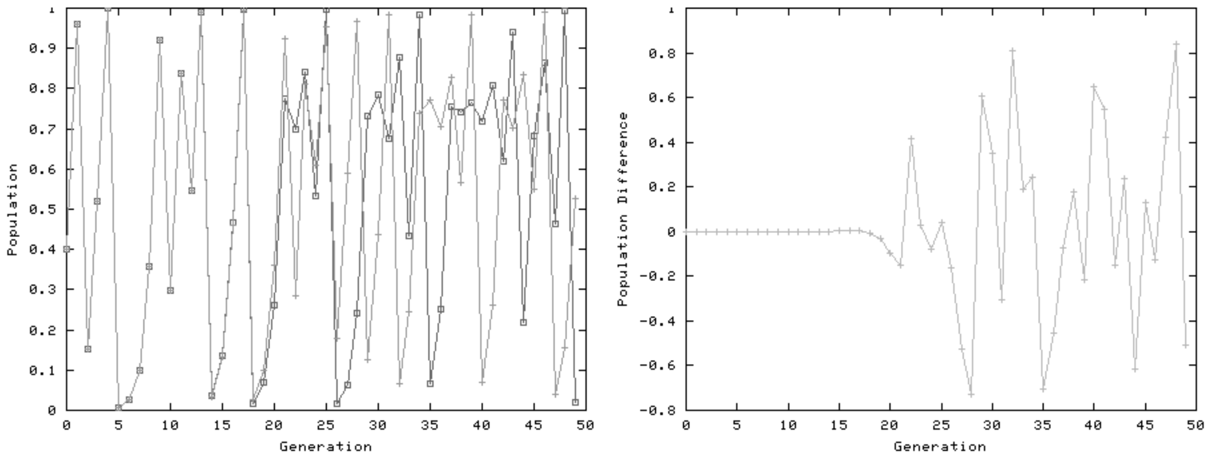


- The right graph plots the difference between the two orbits on the left with slightly different initial conditions.
- Note that the difference between the two orbits grows.
- Can think of one initial condition as the true one, and the other as the measured one.

- The plot on the right then shows what happens to our prediction error over time.
- What happens if the two initial conditions are closer together?

Sensitive Dependence on Initial Conditions

- Logistic equation, $r = 4.0$. Two different initial conditions, $x_0 = 0.4$ and $x_0 = 0.4000001$.



- The two initial conditions differ by one part in one million
- The orbits differ significantly after around 20 iterations, whereas before they differed after around 4 iterations.
- Increasing the accuracy of the initial condition by a factor of 10^5 allow us to predict the outcome 5 times further.

- Thus, for all practical purposes, this system is unpredictable, even though it is deterministic.
- This phenomena is known as **Sensitive Dependence on Initial Conditions**, or, more colloquially, **The Butterfly Effect**.

Definition of Sensitive Dependence on Initial Conditions

- A dynamical system has sensitive dependence on initial conditions (SDIC) if arbitrarily small differences in initial conditions eventually lead to arbitrarily large differences in the orbits.

More formally

- Let X be a metric space, and let f be a function that maps X to itself: $f : X \mapsto X$.
- The function f has SDIC if there exists a $\delta > 0$ such that $\forall x_1 \in X$ and $\forall \epsilon > 0$, there is an $x_2 \in X$ and a natural number $n \in N$ such that $d[x_1, x_2] < \epsilon$ and $d[f^n(x_1), f^n(x_2)] > \delta$.
- In other words, two initial conditions that start ϵ apart will, after n iterations, be separated by a distance δ.

Definition of Chaos

There is not a 100% standard definition of chaos. But here is one of the most commonly used ones:

An iterated function is **chaotic** if:

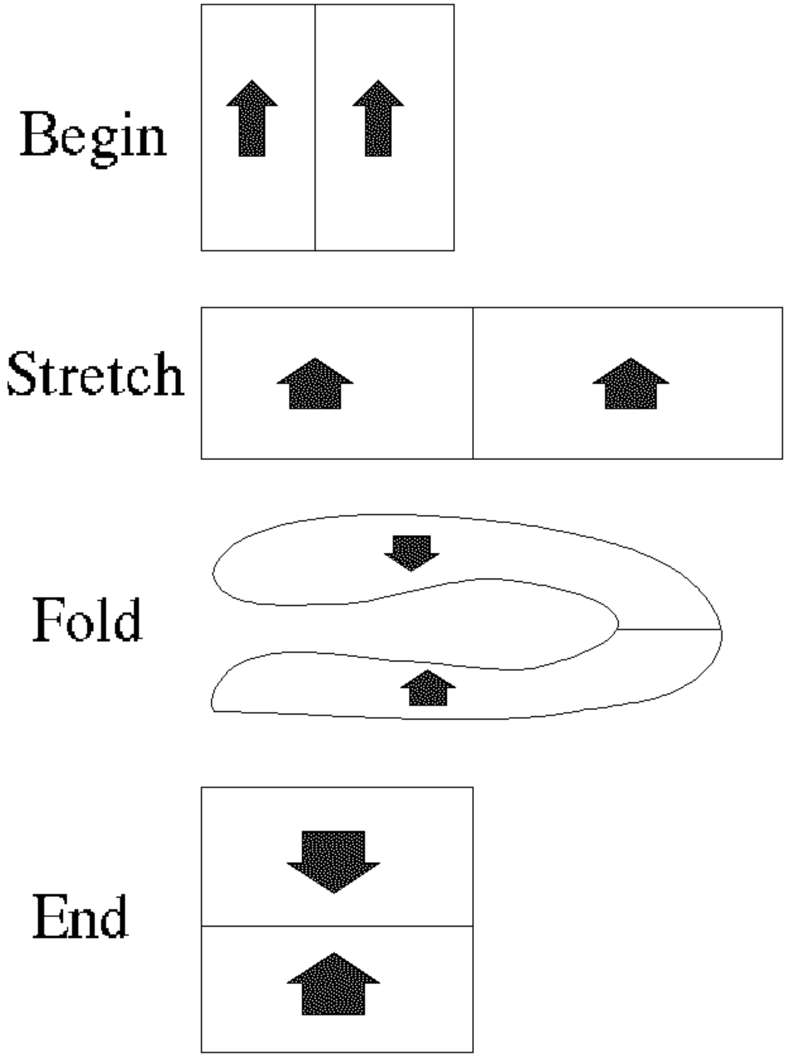
1. The function is **deterministic**.
2. The system's orbits are **bounded**.
3. The system's orbits are **aperiodic**; i.e., they never repeat.
4. The system has **sensitive dependence on initial conditions**.

Other properties of a chaotic dynamical system ($f : X \mapsto X$) that are sometimes taken as defining features:

1. **Dense periodic points:** The periodic points of f are dense in X.
2. **Topological transitivity:** For all open sets $U, V \in X$, there exists an $x \in U$ such that, for some $n < \infty$, $f_n(x) \in V$. I.e., in any set there exists a point that will get arbitrarily close to any other set of points.

Geometry of Chaos

Geometrically, all chaotic systems involve stretching and folding:

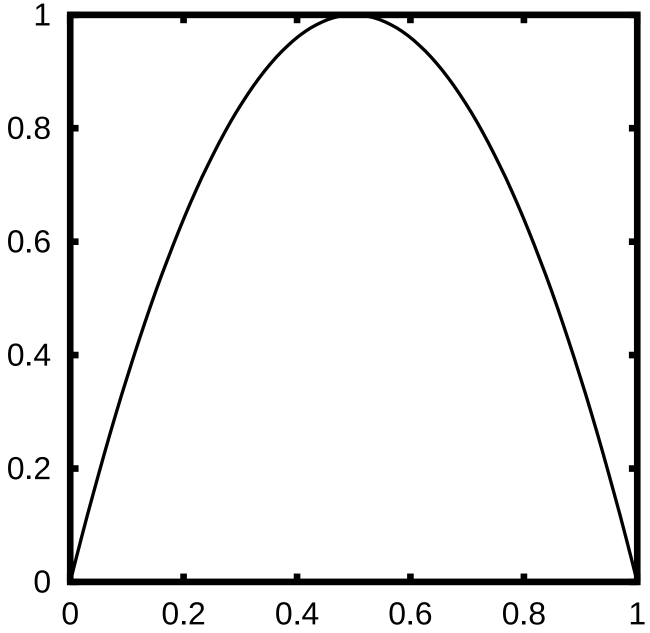


- Stretching pulls nearby points apart, leading to **sensitive dependence on initial conditions.**

- Folding keeps the orbits **bounded**.

Geometry of Chaos, continued

The logistic equation may be viewed as stretching and folding the unit interval onto itself:



- Note that the amount of stretching is captured by the slope of the function.
- We shall see that the "average slope" is related to the degree of SDIC, which is in turn related to the unpredictability.

- Thus, SDIC is a geometric property of the system.
- We will make this idea precise in the next set of lectures.

Chaos and Dynamical Systems: Selected References

There are many excellent references and textbooks on dynamical systems. Some of my favorites:

- Peitgen, et al. *Chaos and Fractals: New Frontiers of Science*. Springer-Verlag. 1992. *Huge (almost 1000 pages), and very clear. Excellent balance of rigor and intuition.*
- Cvitanović, *Universality in Chaos, second edition*, World Scientific. 1989. *Comprehensive collection of reprints. Very handy. Nice introduction by Cvitanović*.
- Gleick, *Chaos: Making a New Science*. Penguin Books. 1988. *Popular science book. Very good. Extremely well written and accurate.*
- Devaney. *An Introduction to Chaotic Dynamical Systems, second edition*. Perseus Publishing. 1989. *Advanced undergrad math textbook. Very clear.*
- Strogatz. *Nonlinear Dynamics and Chaos*. Perseus Books Group. 2001.
- Smith. *Chaos: A Very Short Introduction*. Oxford. 2007.
- I am working on an algebra-based introduction to chaos and fractals textbook. If you are interested in seeing a draft—especially if you might use it in a class—please let me know.

Part II

Part III

More Chaos: Period Doubling, Universality, An Interlude about Power Laws, Lyapunov Exponents

Introduction to Chaos Part II

We have seen several possible long-term behaviors for the logistic equation:

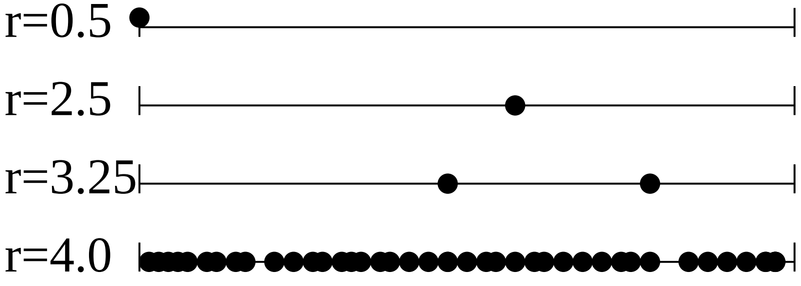
1. $r = 0.5$: attracting fixed point at 0.

2. $r = 2.5$: attracting fixed point at 0.6.

3. $r = 3.25$: attracting cycle of period 2.

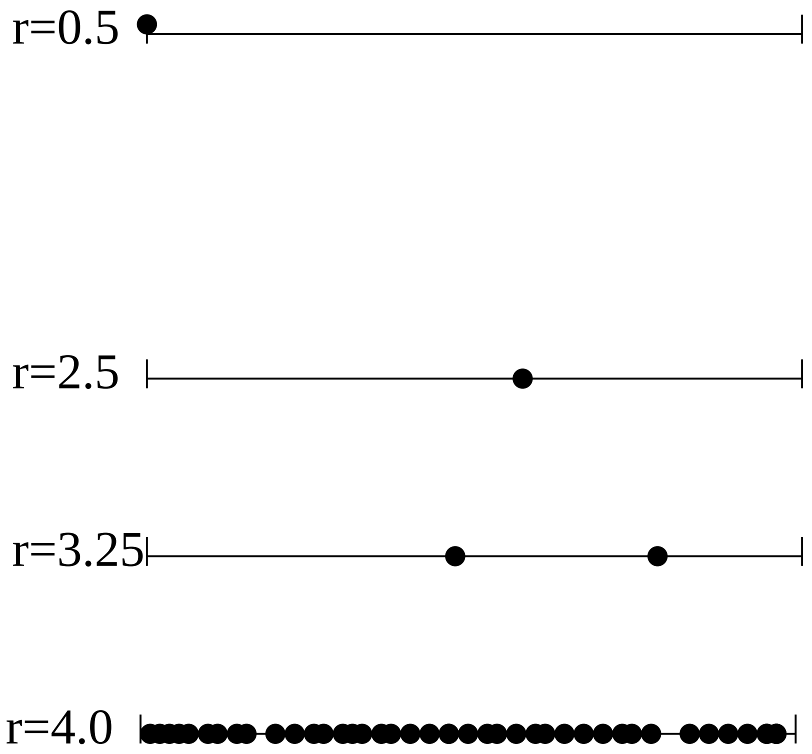
4. $r = 4.0$: chaos.

Graphically, we can illustrate this as follows:



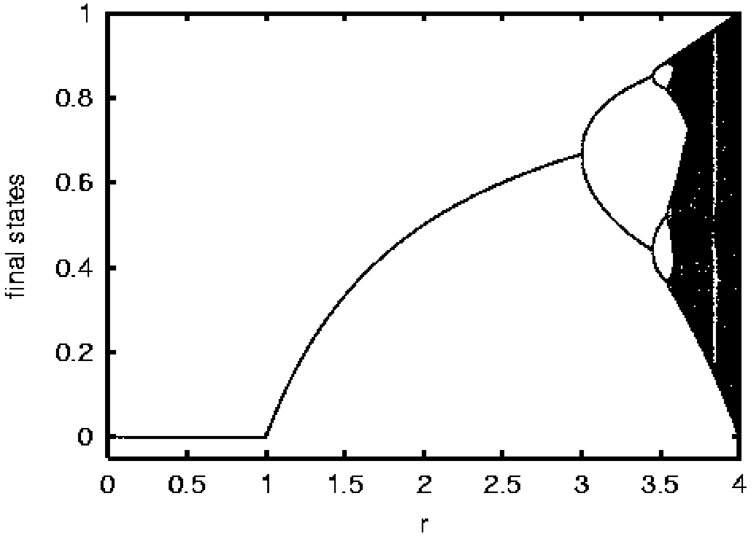
- I.e., for each r, iterate and plot the final x values as dots on the number line.

- What else can the logistic equation do??



- Do this for more and more r values and “glue” the lines together.
- Turn sideways and ...

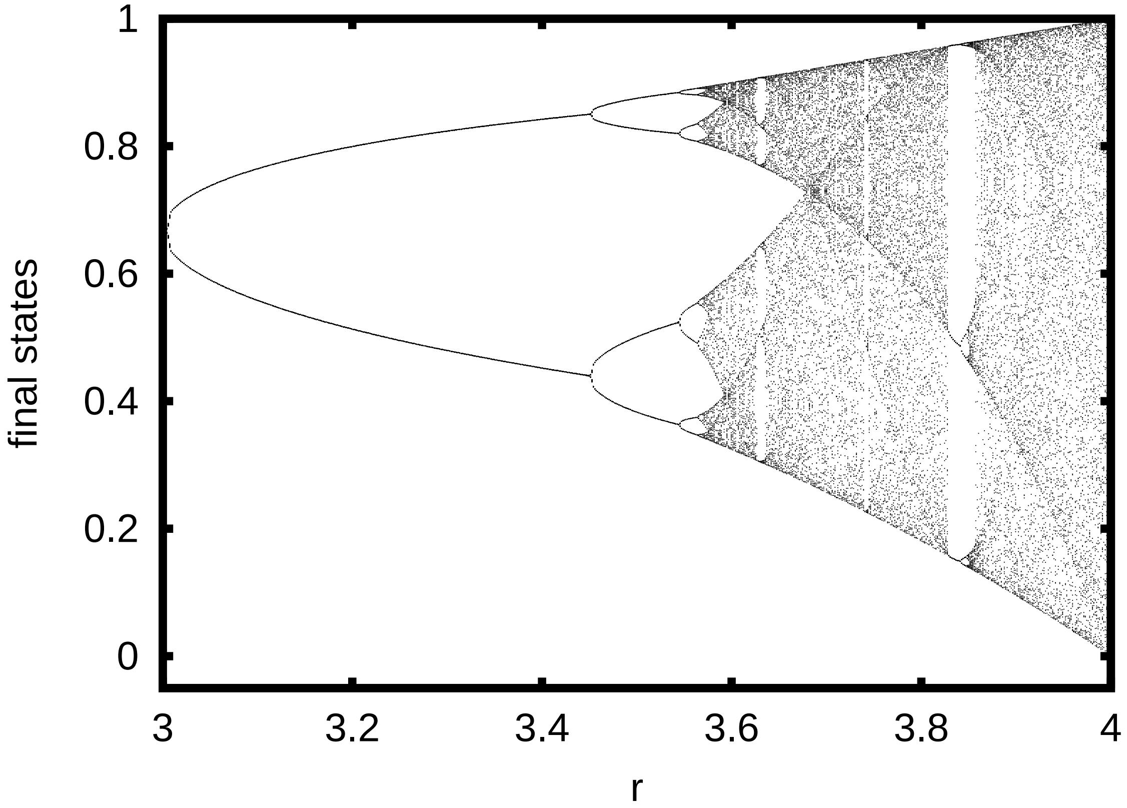
Bifurcation Diagram



- The bifurcation diagram shows the all the the possible long-term behaviors for the logistic map.
- $0 < r < 1$, the orbits are attracted to zero.
- $1 < r < 3$, the orbits are attracted to a non-zero fixed point.
- $3 < r < 3.45$, orbits are attracted to a cycle of period 2.
- Chaotic regions appear as dark vertical lines.

Bifurcation diagram, continued

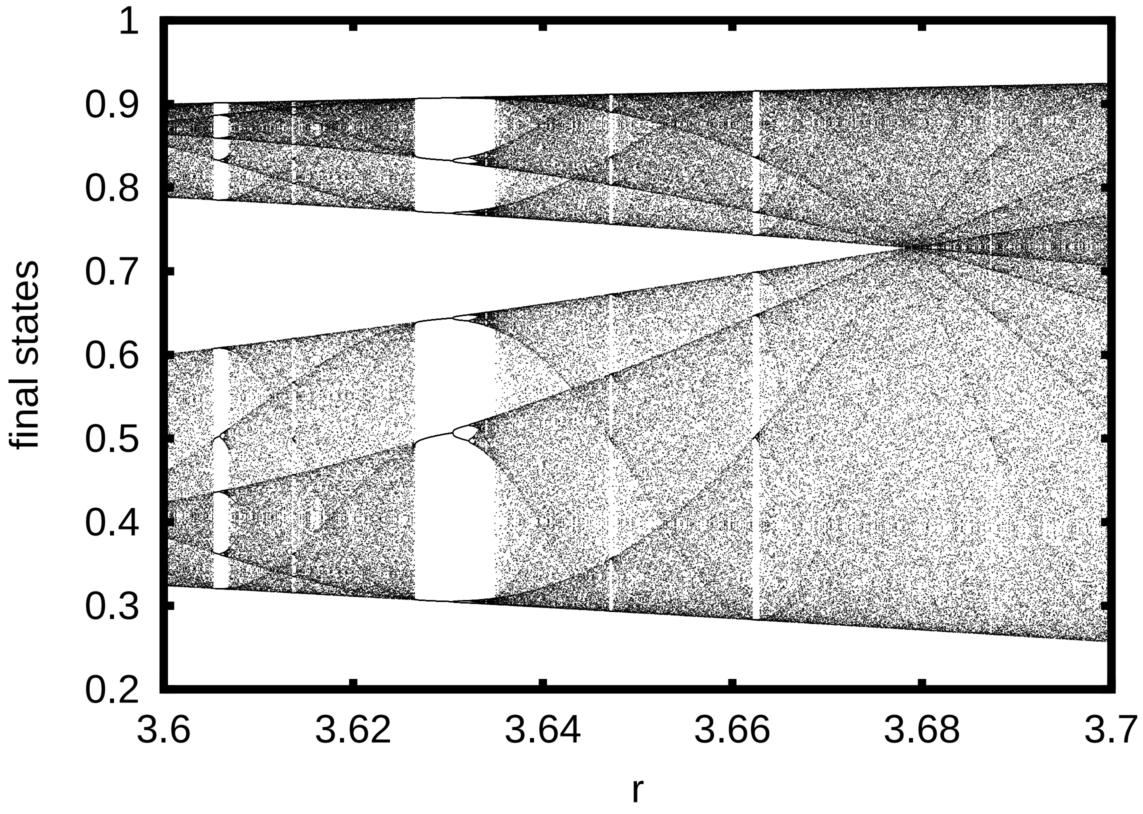
Let’s zoom in on a region of the bifurcation diagram:



- The sudden qualitative changes are known as **bifurcations**.
- There are **period-doubling bifurcations** at $r \approx 3.45$, $r \approx 3.544$, etc.
- Note the window of period 3 near $r = 3.83$.

Bifurcation diagram, continued

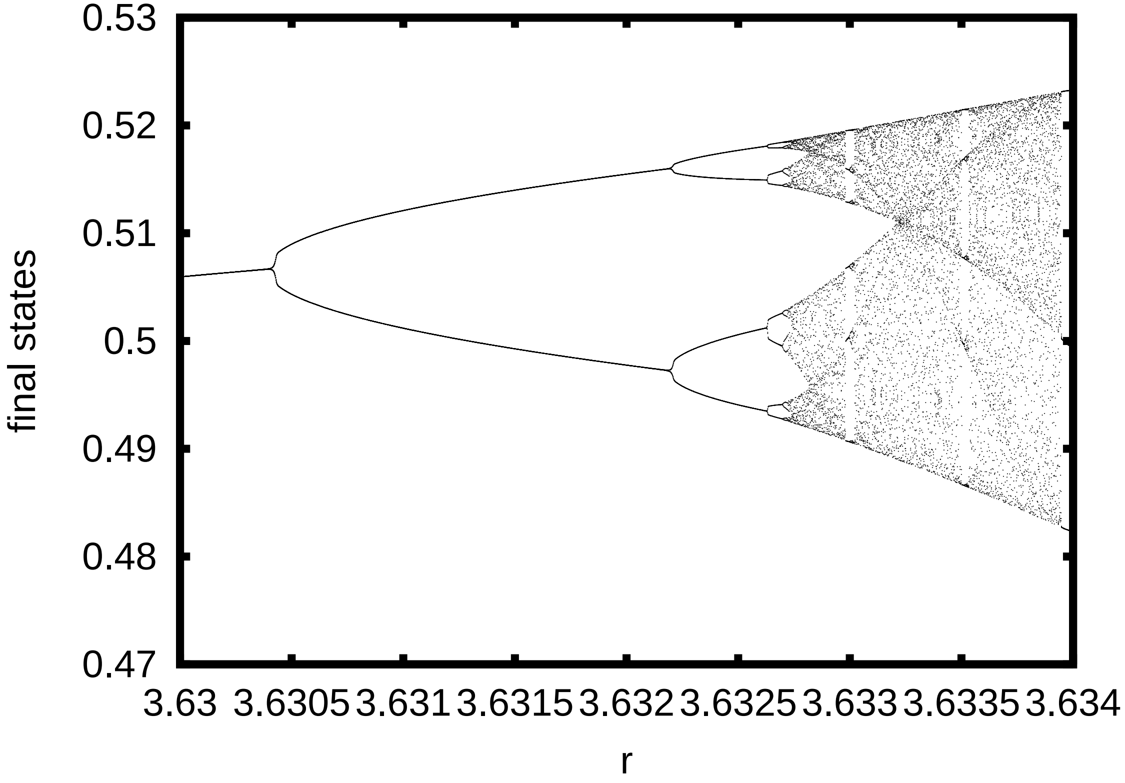
Let's zoom in again:



- Note the sudden changes from chaotic to periodic behavior.

Bifurcation diagram, continued

Let’s zoom in once more:



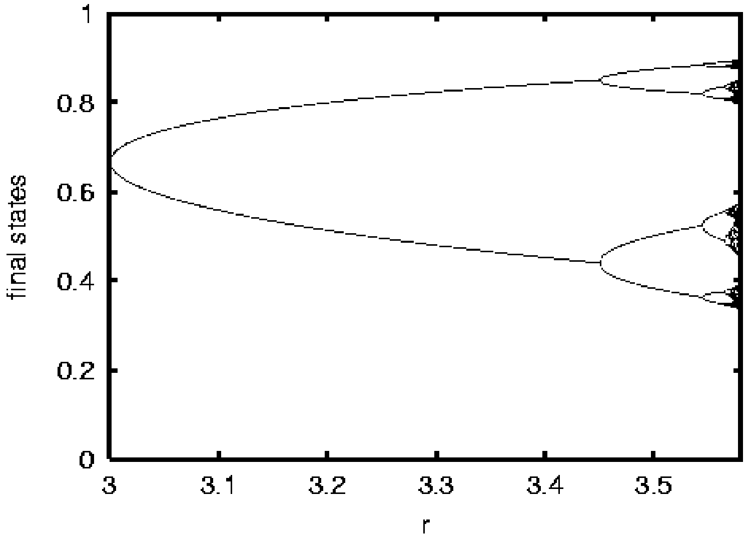
- Note the small scales on the vertical axis, and the tiny scale on the y axis.
- Note the self-similar structure. As we zoom in we keep seeing sideways pitchforks.

Bifurcation Diagram Summary

- As we vary r, the logistic equation shuffles suddenly between chaotic and periodic behaviors, but the bifurcation diagram reveals that these transitions appear in a structured, or regular, way.
- In the next few slides we'll examine one of the regularities in the bifurcation diagram: The **period-doubling route to chaos**.

Period-Doubling Route to Chaos

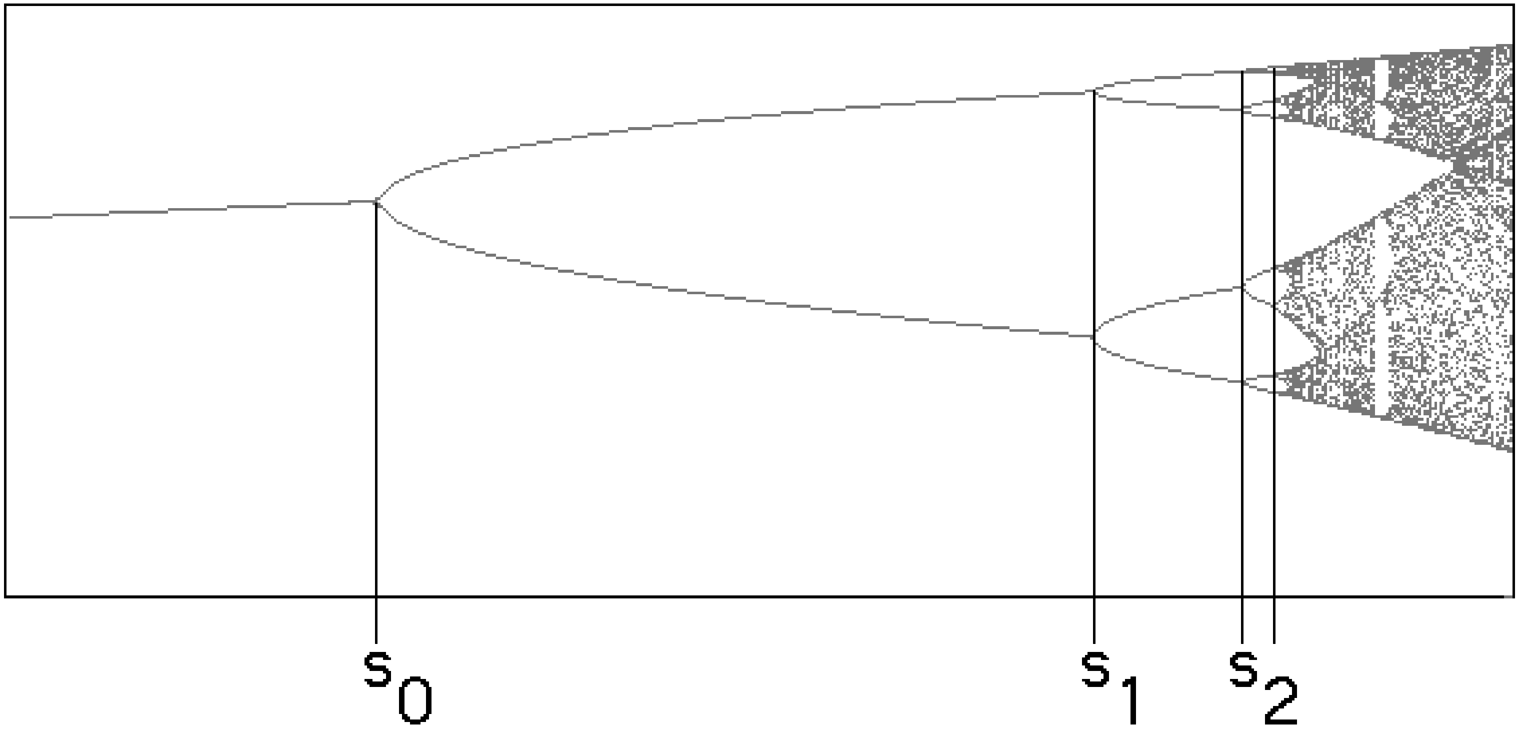
- As r is increased from 3, a sequence of period doubling bifurcations occur.



- At $r = r_\infty \approx 3.569945672$ the periods “accumulate” and the map becomes chaotic.
- For $r > r_\infty$ it has SDIC. For $r < r_\infty$ it does not.
- This is a type of **phase transition**: a sudden qualitative change in a system’s behavior as a parameter is varied continuously.

Period-Doubling Route to Chaos: Geometric Scaling

- Let's examine the ratio of the lengths of the pitchfork tines in the bifurcation diagram.

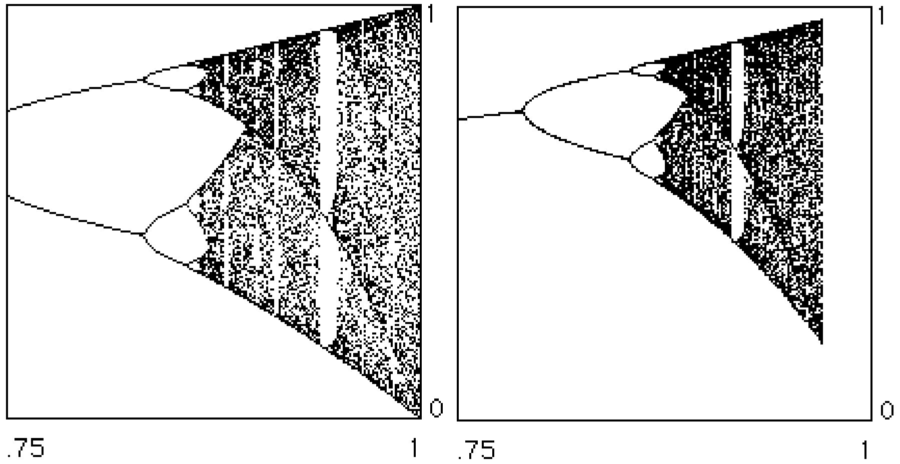


- The first ratio is: $\delta_1 = \frac{s_1 - s_0}{s_2 - s_1}$.
- The n$^{\rm th}$ ratio is: $\delta_n = \frac{s_n - s_{n-1}}{s_{n+1} - s_n}$.

Feigenbaum's Constant

- This ratio approaches a limit: $\lim_{n\to\infty} \delta_n = 4.669201609\ldots$. This is known as **Feigenbaum's constant** δ.
- This means that the bifurcations occur in a regular way.
- Amazingly, the value of δ is **universal**: it is the same for any period-doubling route to chaos!
- Figure Source: `http://classes.yale.edu/fractals/Chaos/Feigenbaum/Feigenbaum.html`

Universality



- The figure on the left is the bifurcation diagram for $f(x) = r\sin(\pi x)$.
- The figure on the right is the bifurcation diagram for $f(x) = \frac{27}{4} r x^2 (1-x)$.
- The bifurcation diagrams are very similar: **both have** $\delta \approx 4.6692$.
- Mathematically, things are constrained so that there is, in some sense, only one possible way for a system to undergo a period-doubling to chaos.
- Figure Source:

  `http://classes.yale.edu/fractals/Chaos/LogUniv/LogUniv.html`

Experimental Verification of Universality

- Universality isn't just a mathematical curiosity. Physical systems undergo period-doubling order-chaos transitions. Almost miraculously, these systems also appear to have a universal δ.
- Experiments have been done on fluids, circuits, acoustics:
  - Water: $4.3 \pm .8$
  - Mercury: $4.4 \pm .1$
  - Diode: $4.5 \pm .6$
  - Transistor: $4.5 \pm .3$
  - Helium: $4.8 \pm .6$

  Data from Cvitanović, *Universality in Chaos*, World Scientific, 1989.
- A very simple equation, the logistic equation, has produced a quantitative prediction about complicated systems (e.g., fluid turbulence) that has been verified experimentally.
- Nature is somehow constrained.

Detour: A Little Bit More About Universality

- The order-disorder phase transition in the logistic map is not the only sort of phase transition that is universal.
- Second order (aka continuous) phase transitions are also universal.
- There are several different universality classes, each of which has different values for quantities analogous to δ.
- The symmetry of the order parameter and the dimensionality of the space of the system determine the universality class.
- The order parameter is a quantity which is zero on one side of the transition and non-zero on the other.
- The theory of critical phenomena does not tell one how to find order parameters. Sometimes order parameters are not obvious.

A Little Bit More About Universality, continued

- At the transition point, or **critical point** T_c, some quantities (e.g., specific heat c) usually diverge. The divergence is described by a power law. The exponents for these power laws are called **critical exponents** (e.g, α).

$$c \sim \left(\frac{T_c}{|T - T_c|} \right)^{\alpha} \; , \quad T \approx T_c \; . \tag{1}$$

- At the critical point, the correlations between components of the system usually decay with a power law as the distance increases. Away from the critical point, the decay is exponential—much faster.

- Let $\Gamma(r)$ be the correlation function between two degrees of freedom separated by a distance r. Then, at T_c,

$$\Gamma(r) \sim \frac{1}{r^{\mu}} \; . \tag{2}$$

- Note: μ is usually written $\mu = 2 - d + \eta$.

A Little Bit about Power Laws

- At critical points, functions like the specific heat diverge with a power law.
- This divergence arises because the correlations between the system's parts is long range—the corrections decay with a power law, not an exponential.
- Power-law decay of correlations is an indication that the system is organized or complex.
- However, this does not mean that the only way that power law distributions can be formed is via long-range order or correlations or complexity.
- In fact, there are very simple mechanisms for producing power laws.

Simple Ways to Make A Power Law Distribution

Exponentially Observing Exponential Distribution

- Suppose a quantity is growing exponentially: $X(t) = e^{\mu t}$.
- Suppose we measure the quantity at a random time T, obtaining the value $\bar{X} = e^{\mu T}$.
- Let T also be exponentially distributed: $Pr(T > t) = e^{-\nu T}$.
- Then the probability density for $\bar{X}$ is given by $f_{\bar{X}}(x) = kx^{-\mu/\nu - 1}$.
- Like magic, a power law has appeared.
- In general, there are lots of ways to make power laws by combining exponential distributions in different ways.
- See Reed and Hughes, Why power laws are so common in nature. Physical Review E 66:067103. 2002. `http://www.math.uvic.ca/faculty/reed/`.

Simple Ways to Make A Power Law Distribution, Continued

Multiplicative Noise:

- Define a random variable X as the product of a number of other random variables. In many cases X will be distributed with a power law.
- See, e.g., Sornette, Multiplicative processes and power laws. Physical Review E 57, 4811. 1998. `arXiv.org/cond-mat/9708213`.

Warning about Empirical Power Laws!

- Concluding that a distribution is actually a power is a potentially subtle matter.
- It is **not** a good idea to do a linear fit on a log-log plot!!
- For much more, see A. Clauset, C.R. Shalizi, and M.E.J. Newman, ”Power-law distributions in empirical data” E-print (2007). `arXiv.org/physics/0706.1062`. This is an extraordinary paper.
- Source code in matlab and R for Clauset, et al, is available at `http://www.santafe.edu/~aaronc/powerlaws/`.
- For additional commentary, see Shalizi, “So you think you have a power law—Well isn’t that special?” `http://cscs.umich.edu/~crshalizi/weblog/491.html`.
- See also, M.E.J. Newman, Power laws, Pareto distributions and Zipf’s law, *Contemporary Physics* **46**, 323-351 (2005). `arXiv.org/cond-mat/0412004`, and D. Sornette, Probability Distributions in Complex Systems, `http://arxiv.org/abs/0707.2194`.

Power Law Conclusions

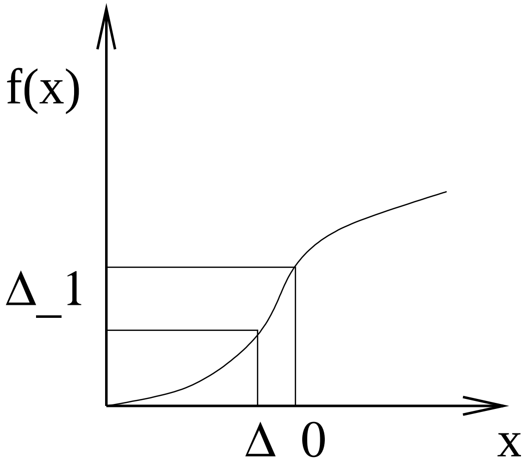
- There are many simple, non-complex ways to make power laws.
- They are not necessarily an indicator of complexity or correlation or organization.
- They are not necessarily an indicator of criticality—of a system on the edge of a phase transition.
- Many of the claims in the literature for the existence of power laws may be based on faulty data analysis.

This ends the interlude on power laws. We now return to chaotic dynamics...

Measuring Sensitive Dependence: Lyapunov Exponent

SDIC arises because the function pushes nearby points apart. The Lyapunov exponent measures this pushing.

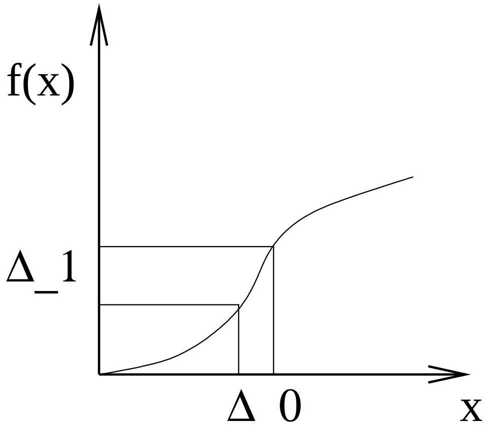
- Consider an initial small interval Δ_0 of initial conditions centered at x_0.



- After one iteration, this interval becomes $\Delta_1 \approx |f'(x_0)|\Delta_0$.
- $|f'(x_0)|$ is the local stretch (or shrink) factor.
- After n iterations, $\Delta_n = \prod_n |f'(x_n)|\Delta_0$.
- The idea is that for the n$^{\rm th}$ iterate interval is getting stretched (shrunk) by the stretch factor $f'(x)$ evaluated at the x_n, the location of the n$^{\rm th}$ iterate of x_0.

Lyapunov Exponent, continued

- We expect the growth of the interval Δ_0 to be exponential, since we're multiplying the interval at each time step.



- That is, we expect that $\frac{\Delta_n}{\Delta_0} = e^{\lambda n}$, where λ is the exponential growth rate.
- The exponential growth is just the the product of all the local stretch factors along an itinerary:

$$e^{\lambda n} = \prod_n |f'(x_n)| \ .$$

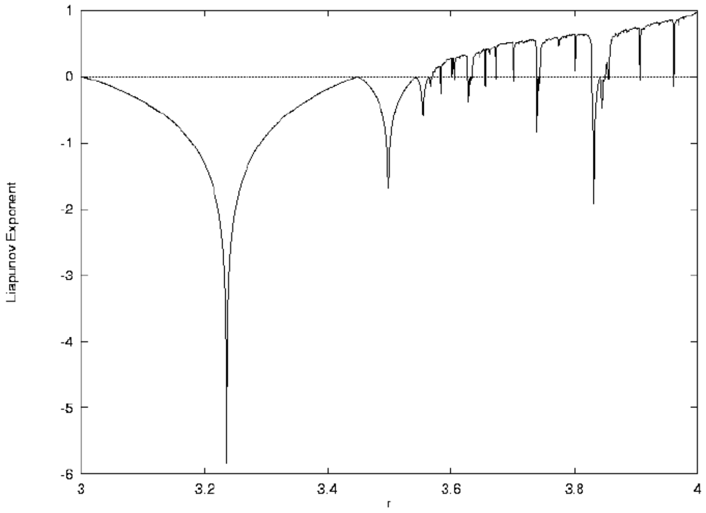
Lyapunov Exponent, continued

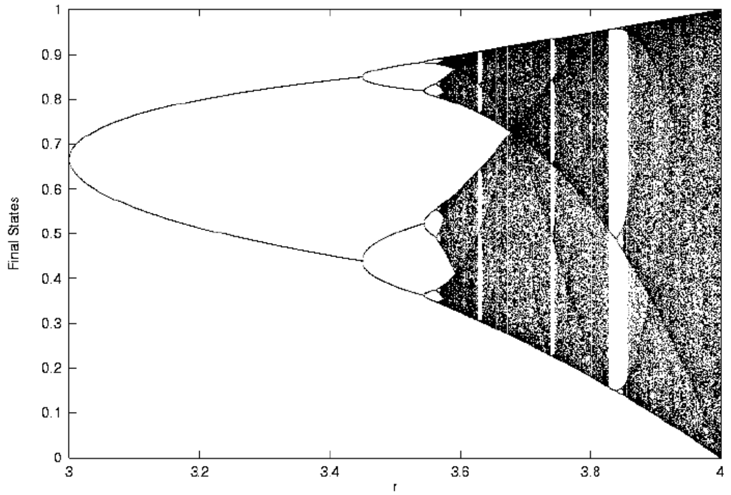
- Solving for λ:

$$\lambda \equiv \lim_{N\to\infty} \left[\frac{1}{N} \sum_n \ln |f'(x_n)| \right] \ . \tag{3}$$

- λ is the **Lyapunov exponent**. It measures the degree of SDIC.
- If $\lambda > 0$, the function has SDIC.

Lyapunov Exponent for the Logistic Equation

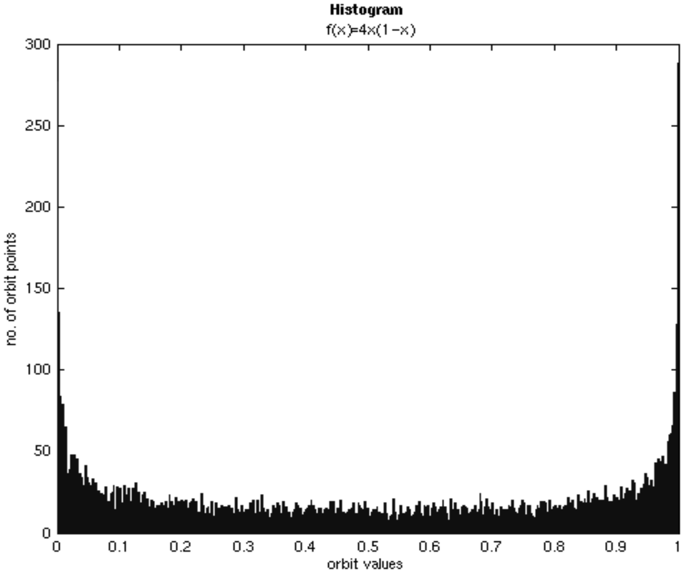




- The top graph shows the Lyapunov exponent as a function of r.
- Note that $\lambda > 0$ in the chaotic regions of the bifurcation diagram.

Initial Conditions?

- It seems like the definition of λ depends on the initial condition. If so, λ is a property of x_o, and not a global property of f.
- It turns out that for many dynamical systems you will get the same λ for almost all x_0. Why is this?
- Imagine building a histogram for orbits. For $r = 4$, this will look like:



- `http://www-m8.mathematik.tu-muenchen.de/personen/hayes/chaos/Hist.html`

Natural Invariant Densities and Ergodicity

- The distribution in this histogram $\rho(x)$ will be obtained by iterating almost any initial condition x_0.
- This distribution is known as the **Natural Invariant Density**.
- If we can figure it out, we can determine the Lyapunov exponent by integrating:
$$\lambda = \int \ln(|f'(x)|)\rho(x)\,dx \ .$$
- In general, if a dynamical property like the lyapunov exponent can be determined by integrating over x instead of performing a dynamical average, the system is **ergodic**.
- Proving that a system is ergodic is usually very hard.
- Trivia: for $r = 4$, $\rho(x) = \frac{\pi}{\sqrt{x(1-x)}}$.
- For other r values, an expression for $\rho(x)$ is not known. Generally, $\rho(x)$ is non-smooth.

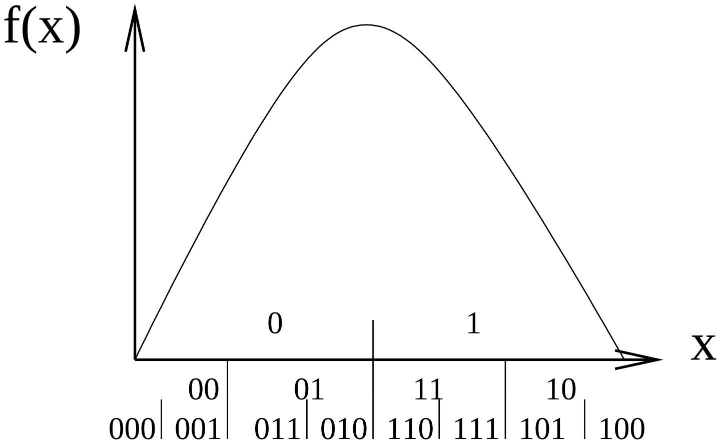
Symbolic Dynamics

- It is often easier to study dynamical systems via symbolic dynamics.
- The idea is to encode the continuous variable x with a discrete variable in some clever way that doesn't entail a loss of information.
- For the logistic equation:

$$s_i = \begin{cases} 0 & x \leq \frac{1}{2} \\ \\ 1 & x > \frac{1}{2} \end{cases} .$$

- Why is this ok? It seems that we're throwing out a lot of information!
  - The the function is deterministic, the initial condition contains all information about the itinerary.
  - For the coding above, longer and longer sequences of 1's and 0's code for smaller and smaller regions of initial conditions.
  - Codings that have this property are known as **generating partitions**.

Symbolic Dynamics, continued



- If we find a generating partition, we can use the symbols to explore the function's properties.

- The symbol sequences are "the same" as the orbits of x: they have the same periodic points, the same stability, etc.

- For $r = 4$, the symbolic dynamics of the logistic equation produce a sequence of 0's and 1's that is indistinguishable from a fair coin toss.

- In general, finding a good symbolic mapping is difficult and may be impossible and/or ill-advised.

Chaos Conclusions

- Deterministic systems can produce random, unpredictable behavior. E.g., logistic equation with $r = 4$.
- Simple systems can produce complicated behavior. E.g., long periodic behavior in logistic equation.
- Some features of dynamical systems are universal—the same for many different systems.
- More generally, dynamics are important. Considering only static averages can be misleading.

Chaos ⇒ Complex Systems

Some of the roots of complex systems are in chaos:

- Universality gives us some reason to believe that we can understand complicated systems with simple models.
- Appreciation that complex behavior can have simple origins.
- Awareness that there's more to dynamical systems than randomness. These systems also make patterns, organize, do cool stuff.
- Is there a way we can describe or quantify these patterns?
- Is there a quantity like the Lyapunov exponent that measures complexity or pattern or structure?
- What is a pattern?

Part IV

Information Theory: Motivation, Basic Definitions, Noiseless Coding Theorem

Information Theory

- Originally developed by Shannon in 1948 as he was figuring out how to efficiently transmit communication signals over a possibly noisy communication channel.

- I am not so much interested in its original uses in communication theory, but in its development and application as a broadly applicable tool for describing probability distributions.

- Information theory lets us ask and answer questions such as:
  1. How random is a sequence of measurements?
  2. How much memory is needed to store the outcome of measurements?
  3. How much information does one measurement tell us about another?

Some Info Theory References

1. T.M. Cover and J.A. Thomas, Elements of Information Theory. John Wiley & Sons, Inc., 1991. By far the best information theory text around.

2. C.E. Shannon and W. Weaver. The Mathematical Theory of Communication. University of Illinois Press. 1962. Shannon's original paper and some additional commentary. Very readable.

3. J.P. Crutchfield and D.P. Feldman, "Regularities Unseen, Randomness Observed: Levels of Entropy Convergence." *Chaos* **15**:25–53. 2003.

4. David MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003. Full text available at: `http://www.inference.phy.cam.ac.uk/mackay/itila/`.

5. D.P. Feldman. A Brief Tutorial on: Information Theory, Excess Entropy and Statistical Complexity: Discovering and Quantifying Statistical Structure. `http://hornacek.coa.edu/dave/Tutorial/index.html`.

Notation for Probabilities

Information theory is concerned with probabilities. We first fix some notation.

- X is a random variable. The variable X may take values $x \in \mathcal{X}$, where $\mathcal{X}$ is a finite set.
- likewise Y is a random variable, $Y = y \in \mathcal{Y}$.
- The probability that X takes on the particular value x is $\Pr(X = x)$, or just $\Pr(x)$.
- Probability of x and y occurring: $\Pr(X = x, Y = y)$, or $\Pr(x, y)$
- Probability of x, given that y has occurred: $\Pr(X = x|Y = y)$ or $\Pr(x|y)$

Example: A fair coin. The random variable X (the coin) takes on values in the set $\mathcal{X} = \{h, t\}$.

$\Pr(X = h) = 1/2$, or $\Pr(h) = 1/2$.

Different amounts of uncertainty?

- Anytime we describe a situation with probabilities, it's because we're uncertain of the outcome.
- However, some probability distributions indicate more uncertainty than others.
- We seek a function $H[X]$ that measures the amount of uncertainty associated with outcomes of the random variable X.
- What properties should such an uncertainty function have?
  1. Maximized when the distribution over X is uniform.
  2. Continuous function of the probabilities of the different outcomes of X
  3. Independent of the way in which we might group probabilities.

Entropy of a Single Variable

The requirements on the previous page *uniquely* determine $H[X]$, up to a multiplicative constant.

The Shannon entropy of a random variable X is given by:

$$H[X] \equiv -\sum_{x \in \mathcal{X}} \Pr(x) \log_2(\Pr(x)) \ . \qquad (4)$$

Using base-2 logs gives us units of *bits*.

Examples

- **Fair Coin:** $\Pr(h) = \frac{1}{2}, \Pr(t) = \frac{1}{2}$. $H = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$ bit.
- **Biased Coin:** $\Pr(h) = 0.6, \Pr(t) = 0.4$.
  $H = -0.6\log_2 0.6 - 0.4\log_2 0.4 = 0.971$ bits.
- **More Biased Coin:** $\Pr(h) = 0.9, \Pr(t) = 0.1$.
  $H = -0.9\log_2 0.9 - 0.1\log_2 0.1 = 0.469$ bits.

We now consider various interpretations for the entropy.

Average Surprise

- $-\log_2 \Pr(x)$ may be viewed as the *surprise* associated with the outcome x.
- Thus, $H[X]$ is the average, or expected value, of the surprise:

$$H[X] = \sum_x \left[-\log_2 \Pr(x) \right] \Pr(x) \ .$$

- The more surprised you are about a measurement, the more informative it is.
- The greater $H[X]$, the more informative, on average, a measurement of X is.

Difficulty of Guessing

For the next few slides, we'll focus on two examples.

1. A random variable X with four equally likely outcomes: $\Pr(a) = \Pr(b) = \Pr(c) = \Pr(d) = \frac{1}{4}$.
2. A random variable Y with four outcomes: $\Pr(\alpha) = \frac{1}{2}, \Pr(\beta) = \frac{1}{4}$, $\Pr(\gamma) = \frac{1}{8}, \Pr(\delta) = \frac{1}{8}$.

What is the optimal strategy for guessing (via yes-no questions) the outcome of a random variable?

- In general, try to divide the probability in half with each guess.
- Example: Guessing X:
  1. "is X equal to a or b?"
  2. If yes, "is $X = a$?" If no, "is $X = c$?"
- Using this strategy, it will always take 2 guesses.
- $H[X] = 2$. Coincidence???

Guessing games, continued

What's the best strategy for guessing Y?

$\Pr(\alpha) = \frac{1}{2}, \Pr(\beta) = \frac{1}{4}, \Pr(\gamma) = \frac{1}{8}, \Pr(\delta) = \frac{1}{8}$.

1. Is it α? If yes, then done, if no:
2. Is it β? If yes, then done, if no:
3. Is it γ? Either answer, done.

Ave # of guesses = $\frac{1}{2}(1) + \frac{1}{4}(2) + \frac{1}{4}(3) = 1.75$.

Not coincidentally, $H[Y] = 1.75$!!

General result: Average number of yes-no questions needed to guess the outcome of X is between $H[X]$ and $H[X] + 1$.

- This is consistent with the interpretation of H as uncertainty.
- If the probability is concentrated more on some outcomes than others, we can exploit this regularity to make more efficient guesses.

Coding

- A *code* is a mapping from a set of symbols to another set of symbols.
- Here, we are interested in a code for the possible outcomes of a random variable that is as short as possible while still being decodable.
- Strategy: use short code words for the more common occurrences of X.
- This is identical to the strategy for guessing outcomes.

Example: Optimal binary code for Y:

$$\alpha \longrightarrow 1\ , \quad \beta \longrightarrow 01$$
$$\gamma \longrightarrow 001\ , \quad \delta \longrightarrow 000$$

Note: This code is unambiguously decodable:

$$0110010000000101 = \beta\alpha\gamma\delta\delta\beta\beta$$

This type of code is called an *instantaneous* code.

Coding, continued

General Result: Average number of bits in optimal binary code for X is between $H[X]$ and $H[X]+1$.

This result is known as Shannon's noiseless source coding theorem or Shannon's first theorem.

- Thus, $H[X]$ is the average memory, in bits, needed to store outcomes of the random variable X.

Summary of interpretations of entropy

- $H[X]$ is *the* measure of uncertainty associated with the distribution of X.
- Requiring H to be a continuous function of the distribution, maximized by the uniform distribution, and independent of the manner in which subsets of events are grouped, uniquely determines H.
- $H[X]$ is the expectation value of the surprise, $-\log_2 \Pr(x)$.
- $H[X] \ \leq$ Average number of yes-no questions needed to guess the outcome of $X \ \leq \ H[X] + 1$.
- $H[X] \ \leq$ Average number of bits in optimal binary code for X $\leq \ H[X] + 1$.
- $H[X] = \lim N \to \infty \, \frac{1}{N} \times$ average length of optimal binary code of N copies of X.

Joint and Conditional Entropies

Joint Entropy

- $H[X,Y] \equiv -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \Pr(x,y)\log_2(\Pr(x,y))$
- $H[X,Y]$ is the uncertainty associated with the outcomes of X **and** Y.

Conditional Entropy

- $H[X|Y] \equiv -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \Pr(x,y)\log_2 \Pr(x|y)$.
- $H[X|Y]$ is the average uncertainty of X given that Y is known.

Relationships

- $H[X,Y] = H[X] + H[Y|X]$
- $H[Y|X] = H[X,Y] - H[X]$
- $H[Y|X] \neq H[X|Y]$

Mutual Information

Definition

- $I[X;Y] \,=\, H[X] - H[X|Y]$
- $I[X;Y]$ is the average reduction in uncertainty of X given knowledge of Y.

Relationships

- $I[X;Y] \,=\, H[X] - H[X|Y]$
- $I[X;Y] \,=\, H[Y] - H[Y|X]$
- $I[X;Y] \,=\, H[Y] + H[X] - H[X,Y]$
- $I[X;Y] = I[Y;X]$

Example 1

Two independent, fair coins, C_1 and C_2.

C_1	C_2	
	h	t
h	$\frac{1}{4}$	$\frac{1}{4}$
t	$\frac{1}{4}$	$\frac{1}{4}$

- $H[C_1] = 1$ and $H[C_2] = 1$.
- $H[C_1, C_2] = 2$.
- $H[C_1|C_2] = 1$. Even if you know what C_2 is, you're still uncertain about C_1.
- $I[C_1; C_2] = 0$. Knowing C_1 does not reduce your uncertainty of C_2 at all.
- C_1 carries no information about C_2.

Example 2

Weather (rain or sun) yesterday W_0 and weather today W_1.

	W_1	
W_0	r	s
r	$\frac{5}{8}$	$\frac{1}{8}$
s	$\frac{1}{8}$	$\frac{1}{8}$

- $H[W_0] = 0.811$ and $H[W_1] = 0.811$.
- $H[W_0, W_1] = 1.549$.
- Note that $H[W_0, W_1] \neq H[W_0] + H[W_1]$.
- $H[W_1|W_0] = 0.738$.
- $I[W_0; W_1] = 0.074$. Knowing the weather yesterday, W_0, reduces your uncertainty about the weather today W_1.
- W_0 carries 0.074 bits of information about W_1.

Example 2, continued

- Note: The above statistics are consistent with the perfectly periodic pattern: $\cdots rrrrrrssrrrrrrssrrrrrrss\cdots$.
- How could we detect if this was the actual pattern?

Application: Maximum Entropy

- A common technique in statistical inference is the **maximum entropy method**.
- Suppose we know a number of average properties of a random variable. We want to know what distribution the random variable comes from.
- This is an underspecified problem. What to do?
- Choose the distribution that maximizes the entropy while still yielding the correct average values.
- This is usually accomplished by using Lagrange multipliers to perform a constrained maximization.
- The justification for the maximum entropy method is that it assumes no information beyond what is already known in the form of the average values.

Another Application: Mutual Information

- In settings in which one wants to design a maximally predictive model, one often adjusts parameters to maximize the mutual information between input variables and those variables that are to be predicted.
- A particularly nice example of this is the *Information Bottleneck Method*. N. Tishby, F. Pereira and W. Bialek, In Proc. 37th Annual Allerton Conf. Eds.: B. Hajek and R. S. Sreenivas (1999) University of Illinois, physics/0004057.
- Also see, S. Still and W. Bialek. How Many Clusters? An Information Theoretic Perspective. *Neural Computation*, 16(12):2483-2506, 2004.

Information Theory Summary

- Information theory quantifies how much uncertainty is associated with a probability distribution.
- The entropy also measures amount of memory needed to store outcomes of a random variable.
- Information theory provides a natural language for working with probabilities.
- Information theory is *not* a theory of semantics or meaning.

Part V

Information Theory Applied to Stochastic Processes: Entropy Rate, Excess Entropy, Transient Information

Information Theory: Part II

Applications to Stochastic Processes

- We now consider applying information theory to a long sequence of measurements.

$$\cdots 00110010010101101001100111010110 \cdots$$

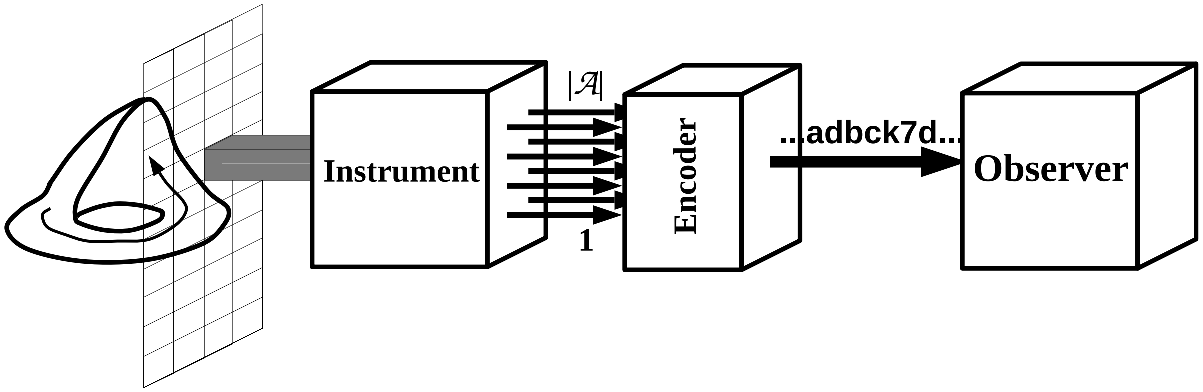
- In so doing, we will be led to two important quantities
  1. **Entropy Rate:** The irreducible randomness of the system.
  2. **Excess Entropy:** A measure of the complexity of the sequence.

Context: Consider a long sequence of discrete random variables. These could be:

1. A long time series of measurements
2. A symbolic dynamical system
3. A one-dimensional statistical mechanical system

The Measurement Channel

- Can also picture this long sequence of symbols as resulting from a generalized measurement process:



- On the left is “nature”—some system’s state space.
- The act of measurement projects the states down to a lower dimension and discretizes them.
- The measurements may then be encoded (or corrupted by noise).
- They then reach the observer on the right.
- Figure source: Crutchfield, “Knowledge and Meaning ... Chaos and Complexity.” In Modeling Complex Systems. L. Lam and H. C. Morris, eds. Springer-Verlag, 1992: 66-10.

Stochastic Process Notation

- Random variables S_i, $S_i = s \in \mathcal{A}$.
- Infinite sequence of random variables: $\overleftrightarrow{S} = \ \ldots \ S_{-1}\, S_0\, S_1\, S_2\ \ldots$
- Block of L consecutive variables: $S^L \ = \ S_1, \ldots, S_L$.
- $\Pr(s_i, s_{i+1}, \ldots, s_{i+L-1}) \ = \ \Pr(s^L)$
- Assume translation invariance or stationarity:

$$\Pr(\, \mathrm{s_i}, \mathrm{s_{i+1}}, \cdots, \mathrm{s_{i+L-1}} \,) \ = \ \Pr(\, \mathrm{s_1}, \mathrm{s_2}, \cdots, \mathrm{s_L} \,) \ .$$

- Left half (“past”): $\overleftarrow{s} \equiv \cdots S_{-3}\, S_{-2}\, S_{-1}$
- Right half (“future”): $\overrightarrow{s} \equiv S_0\, S_1\, S_2 \cdots$

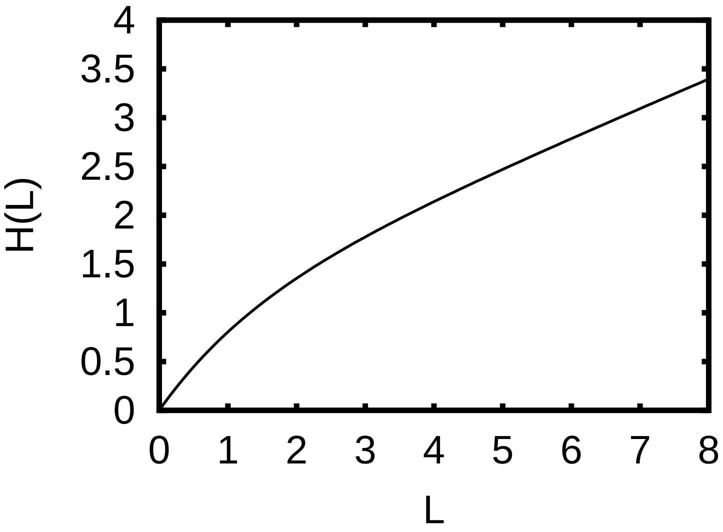
$$\cdots 11010100101101010101001001010010 \cdots$$

Entropy Growth

- Entropy of L-block:

$$H(L) \equiv - \sum_{s^L \in \mathcal{A}^L} \Pr(s^L) \log_2 \Pr(s^L) \; .$$

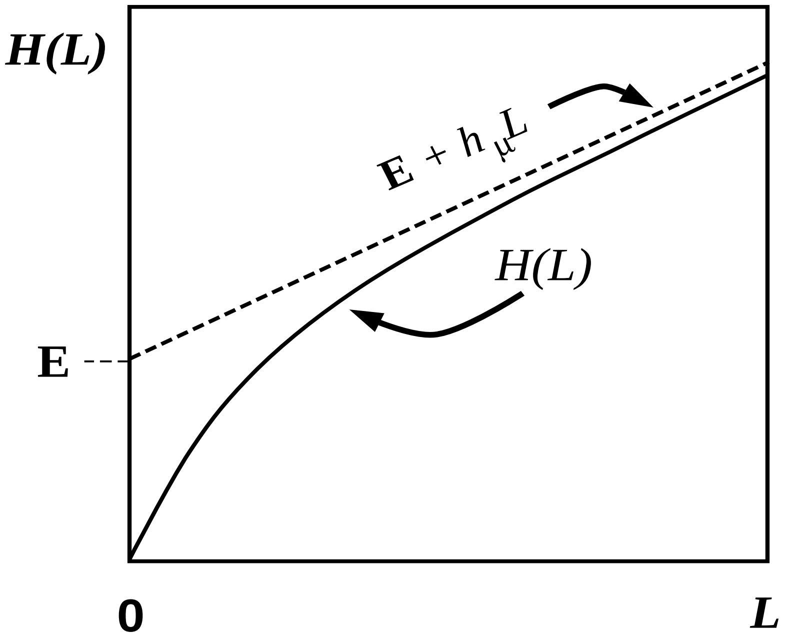
- $H(L) =$ average uncertainty about the outcome of L consecutive variables.



- $H(L)$ increases monotonically and asymptotes to a line
- We can learn a lot from the shape of $H(L)$.

Entropy Rate

- Let’s first look at the slope of the line:



- Slope of $H(L)$: $h_\mu(L) \equiv H(L) - H(L{-}1)$
- Slope of the line to which $H(L)$ asymptotes is known as the *entropy rate:*

$$h_\mu = \lim_{L\to\infty} h_\mu(L).$$

Entropy Rate, continued

- Slope of the line to which $H(L)$ asymptotes is known as the *entropy rate:*

$$h_\mu = \lim_{L\to\infty} h_\mu(L).$$

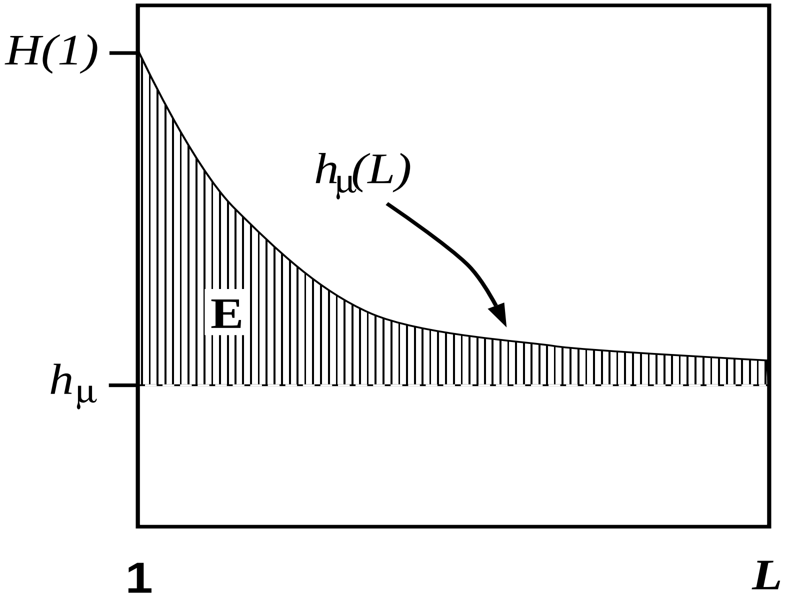
- $h_\mu(L) = H[S_L|S_1 S_1 \ldots S_{L-1}]$
- I.e., $h_\mu(L)$ is the average uncertainty of the next symbol, given that the previous L symbols have been observed.

Interpretations of Entropy Rate

- Uncertainty per symbol.
- Irreducible randomness: the randomness that persists even after accounting for correlations over arbitrarily large blocks of variables.
- The randomness that cannot be "explained away".
- Entropy rate is also known as the Entropy Density or the Metric Entropy.
- $h_\mu =$ Lyapunov exponent for many classes of 1D maps.
- The entropy rate may also be written: $h_\mu = \lim_{L\to\infty} \frac{H(L)}{L}$.
- h_μ is equivalent to thermodynamic entropy.
- These limits exist for all stationary processes.

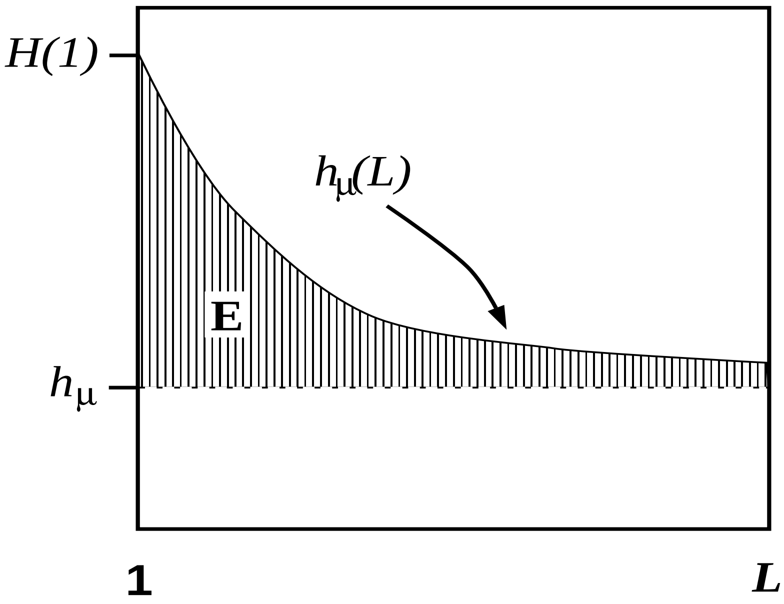
How does $h_\mu(L)$ approach h_μ?

- For finite L , $h_\mu(L) \geq h_\mu$. Thus, the system appears more random than it is.



- We can learn about the complexity of the system by looking at *how* the entropy density converges to h_μ.

The Excess Entropy



- The **excess entropy** captures the nature of the convergence and is defined as the shaded area above:

$$\mathbf{E} \equiv \sum_{L=1}^{\infty} [h_\mu(L) - h_\mu] \ .$$

- $\mathbf{E}$ is thus the total amount of randomness that is “explained away” by considering larger blocks of variables.

Excess Entropy: Other expressions and interpretations

Mutual information

- One can show that $\mathbf{E}$ is equal to the mutual information between the “past” and the “future”:

$$\mathbf{E} = I(\overleftarrow{S};\overrightarrow{S}) \equiv \sum_{\{\overleftrightarrow{s}\}} \Pr(\overleftrightarrow{s}) \log_2 \left[\frac{\Pr(\overleftrightarrow{s})}{\Pr(\overleftarrow{s})\Pr(\overrightarrow{s})} \right] \ .$$

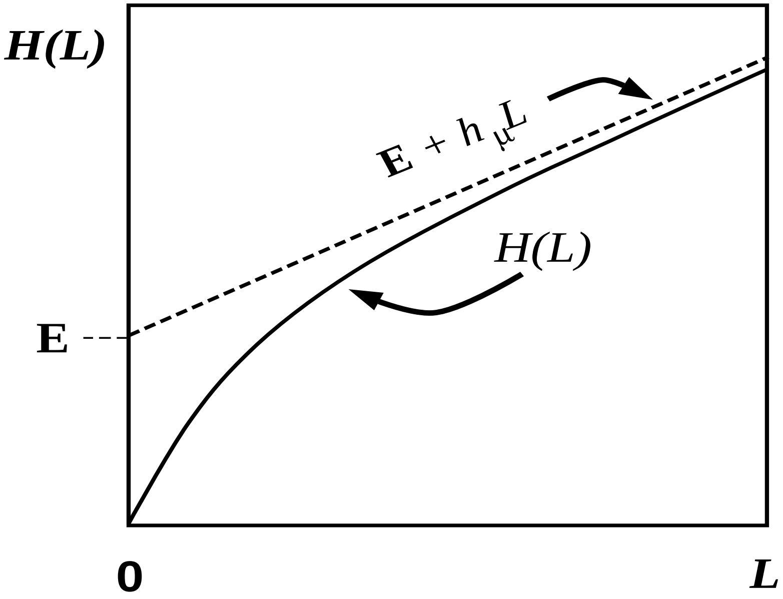
- $\mathbf{E}$ is thus the amount one half “remembers” about the other, the reduction in uncertainty about the future given knowledge of the past.

- Equivalently, $\mathbf{E}$ is the “cost of amnesia:” how much more random the future appears if all historical information is suddenly lost.

Excess Entropy: Other expressions and interpretations

Geometric View

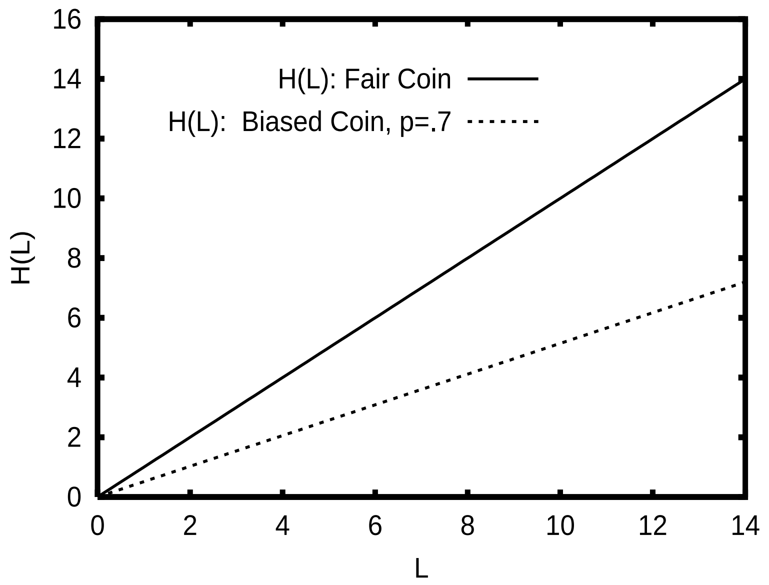
- $\mathbf{E}$ is the y-intercept of the straight line to which $H(L)$ asymptotes.
- $\mathbf{E} = \lim_{L\to\infty} \left[H(L) - h_\mu L \right]$.



Excess Entropy Summary

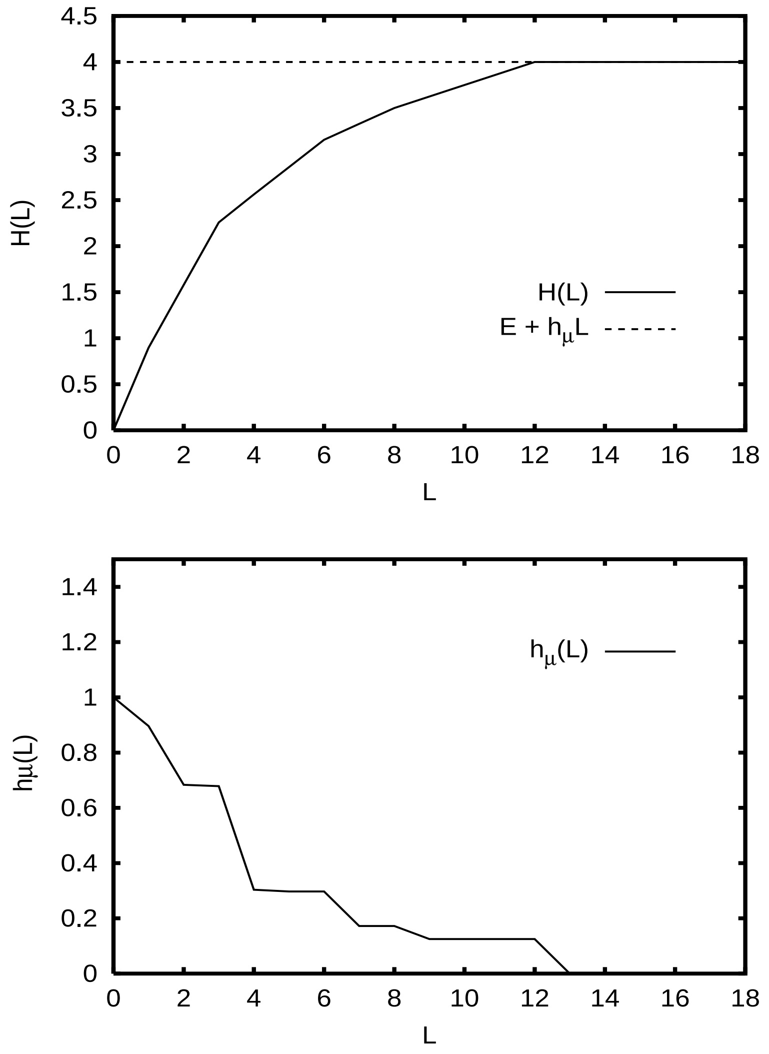
- Is a structural property of the system — measures a feature complementary to entropy.
- Measures memory or spatial structure.
- Lower bound for statistical complexity, minimum amount of information needed for minimal stochastic model of system

Example I: Fair Coin



- For fair coin, $h_\mu = 1$.
- For the biased coin, $h_\mu \approx 0.8831$.
- For both coins, $\mathbf{E} = 0$.
- Note that two systems with different entropy rates have the same excess entropy.

Example II: Periodic Sequence



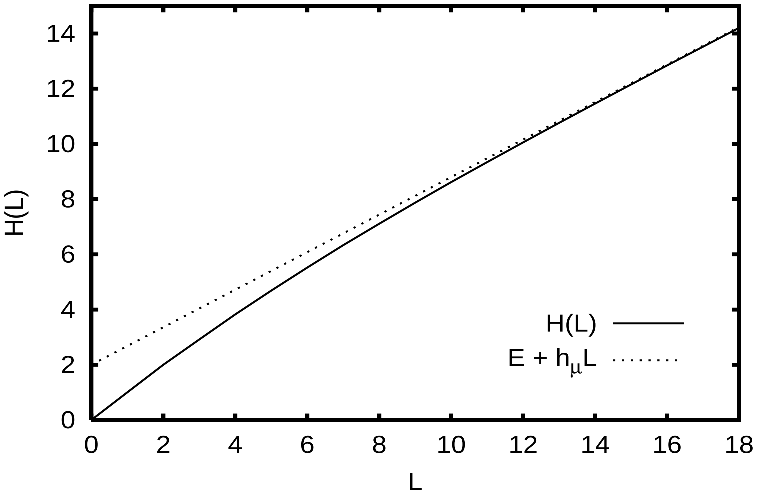
- Sequence: $\ldots 1010111011101110 \ldots$

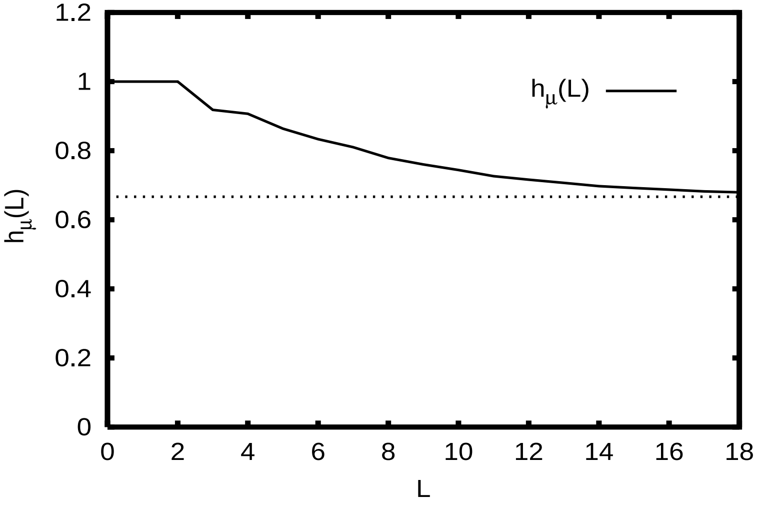
Example II, continued

- Sequence: $\ldots 1010111011101110 \ldots$
- $h_\mu \approx 0$; the sequence is perfectly predictable.
- $\mathbf{E} = \log_2 16 = 4$: four bits of phase information
- For any period-p sequence, $h_\mu = 0$ and $\mathbf{E} = \log_2 p$.

For more than you probably ever wanted to know about periodic sequences, see Feldman and Crutchfield, Synchronizing to Periodicity: The Transient Information and Synchronization Time of Periodic Sequences. *Advances in Complex Systems*. **7**(3-4): 329-355, 2004.

Example III: Random, Random, XOR





- Sequence: two random symbols, followed by the XOR of those symbols.

Example III, continued

- Sequence: two random symbols, followed by the XOR of those symbols.
- $h_\mu = \frac{2}{3}$; two-thirds of the symbols are unpredictable.
- $\mathbf{E} = \log_2 4 = 2$: two bits of phase information.
- For many more examples, see Crutchfield and Feldman, Chaos, 15: 25-54, 2003.

Excess Entropy: Notes on Terminology

All of the following terms refer to essentially the same quantity.

- **Excess Entropy:** Crutchfield, Packard, Feldman
- **Stored Information:** Shaw
- **Effective Measure Complexity:** Grassberger, Lindgren, Nordahl
- **Reduced (Rényi) Information:** Szépfalusy, Györgyi, Csordás
- **Complexity:** Li, Arnold
- **Predictive Information:** Nemenman, Bialek, Tishby

Excess Entropy: Selected References and Applications

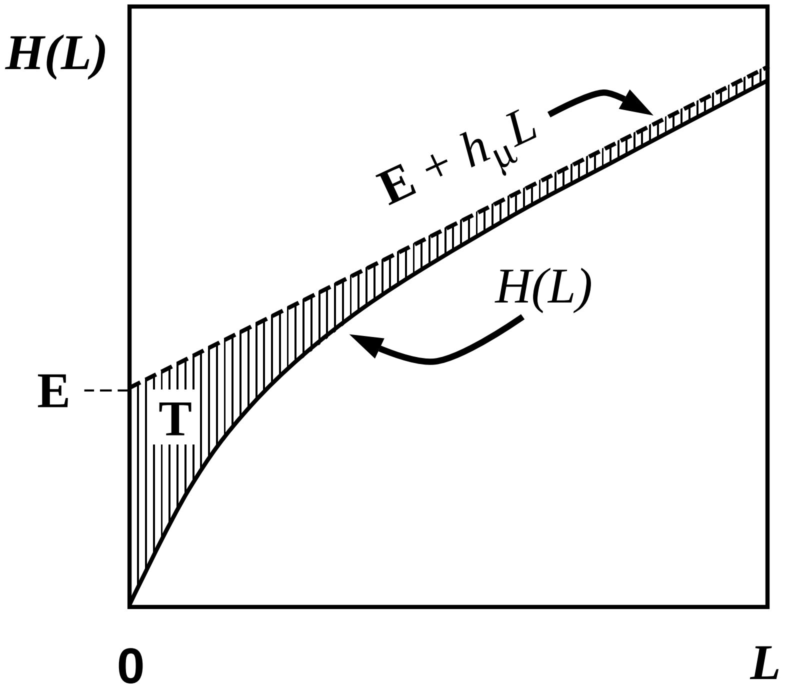
- Crutchfield and Packard, *Intl. J. Theo. Phys*, 21:433-466. (1982); *Physica D*, 7:201-223, 1983. [Dynamical systems]
- Shaw, “The Dripping Faucet ..., ” Aerial Press, 1984. [A dripping faucet]
- Grassberger, *Intl. J. Theo. Phys*, 25:907-938, 1986. [Cellular automata (CAs), dynamical systems]
- Szépfalusy and Györgyi, *Phys. Rev. A*, 33:2852-2855, 1986. [Dynamical systems]
- Lindgren and Nordahl, *Complex Systems*, 2:409-440. (1988). [CAs, dynamical systems]
- Csordás and Szépfalusy, *Phys. Rev. A*, 39:4767-4777. 1989. [Dynamical Systems]
- Li, *Complex Systems*, 5:381-399, 1991.
- Freund, Ebeling, and Rateitschak, *Phys. Rev. E*, 54:5561-5566, 1996.
- Feldman and Crutchfield, SFI:98-04-026, 1998. Crutchfield and Feldman, *Phys. Rev. E* 55:R1239-42. 1997. [One-dimensional Ising models]

Excess Entropy: Selected References and Applications, continued

- Feldman and Crutchfield. *Physical Review E*, 67:051104. 2003. [Two-dimensional Ising models]
- Feixas, et al, *Eurographics*, Computer Graphics Forum, 18(3):95-106, 1999. [Image processing]
- Ebeling. *Physica D*, 1090:42-52. 1997. [Dynamical systems, written texts, music]
- Bialek, et al, *Neur. Comp.*, 13:2409-2463. 2001. [Long-range 1D Ising models, machine learning]

Transient Information $\mathbf{T}$

- $\mathbf{T} \equiv \sum_{L=1}^{\infty} [\mathbf{E} + h_\mu L - H(L)]$.
- $\mathbf{T}$ is related to the total uncertainty experienced while synchronizing to a process.



- The shaded area is the transient information T.
- $\mathbf{T}$ measures how difficult it is to synchronize to a sequence.

Some Applications in Agent-Based Modeling Settings

1. If an agent doesn't have sufficient memory, its environment will appear more random. In a quantitative sense, regularities that are missed (as measured by the excess entropy) are converted into randomness (as measured by the entropy rate).

   - Crutchfield and Feldman, Synchronizing to the Environment: Information Theoretic Constraints on Agent Learning. *Advances in Complex Systems.* 4. 251–264. 2001.

2. The average-case difficulty for an agent to synchronize to a periodic environment is measured by the transient information.

   - Feldman and Crutchfield. Synchronizing to a Periodic Signal: The Transient Information and Synchronization Time of Periodic Sequences. *Advances in Complex Systems.* 7. 329–355. 2004.

   More on this in Part XIII.

Some Applications in Agent-Based Modeling Settings, continued

3. More generally it seems likely that the entropy and mutual information are useful tools for quantifying

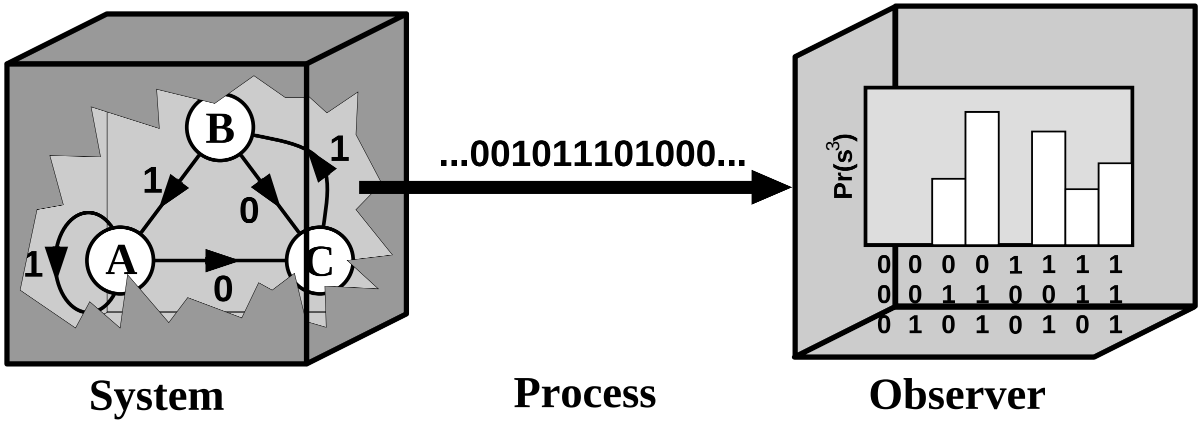
    (a) properties of agents: e.g., how much memory they have

    (b) the behavior of agents: e.g, how unpredictably they act

    (c) properties of the environment: e.g., how structured it is

Estimating Probabilities

- $\mathbf{E}$ and h_μ can be estimated empirically by observing a process.



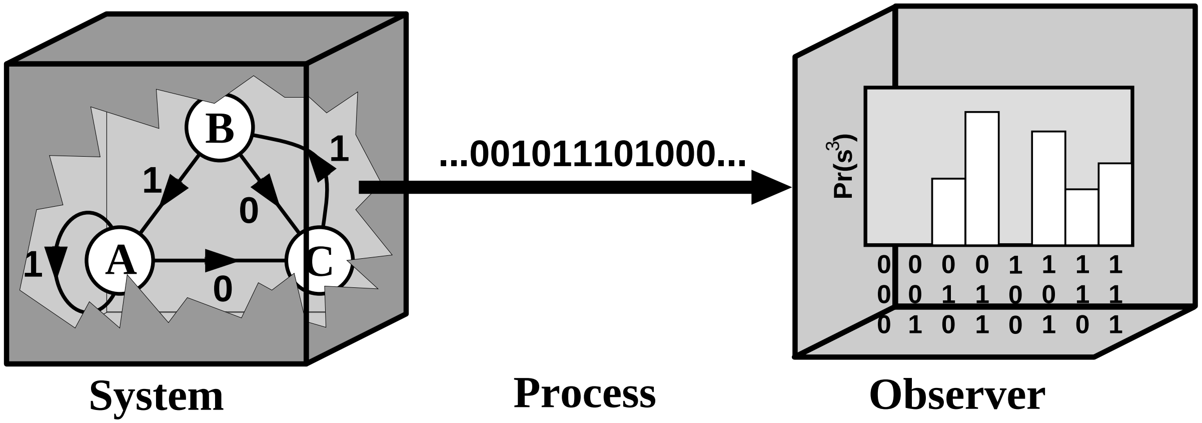
- One simply forms histograms of occurrences of particular sequences and uses these to estimate $\Pr(s^L)$, from which $\mathbf{E}$ and h_μ may be readily calculated.

However, this will lead to a biased under-estimate for h_μ. For more sophisticated and accurate ways of inferring h_μ, see, e.g.,

- Schürmann and Grassberger. Chaos 6:414-427. 1996.
- Nemenman. `http://arXiv.org/physics/0207009`. 2002.

A look ahead

- Note that the observer sees measurement symbols: 0's and 1's.



- It doesn't see inside the "black box" of the system.
- In particular, it doesn't see the internal, hidden states of the system, A, B, and C.
- Is there a way an observer can infer these hidden states?
- What is the meaning of *state*?

Part VI

Extensions to Shannon Entropy, Rényi Entropyies, Multifractals, Tsallis Entropies

Extensions to Shannon Entropy

- One of the requirements on the Shannon entropy H that is used to derive it is that H is independent of the way we group probabilities.
- Let's state this more precisely. We'll do so via an example.
- Consider the random variable X that can take on three outcomes, a, b, and c:
- $\Pr(a) = 1/2$, $\Pr(b) = 1/4$, and $\Pr(c) = 1/4$.
- It turns out that $H[X] = 3/2$.
- We can also view this as follows: Y can be a or Z, each with probability $1/2$. And Z can be b or c with probability $1/2$.
- $H[Y] = 1$, and $H[Z] = 1$.
- $H[X] = H[Y] + \frac{1}{2}H[Z]$.
- This last condition is an example of requiring H be independent of the way we group probabilities.

Rényi Entropy

- Let's relax the condition that H be independent of grouping.
- But still require that the entropy of independent variables be additive:

$$\begin{aligned} \mathrm{Pr}(X,Y) = \mathrm{Pr}(X)\mathrm{Pr}(Y) \quad &\Longrightarrow \\ H[X,Y] = H[X] + H[Y] \quad & . \end{aligned}$$

- The result is a one-parameter family, the Rényi entropies:

$$H_q \equiv \frac{1}{q-1} \log_2 \sum_i p_i^q \ . \tag{5}$$

- This can be rewritten in the following, slightly less odd-looking way.

$$H_q = \frac{1}{q-1} \log_2 \sum_i p_i p_i^{q-1} \ . \tag{6}$$

$$H_q = \frac{1}{q-1} \log_2 \langle p_i^{q-1} \rangle \ . \tag{7}$$

Rényi Entropies: Properties and Comments

- H_q is a non-increasing function of q.
- H_1 is the Shannon entropy.
- H_0 is the topological entropy, the log of the number of states.
- There are coding theorems for Rényi entropy. Campbell. *Information and Control.* 8:423. 1966; Aggarwal and Bansal, `arXiv.cs.IT/0607029`. 2006.

Rényi and Thermodynamics

- The Rényi entropy allows one to apply the formalism of thermodynamics to any probability distribution. q plays a role similar to inverse temperature.

Escort Distributions

- Given a set of probabilities, we can always make a new set of probabilities as follows:

$$p_i \longrightarrow \frac{p_i^\beta}{Z} \ .$$

- β is a number that acts like $1/\mathrm{Temperature}$.
- $\beta = 1$: initial distribution
- $\beta = 0$: all states equally likely $\Rightarrow T = \infty$.
- $\beta = \infty$: only most probable state remains. This is the $T = 0$ “ground state.”
- $\beta = -\infty$: only least probable state remains. This is the $T = 0^-$ “anti-ground” state.
- Loosely speaking, the Rényi entropy can be thought of the average surprise of the escort distribution with $\beta = q - 1$.
- The parameter β allows one to probe different regions of the distribution.

Thermodynamic Formalism

- The ideas on the previous slide can be extended in an elegant and fun way to apply thermodynamics to any probability distribution.
- This goes by many names; thermodynamic formalism, $S(u)$, $f(\alpha)$, multifractals, fluctuation spectrum, large deviation theory.
- This is a well developed, well understood approach. It is very enticing and very cool.
- In my experience, this approach doesn't speak directly to complexity or pattern, largely because thermodynamics doesn't have direct measures of complexity.
- For example, a biased coin (i.e. no correlations), has a "multifractal spectrum."
- This doesn't mean the multifractals are uninteresting. They are a natural way of quantifying the frequency with which deviations from typical behavior occurs.

Thermodynamic Formalism References

There are many confusing things written about the thermodynamic formalism. Some clear references:

- Young and Crutchfield. *Chaos, Solitons, and Fractals*. **4**:5. 1993.
- Beck and Schlögl, *Thermodynamics of Chaotic Systems*. Cambridge University Press. 1993.

Note: If you wish to estimate the fluctuation spectrum $S(u)$, a good way to do it is to first estimate the ϵ-machine and then calculate $S(u)$. Numerically calculating $S(u)$ directly can be inaccurate. See Young and Crutchfield.

Tsallis Entropy

- Define the following generalized entropy

$$S_q \equiv \frac{1 - \sum_i p^q}{q-1} \ . \tag{8}$$

- This q is not the same as Rényi's q.
- S_q has the property that if $\mathrm{Pr}(X,Y) = \mathrm{Pr}(X)\mathrm{Pr}(Y)$, then:

$$S_q[X,Y] = S_q[X] + S_q[Y] + (1{-}q)S_q[X]S_q[Y] \ . \tag{9}$$

- I.e., S_q is not additive for independent events.
- One can generate a statistical mechanics and thermodynamics using Eq. (8) as a starting point.

Tsallis Entropy: Doubts and Concerns

- However, it is hard to see how a non-additive entropy can be physical.
- It has been claimed that S_q works well for systems with strong correlations. But it seems to me that the non-additivity creates spurious correlations rather than measuring correlations that are really there.
- My sense is that q is basically a fitting parameter. I don't know that it has a clear physical or mathematical meaning.
- Overall, I don't understand why Tsallis entropy is a big deal. I think this is the position of most, but by no means all, physicists.

Tsallis Entropy, references

But... you should read the papers and decide for yourself.

Some reviews:

- Tsallis. *Physica D*, **193**:3. 2004. `http://arxiv.org/cond-mat/0403012`.
- Tsallis, et al. `arXiv.org/cond-mat/0309093`. 2003.
- Tsallis and Brigatti, *Continuum Mechanics & Thermodynamics*, **16**:223 `arXiv.org/cond-mat/0305606`. 2004.

Some strenuous (and entertaining) critiques, and responses:

- Grassberger. *Physical Review Letters*, 95. 140601. 2005. `http://arxiv.org/cond-mat/0508110`
- Responses to Grassberger:
  - Robledo. `arxiv.org/cond-mat/0510293`
  - Tsallis. `arxiv.org/cond-mat/0511213`

- Nauenberg. *Physical Review E*. **67**:036114. 2003.
  `arxiv.org/cond-mat/0210561`
- Responses to Nauenberg and discussion:
  - Tsallis. *Physical Review E*. **69**:038102. 2004.
    `arxiv.org/cond-mat/0304696`
  - Nauenberg. *Physical Review E*. **69**:038102. 2004
    `arxiv.org/cond-mat/0305365`

See also:

- `www.cbpf.br/GrupPesq/StatisticalPhys/biblio.htm`
  for an extensive bibliography on Tsallis entropy.

Part VII

A Sketch of Computation Theory: Machines, Languages, and the Computation Hierarchy

A (Mostly) Informal Introduction to Computation Theory

- Computation theory is a different, more structural and less statistical approach to complexity, emergence, organization.
- Computation theory can be very elegant, rigorous, and mathematical.
- But I'll present little of the formalism. I think the math can obscure some of the basic ideas, which are really quite simple.

We'll begin with some examples in the form of a game:

- I'll give you the specification for a set
- I'll then show you an object, and you need to tell me if it's in the set or not

Example 1

The set $\mathcal{L}$ consists of all sequences of 0's and 1's of any length, except for those that have two 00's in a row.

Accept all sequences of 1's and 0's except for those which have two or more 0's in a row.

1110101101

1101101001

110110101011

Example 2:

The set $\mathcal{L}$ consists of all sequences of correctly balanced parentheses.

$$(\;\;(\;\;)\;\;(\;\;)\;\;)$$

$$(\,(\,)\,(\,)\,(\,)\,)\,(\,(\,)\,)$$

$$\Bigg(\Bigg(\Bigg) \Bigg(\Bigg(\Bigg) \Bigg(\Bigg) \Bigg(\Bigg) \Bigg(\Bigg)$$

Example 3:

The set $\mathcal{L}$ consists of all sequences of 0's and 1's, except for those that contain a **prime** number of consecutive 0's!

1100011000001

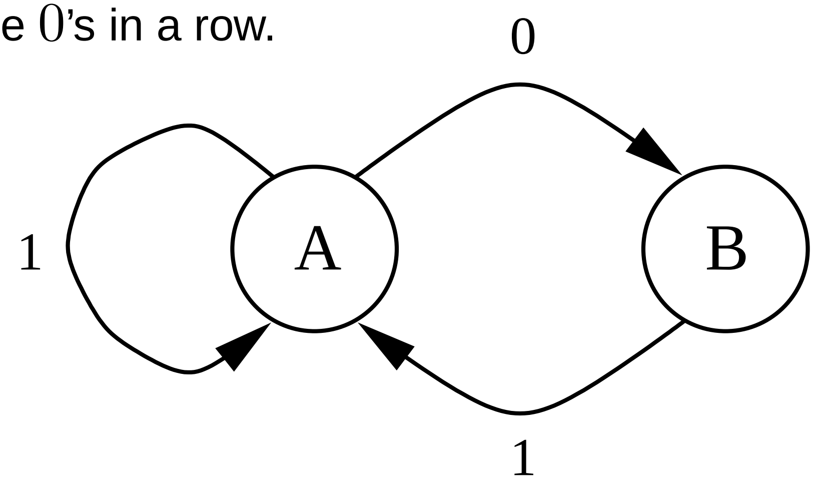
11000011

111000000000001

$$11 \quad \overbrace{00\cdots00}^{1031\text{ elements}} \quad 11$$

What to learn from the examples

- There are qualitative differences between the procedures you just used to identify the strings on the previous slides.
- These distinctions lie at the heart of computation theory.
- We'll start by focusing on example 1.
- Your task was to accept all sequences of 1's and 0's except for those which have two or more 0's in a row.



- Sequence is OK if there exists a path through this machine
- Example: 1011001 is not in the set.

Finite State Machines

- The mathematical object on the previous page is known as a **Finite State Machine** or a **Finite Automaton**.
- Note that this two-state machine can correctly identify arbitrarily long sequences.
- The machine is a finite representation of the infinite set $\mathcal{L}$.

Some terminology and definitions

- A **Language** $\mathcal{L}$ is a set of words (symbol strings) formed from an **Alphabet** $\mathcal{A}$.
- We'll always assume a binary alphabet, $\mathcal{A} = \{0, 1\}$.

Big Idea: There is a correspondence between the rules needed to generate or describe a language, and the type of machine needed to recognize it.

Regular Expressions

- A **Regular Expression** is a way of writing down rules that generate a language.
- To generate a regexp, start with the symbols in $\mathcal{A}$.
- You can make new expressions via the following operations: grouping, concatenating, logical OR (denoted $+$), and closure $*$.
- Closure means 0 or more concatenations.
- Examples:
  1. $(0+1)=\{0,1\}$
  2. $(0+1)^*=\{\epsilon,0,1,00,01,10,11,000,001,\ldots\}$
  3. $(01)^*=\{\epsilon,01,0101,010101,\ldots\}$
- (ϵ is the empty symbol.)

Regular Languages and FSM

- A language $\mathcal{L}$ is a **Regular Language** if and only if it can be generated by a regular expression.
- A puzzle: what is the regular expression that generates the language of example 1?

Two important results:

1. For any regular language, there is an FSM that recognizes it.
2. Any language generated by an FSM is regular.

Notes on terminology:

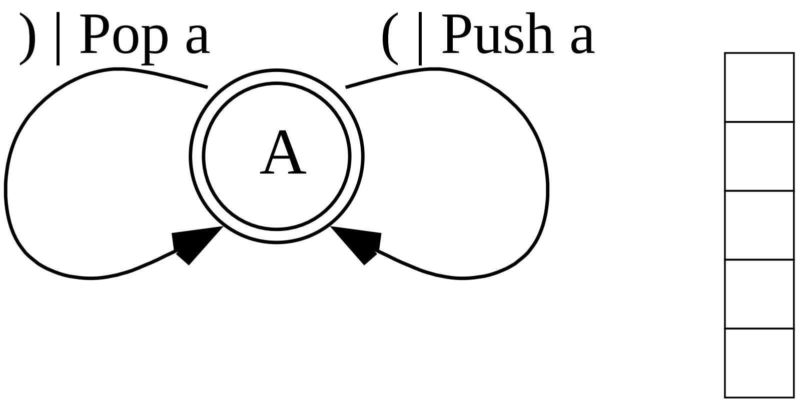
- A regular expression is a **rule**.
- A regular language is a **set**.
- A FSM is a **machine**.

Regular languages ↔ FSM's is the first example of the correspondence between sets and the procedures or machines needed to recognize them.

Revisiting Parentheses

- This example is different than the last—you can't scan left to right unless you remember stuff.
- There is no FSM that can recognize this language. The problem is that as the sequence grows in length, the number of states necessary also grows.
- This task requires infinite memory. However, the memory only needs to be organized in a simple way.
- The parentheses language can be recognized by a device known as a **Pushdown Automata**.
- Put an object on the stack if you see a left paren $($ and take it off if you see a right paren $)$.
- If the stack is empty after scanning the sequence, then it is ok.

Pushdown Automata



- This is the PDA for the parentheses example
- If you see a “(“, write (push) a symbol to the stack.
- If you see a “)“, erase (pop) a symbol from the stack.
- The machine can only write to the top of the stack.
- This PDA can recognize balanced parentheses of any length.

Context-Free Languages

- The languages recognized by PDA are **context-free languages.**
- Regular languages are generated sequentially—one symbol after the next.
- CFL's are generated by writing rules applied in parallel.
- For example, to generate the parentheses language, apply the following:

$$\begin{aligned} W &\rightarrow (V \\ V &\rightarrow (VV \ \text{ or } \ ) \end{aligned}$$

- Start with W. The set of all possible applications of the above rules give you the set of all possible balanced parentheses.
- For example:

$$W, \ (V, \ ((VV, \ (()V \ \ (()(VV \ \ (()()V \ \ (()())$$

CFL Terminology

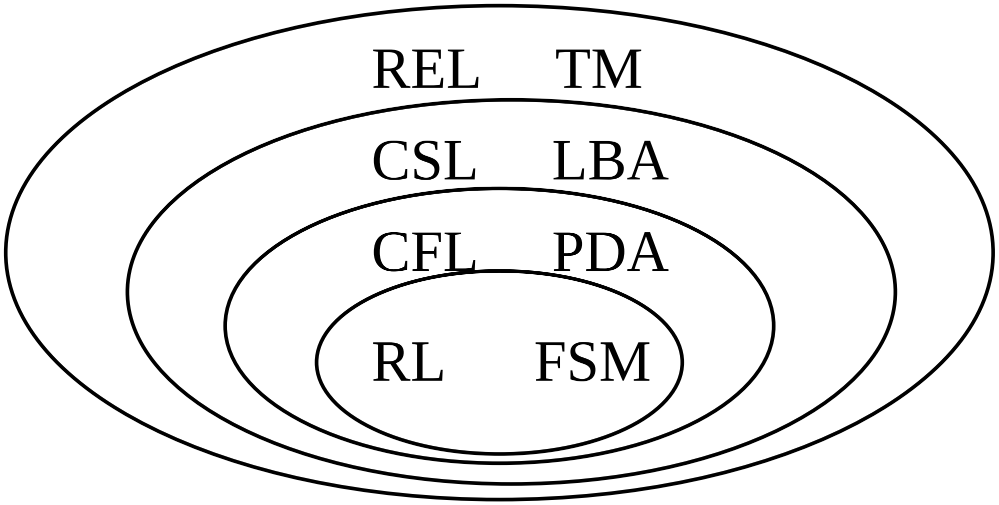
- $(,)$ are **terminals**, symbols in the alphabet $\mathcal{A}$.
- W, V are **variables**, symbols not in $\mathcal{A}$, to be eventually replaced by terminals.
- CFL's are context free in the sense that the production rule depends only on the variable, not on where the variable is in the string.

CFL Summary

- Every CFL can be recognized by a PDA, and every PDA produces a CFL.
- Also, FSM's are a proper subset of PDAs, and
- Regular Languages are a proper subset of CFL's
- We can thus divide languages into two classes, one of which is strictly more complex than the other.
- Are there even more complex languages? Yes ...

Chomsky Hierarchy

- The hierarchy continues:



- This hierarchy of languages/machines is known as the **Chomsky Hierarchy**.
- Each level in the hierarchy contains something new, and also contains all the languages at lower levels of the hierarchy.

Chomsky Hierarchy, terminology

- CSL = **Context Sensitive Language**. These are like CFL's, but allow transitions that depend on the position of the variable in the strings.
- LBA = **Linear Bounded Automata**. These are like PDA's, except:
  1. Controller can write anywhere on work tape.
  2. Work tape restricted to be a linear function of input.
- **Recursively Enumerable Languages** are those languages produced by an unrestricted grammar.
- An **Unrestricted Grammar** is like a CSL, but allows substitutions that shrink the length of the string.
- TM = **Turing Machines**. These are LBA's with linear tape restriction removed. These are the most powerful model of computation. (Example 3 requires a TM.) More on these later.

Chomsky Hierarchy, Conclusions

- Order languages (sets) by the type of machine needed to recognize elements of the language.
- There are qualitative difference between machines at different levels of the hierarchy.
- At lower levels of the hierarchy, there are algorithms for minimizing machines. (I.e., remove duplicate nodes.)
- The minimum machine can be viewed as a representation of the pattern contained in the language. The machine is a description of all the regularities.
- The size of the machine may be viewed as a measure of complexity.
- The machine itself reveals the “architecture” of the information processing.

Other computation theory notes

- It is possible to refine the Chomsky hierarchy with different sorts of machines. The result is a rich partial ordering of languages.
- To use computation theory as a basis for measuring complexity or structure, I think it's important to start at the bottom of the hierarchy and work your way up.

Computation Theory References

The basic material presented is quite standard and there are many references on it. Here are a few:

- Hopcroft and Ullman. Introduction to Automata Theory, Languages and Computation. Addison-Wesley. 1979. *A standard reference. Not my favorite, though. It's thorough and clear, but rather dense.*

- Brookshear. Theory of Computation: Formal Languages, Automata, and Complexity. Benjamin/Cummings. 1989. *I like this book. I find it much clearer than Hopcroft and Ullman.*

Computation theory applied to physical sequences

- Badii and Politi. Complexity: Hierarchical Structures and Scaling in Physics. Cambridge. 1997. *Excellent book, geared toward physics grad students. Closest thing to a textbook that covers topics similar to those I've covered throughout these lectures.*

- Bioinformatics textbooks?

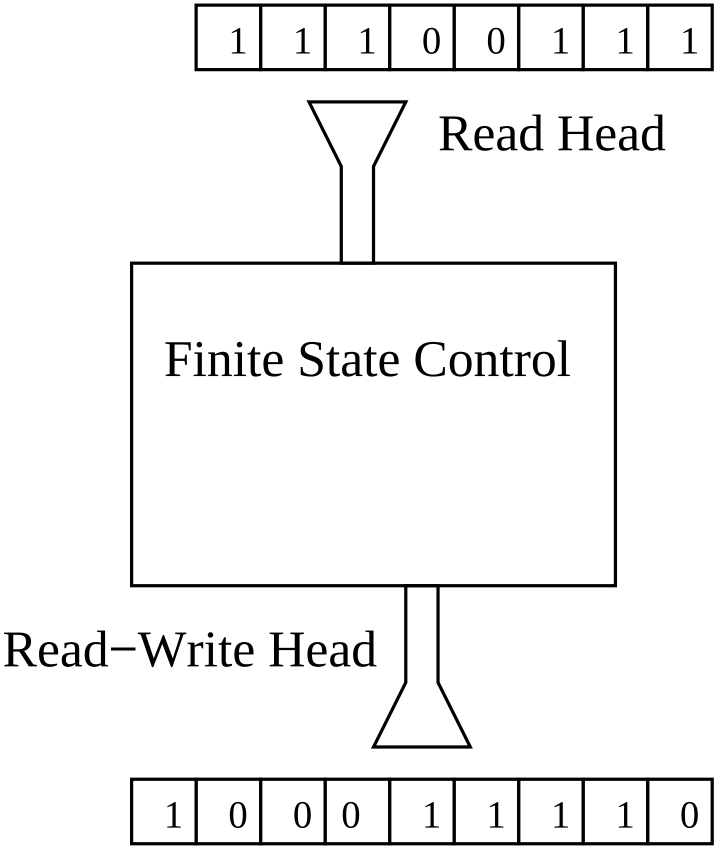
Part VIII

A Very Brief and Informal Introduction to Computability and Computational Complexity

A Very Brief and Informal

Introduction to Computability and Computational Complexity

- **Turing Machines** are at the top of the computational hierarchy.



- A TM has an input tape, a finite state controller, and a working tape. The finite state controller can read and write symbols from/to the working tape.

More about UTMs

- A universal TM is a TM that can simulate any other TM.
- A UTM is the most general model of computation. It can match the power of any other computational method.
- Thus, UTM's are equivalent to C programs, java programs, etc.
- TM's are also taken as being equivalent to algorithms.

More about UTMs, continued

- Recursively Enumerable Languages are recognized by UTMs
- However, there exists some languages that no machine can recognize! Why?
- The number of UTMs is countably infinite.
- The number of languages is uncountably infinite, since it is the set of all subsets of a countably infinite set.
- Thus # of Languages $>$ # of machines.

This has some profound and important implications. We can use this to show that some algorithms do not exist.

We will do so via a paradox

Paradoxes

Paradoxes are unavoidable with self-referencing systems. Examples:

- “I’m lying”
  1. If I’m lying, I’m telling the truth
  2. If I’m telling the truth, I’m lying
- The shortest integer that can’t be described in less than thirteen words.
- Consider the set of all sets that are not members of themselves Should this set be a member of itself?
- etc.

Halting

- One of the problems with UTMs is that sometimes they don't halt. They can get caught in loops.
- Let's see if we can come up with a method for determining if a UTM, with program P and input x will halt.
- This slide and the following is almost directly taken from Randy Wang's lecture on computability: `www.cs.princeton.edu/~rywang/99f126/slides.html`.
- Call this program $\mathrm{HALT}(P, x)$. It takes as input the program P and the output x.
- $\mathrm{HALT}(P, x)$ returns YES if $P(x)$ halts, and NO if it doesn't.
- Will assume that $\mathrm{HALT}(P, x)$ always halts—it does not go into loops.

Halting Problem, continued

Now, let's construct the strange program $\mathrm{XX}(P)$, as follows:

- $\mathrm{XX}(P)$ calls $\mathrm{HALT}(P, P)$.
- $\mathrm{XX}(P)$ halts if $\mathrm{HALT}(P, P)$ outputs NO.
- $\mathrm{XX}(P)$ infinite loops if $\mathrm{HALT}(P, P)$ outputs YES

In other words:

- If $P(P)$ does not halt, $XX(P)$ halts.
- If $P(P)$ halts, $XX(P)$ does not halt.

Let's call XX with *itself* as input

- If $XX(XX)$ does not halt, $XX(XX)$ halts
- If $XX(XX)$ halts, $XX(XX)$ does not halt

Both lead to a contradiction. Therefore, $\mathrm{HALT}(P, x)$ **cannot exist**.

Consequences of Halting Problem

- There does not exist an algorithm to determine if a UTM running program P with input x will halt.
- We say that the Halting problem is **uncomputable** or unsolvable.
- It turns out that many other problems are uncomputable as well.
- Of particular relevance to us: There does not exist an algorithm that will determine the shortest program that will output a given string.
- I.e., there is no general-purpose algorithm for optimal data compression.
- This means that measures of complexity or randomness based on minimal UTM representations are uncomputable.
- More generally, the existence of uncomputable problems means that we'll never be able to find algorithms for everything.

A Very Brief Discussion of Computational Complexity

- The computational complexity of an algorithm is the run time $T(N)$ needed for the algorithm to run, expressed as a function of the size of the problem N.
- The slower $T(N)$ grows with N, the more tractable (and less complex?) the problem is.
- For applications to physics, see, e.g., the work of Machta `www-unix.oit.umass.edu/~machta/`, and Moore, “Computational Complexity in Physics,” `www.santafe.edu/~moore/pubs/nato.html`.
- Computational Complexity is usually concerned with the time scaling of the *worst* cast scenario. The time scaling of the average case may be more relevant. Unlike comp complexity, information theory is concerned with *average* case behavior.
- There has recently been work in applying parallel computational complexity to physical models. This may be more relevant—nature is often parallel.

Part IX

An Introduction to Computational Mechanics

An Introduction to Computational Mechanics

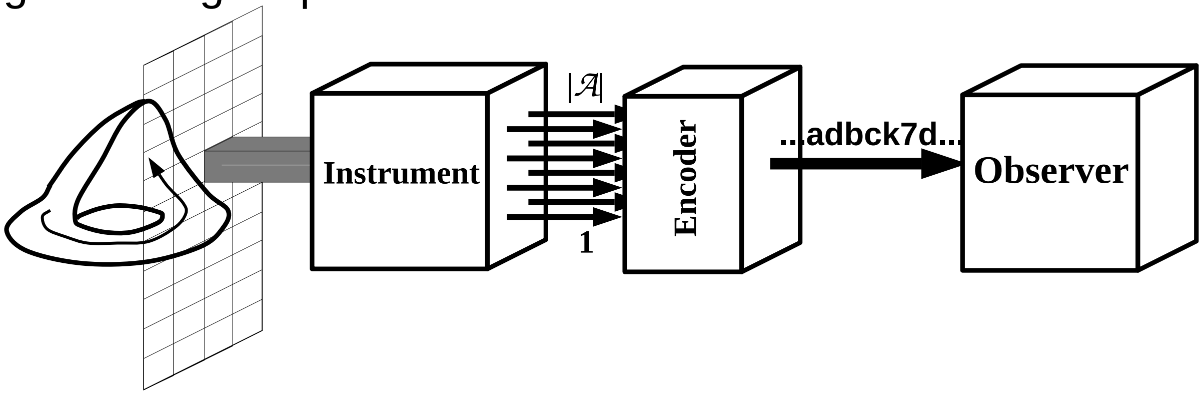
1. Computational Mechanics provides another way of measuring an object's complexity or regularities.
2. Unlike the excess entropy, computational mechanics makes use of the the models of formal computation to provide a direct, structural accounting of a system's intrinsic information processing.
3. Computational Mechanics lets us see how a system stores, transmits, and manipulates information.

Context:

- As before, we have a long sequence of symbols, $s_1, s_2, s_3, \cdots$, from a binary alphabet. Assume a stationary probability distribution over the sequence.

Measurement Channel

Consider again a long sequence of measurements:



- On the left is “nature”—some system’s state space.
- The act of measurement projects the states down to a lower dimension and discretizes them.
- They then reach the observer on the right.
- Figure source: Crutchfield, In Modeling Complex Systems. L. Lam and H.C. Morris, eds. Springer-Verlag, 1992: 66-10.
- Task: What can the observer infer about the intrinsic computation, the pattern or complexity, of the observed process?

An initial example: The Prediction Game

- Your task is to observe a sequence, and then come up with a way of predicting, as best you can, subsequent values of the sequence.
- The sequence might have non-zero entropy rate, so perfect prediction might be impossible.
- We will begin by focusing at some length on the following example:

$$\ldots 10111110101110111010111 \ldots$$

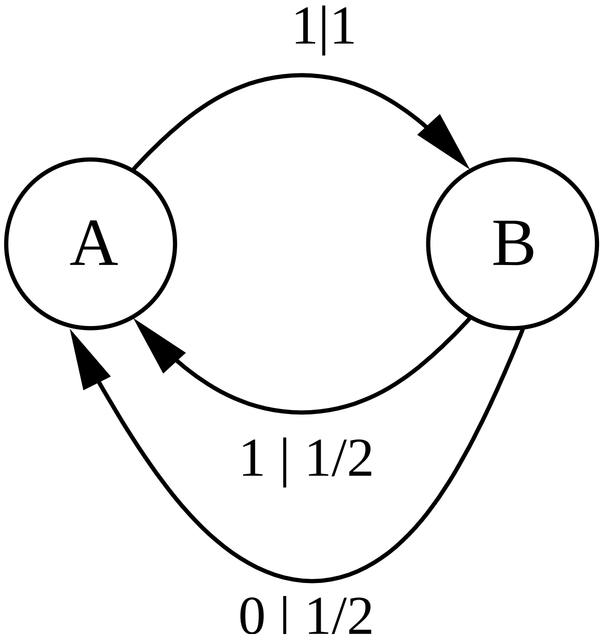
Discovery!

$$\ldots 10111110101110111010111 \ldots$$

- After some squinting, you will probably notice that every other symbol is 1. The other symbols are 0 or 1 with equal probability.
- You discovered a pattern: a regularity.
- Note that this pattern is stochastic.
- Note that you did not *recognize* the pattern.
- Recognition entails searching for a match to a pre-determined set patterns or templates.
- Discovery means finding something new: something not necessarily seen before.
- How can we represent this regularity mathematically, and can we program a computer to do pattern discovery?

Initial example, continued

- The machine that can reproduce this sequence is:

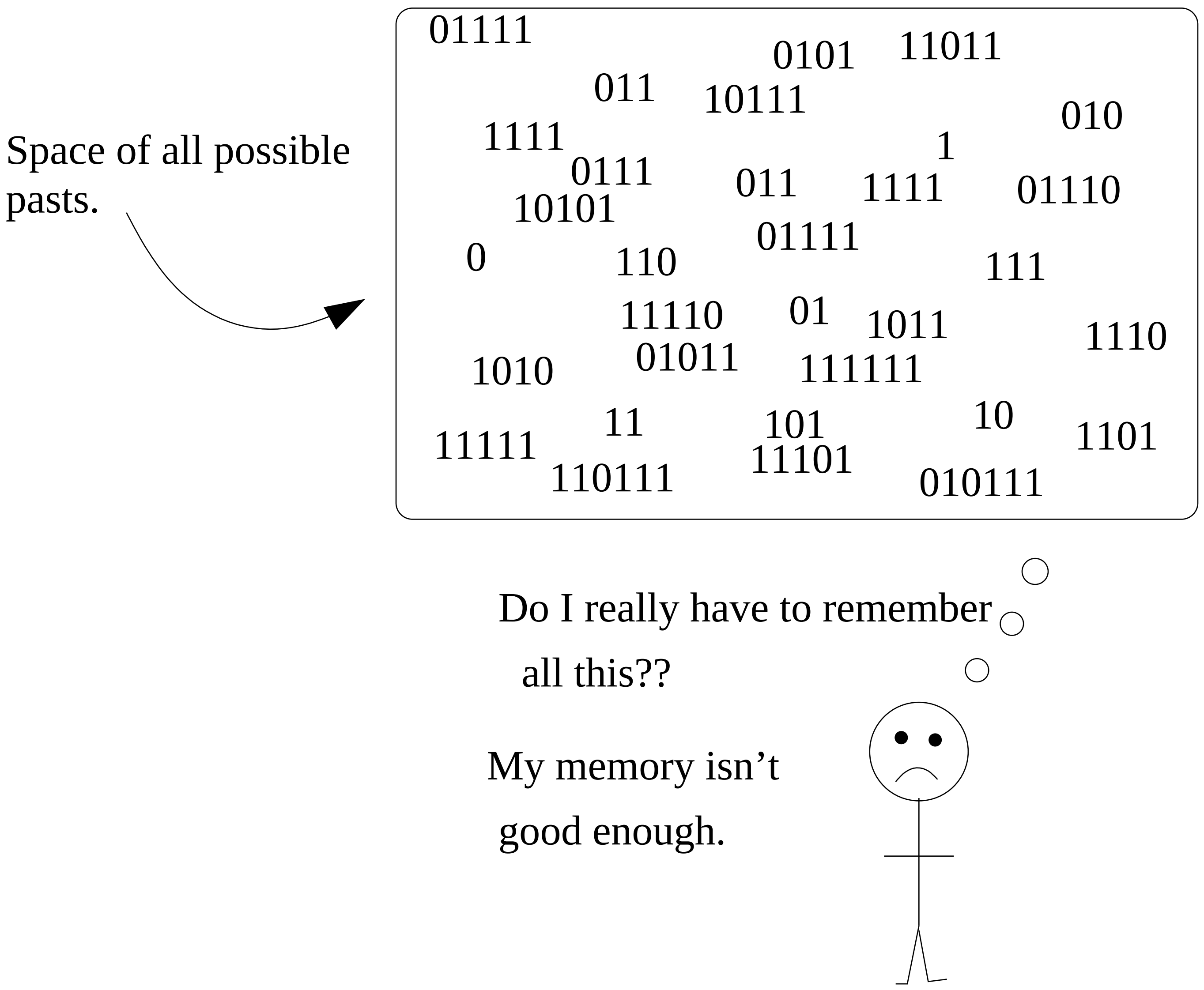


- From state $\mathbf{A}$, one sees a 1 with probability 1.
- From sate $\mathbf{B}$, one sees a 1 with probability $1/2$, and a 0 with probability $1/2$.
- This is a stochastic generalization of a finite state machine.
- Note that it is still *deterministic* in the sense that the output symbol (0 or 1) determines the next state ($\mathbf{A}$ or $\mathbf{B}$).

Initial Example: Why Two States?

- Why are only two states necessary? And what exactly do we mean by “state”?
- There are many particular observed sequences which give one equivalent information about the future sequences
- For example, if you see 1010, or 1110 or simply 0, in all cases you know with certainty that a 1 is next.
- The idea is that it only makes sense to distinguish between historical sequences that give rise to different predictive information.
- There will usually be many sequences that give the same predictive information. Group these sequences together into a **state**.
- These states are known as **causal states**, I will formalize this notion of state below.

What do you Need to Remember in Order to Predict?



Space of all possible pasts.

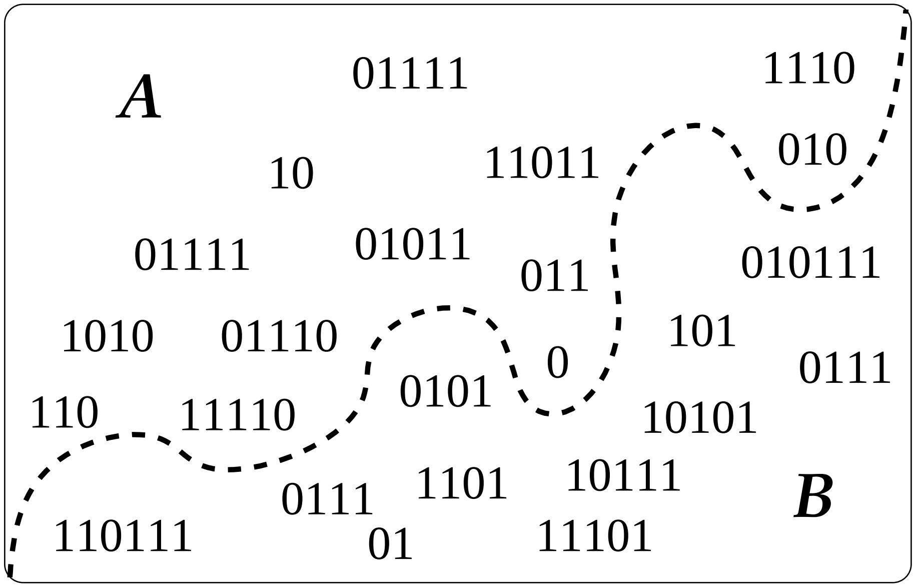
01111 0101 11011 011 10111 010 1111 1 0111 011 1111 01110 10101 01111 0 110 111 11110 01 1011 1110 01011 111111 1010 11 101 10 1101 11111 11101 110111 010111

Do I really have to remember all this??

My memory isn't good enough.

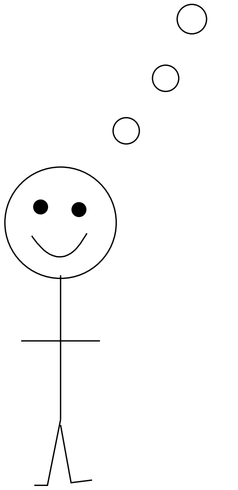
One Only Needs to Remember the Causal States.

Causal states partition the space of all past sequences



This is better!

I only need to remember the causal state, A or B.



Causal States

- The states $\mathbf{A}$ and $\mathbf{B}$ are known as causal states.
- Each causal state is a set containing (usually) very many past sequences $\overleftarrow{s}$.
- E.g., $\mathbf{A} = \{0, 10, 010, 110, 011, 1011, 1110, \ldots\}$.
- Denote the set of causal states by σ. I will index this set with Greek letters ($\alpha, \beta, \ldots$).
- Let ϕ be a function that takes a past history $\overleftarrow{s}$ as input and returns the causal state σ which is associated with it.
- E.g., $\phi(011) = \mathbf{B}$.
- Note: To keep this example simple, I've ignored the question of how an observer might come to know in which state the system is. I will return to this later.

How Might We Find Causal States?

- How much of the left half $\overleftarrow{S}$ is needed to predict the right half $\overrightarrow{S}$?
- Only need to distinguish between $\overleftarrow{S}$'s that give rise to different states of knowledge about $\overrightarrow{S}$.
- Two $\overleftarrow{S}$'s that give rise to the same state of knowledge are equivalent:

$$\overleftarrow{S}_i \sim \overleftarrow{S}_j \text{ iff } \Pr(\overrightarrow{S} \mid \overleftarrow{s}_i) \; = \; \Pr(\overrightarrow{S} \mid \overleftarrow{s}_j) \; .$$

- Equivalence classes induced by $\sim$ are **Causal States**, minimal sets of aggregate variables necessary for optimal prediction of $\overrightarrow{S}$.
- For example, $\Pr(\overrightarrow{S}\ |0) = \Pr(\overrightarrow{S}\ |1011)$. Hence, 0 and 1011 are equivalent under $\sim$.
- This means that the probability over the futures $\overrightarrow{S}$ is the same if you've seen 0 or 1011.

ϵ-Machines

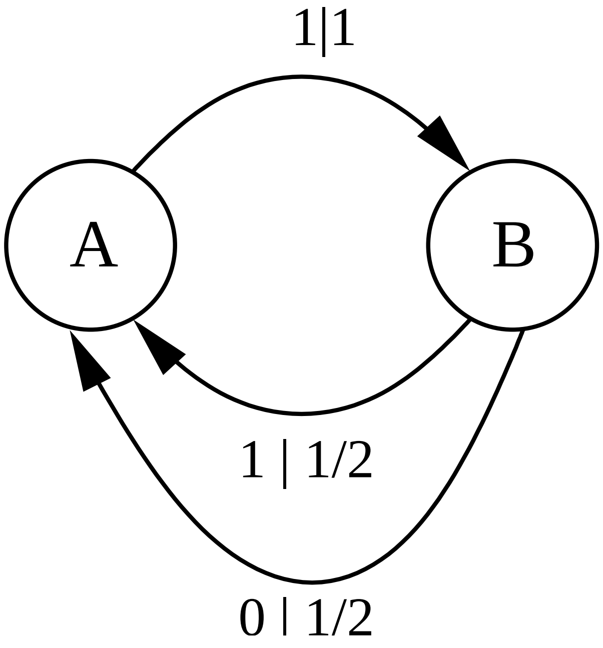
- The causal states together with the probability of transitions between causal states are an **ϵ-machine**, a minimal model capable of statistically reproducing the original configuration.
- The ϵ-machine tells us **how the system computes**.
- The "ϵ" reminds us that the measurement symbols upon which the machine is formed may be distorted via noise or the discretization process.
- Let $T^{(s)}_{\alpha\beta}$ denote the probability of being in causal state α, making a transition to causal state β and emitting the alphabet symbol s:

$$T^{(s)}_{\alpha\beta} = \Pr(\sigma_\beta, s|\sigma_\alpha) \ , \tag{10}$$

or

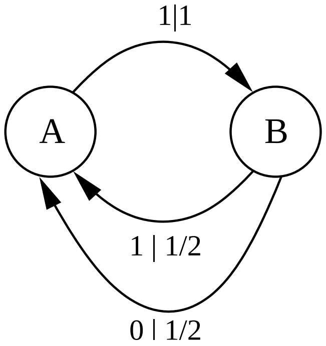
$$T^{(s)}_{\alpha\beta} = \alpha \xrightarrow{s} \beta \ . \tag{11}$$

Initial Example Again



$$T^{(s=0)} = \begin{pmatrix} 0 & 0 \\ \frac{1}{2} & 0 \end{pmatrix} , \quad T^{(s=1)} = \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & 0 \end{pmatrix} . \qquad (12)$$

Causal State Transitions



- Knowing the next signal uniquely determines the next causal state. Thus, the transition probability $T_{\alpha\beta}$ from causal state α to β is given by:

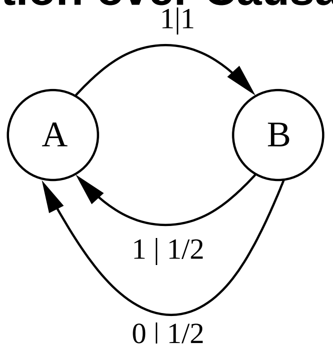
$$T_{\alpha\beta} = \sum_s T^{(s)}_{\alpha\beta} \ . \tag{13}$$

- For our example:

$$T = T^{(s=0)} + T^{(s=1)} \ , \tag{14}$$

$$T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \frac{1}{2} & 0 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & 0 \end{pmatrix} \ . \tag{15}$$

Distribution over Causal States



- Transitions between causal states are Markovian.
- Thus, the stationary (or asymptotic) distribution $p \equiv \Pr(\sigma)$ over the causal states is the left eigenvector of the transition matrix T:

$$pT = p \; . \tag{16}$$

- Normalize p so that $\sum_\alpha p_\alpha = 1$.
- For this example,

$$p = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} \; . \tag{17}$$

- I.e., the ϵ-machine spends an equal amount of time in states $\mathbf{A}$ and $\mathbf{B}$.

Statistical Complexity

- The *statistical complexity* is defined as the Shannon entropy of the asymptotic distribution of the causal states:

$$C_\mu \equiv -\sum_\alpha p_\alpha \log_2 p_\alpha \ . \tag{18}$$

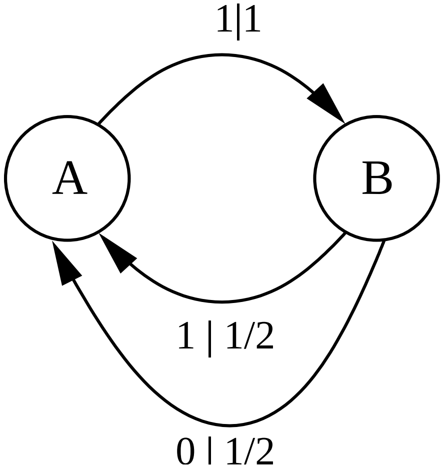
- To perform optimal prediction of the system one needs only to remember the causal states.
- The statistical complexity thus measures the minimum amount of memory needed to perform optimal prediction.
- The statistical complexity is a measure of the pattern or structure or regularity present in the system.
- For our example, $C_\mu = 1$.

Entropy Rate from an ϵ-machine

- The entropy rate h_μ can be easily calculated from an ϵ-machine.
- The entropy rate is just the entropy associated with the next transition at each causal state, weighted by the probability of being in that state:

$$h_\mu = -\sum_\alpha p_\alpha \sum_s \Pr(s|\sigma_\alpha) \log_2 \Pr(s|\sigma_\alpha) \ . \tag{19}$$

- For our example, $h_\mu = \frac{1}{2}$. The entropy from causal state $\mathbf{A}$ is 0 and from causal state $\mathbf{B}$ is 1.



Some Important Properties of ϵ-machines

- (For proofs, see Shalizi and Crutchfield. *J. Statistical Physics*. **104**:819. 2001.)
- The causal states are a *sufficient statistic*:

$$I[\overrightarrow{S};\overleftarrow{S}] = I[\overrightarrow{S};\sigma] \; . \tag{20}$$

  I.e., all the information about the future is contained in the causal states.

- The causal states are minimal.
- The causal states are unique up to trivial relabeling.
- The causal states form a Markov process.
- The ϵ-machine is a semi-group.

Statistical Complexity vs. Excess Entropy

- Both the statistical complexity C_μ and the excess entropy $\mathbf{E}$ are measures of complexity or structure or pattern or organization. However, they are not the same.
- $C_\mu =$ the minimal amount of memory needed to optimally predict the process.
- $\mathbf{E} =$ the amount of information the past carries about the future.

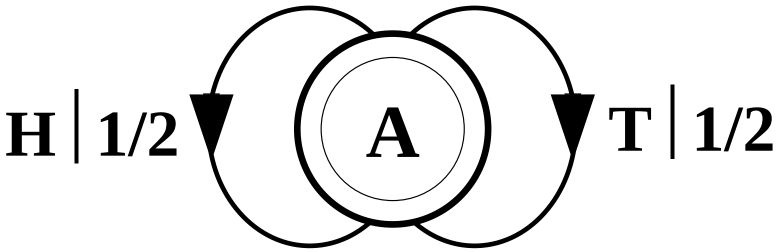
$$C_\mu \geq \mathbf{E} \,. \tag{21}$$

$$\text{Memory needed for model } \geq \text{ Memory of the process itself} \,. \tag{22}$$

- $\mathbf{E}$ is time reversal invariant; C_μ is not.

Example I

Fair Coin:

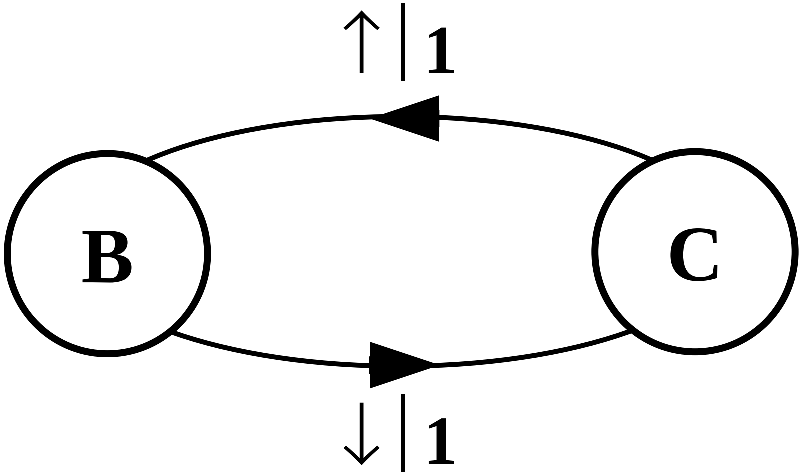


$$\cdots \mathrm{HHTHTHTTTHTHTHTTHTHH} \cdots$$

Entropy rate $h_\mu = 1$, Statistical Complexity $C_\mu = 0$.

Example II

Period 2 Pattern:

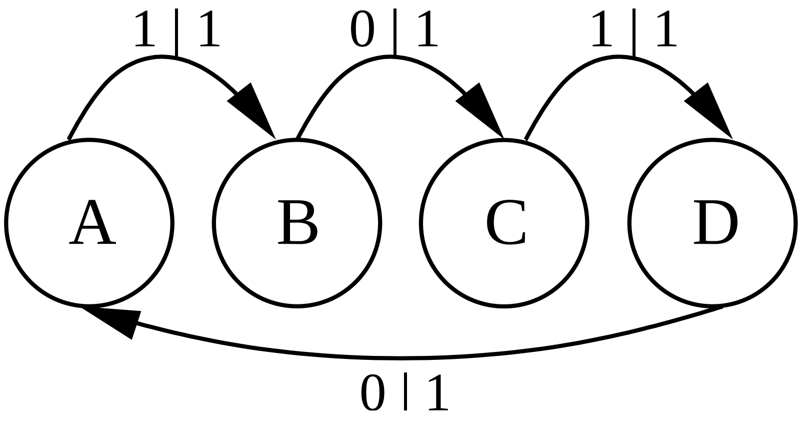


$$\cdots \uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow \cdots$$

Entropy rate $h_\mu = 0$, Statistical complexity $C_\mu = 1$.

A non-minimal example

Consider this machine for a period 2 sequence:



- States A and C are identical—they represent the same state of information about the future.
- So A and C should be merged to make one causal state.
- The same holds for B and D.
- The process of forming equivalence classes described on previous slides ensure that ϵ-machines are minimal.

Algorithms for Inferring ϵ-machines

There are two basic approaches

1. **Merge**
   - Initially distinguish between different histories. Then *merge* states that give rise to the same future distribution. I.e., merge states that are equivalent under $\sim$.
   - See Hanson, *PhD Thesis*, University of California, Berkeley, 1993.
2. **Split:**
   - Start with one state. This is equivalent to assuming a history of length zero. I.e., an IID process.
   - Add a symbol to history length. Split each state only if doing so increases predictability.
   - Repeat.

CSSR

- Shalizi and Shalizi(Klinkner) have implemented a state-splitting algorithm known as CSSR. (Causal State Splitting Algorithm)
- See Shalizi and Shalizi pp. 504–511 of Max Chickering and Joseph Halpern (eds.), *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference*, `http://arxiv.org/abs/cs.LG/0406011`.
- See also Shalizi, Shalizi, and Crutchfield. `http://arxiv.org/abs/cs.LG/0210025`. 2002.
- CSSR source code is available at `http://bactra.org/CSSR`.
- CSSR has been applied to: crystallography, geomagnetic fluctuations, natural languages, anomaly detection, natural languages, and more.

CSSR Details

- The run-time complexity of CSSR is: $\mathcal{O}(N) + \mathcal{O}(k^{2L+1})$.
- N is the number of symbols in your data string.
- L is the block-length you want to gather statistics over.
- k is the size of the alphabet.
- The $\mathcal{O}(N)$ term is associated with making a single pass through the data and filling a parse tree.
- The $\mathcal{O}(k^{2L+1})$ is associated with doing a large number of comparisons to see if states should be split.
- This is an almost-never-attained worst case scenario. It would only happen if essentially every history becomes its own state.
- For a given N, how large an L can you choose? A rough guide: $L \approx \log(N)/h_\mu$. (Marton and Shields, *Annals of Probability* **23**:960. 1994.) Be careful to not choose too large an L!

Computational Mechanics References and Applications

Almost all of the papers below can be found online either on arXiv.org or with a little bit of searching.

Early papers, foundations, reviews:

- Crutchfield and Young, *Phys. Rev. Lett*, 63:105-108, 1989
- Crutchfield and Young, in *Complexity, Entropy and the Physics of Information*, Addison-Wesley, 1990. [Detailed analysis of Logistic and Tent maps]
- Crutchfield, *Physica D*, 75:11-54, 1994. [Long article, good review section, many different examples. A good place to start.]
- Shalizi and Crutchfield. *J. Statistical Physics*. **104**:819. 2001. [Mathematical foundations of causal states. Careful proofs of optimality and minimality.]

Computational Mechanics Extensions: Optimal Causal Filtering

- What if you impose a constraint on your model size, possibly limiting the number of causal states you use?
- The result will be a model that is sub-optimal?
- But how sub-optimal? What states achieve the best possible (sub-optimal) prediction? And how can these states be found?
- These questions, and more, are answered in the following references:
  - Still and Crutchfield, “Structure or Noise?” `arXiv:0708.0654v1`.
  - Still, Crutchfield, Ellison, “Optimal Causal Inference.” `arXiv:0708.1580v1`.

Applications and Extensions of Causal States

- Hanson, *PhD Thesis*, University of California, Berkeley, 1993. [Cellular Automata]
- Hanson and Crutchfield, *Physica D*, 103:169-189, 1997. [Cellular Automata]
- Upper, *PhD Thesis*, University of California, Berkeley, 1997. [Hidden Markov Models]
- Delgado and Solé, *Phys. Rev. E*, 55:2338-2344, 1997. [Coupled Map Lattices]
- Witt, Neiman and Kurths, *Phys. Rev. E*, 55:5050-5059, 1997. [Stochastic resonance]
- Goncavales, et. al., *Physica A*, 257, 385-389. 1998. [Dripping faucets]
- Feldman and Crutchfield, SFI:98-04-026, 1998. [One-dimensional Ising models. Includes lengthy review, calculations of excess entropy, and comparisons to statistical mechanical quantities.]
- Varn, et al. *Physical Review B*. **66**:156. 2002. [Layered Solids]
- Clarke, et al. *Physical Review E*. **67**:016203. 2003 [Geomagnetism]
- Palmer, et al. *Advances in complex systems*. 1:1-16. 2001. [Climate modeling, ϵ-machines inferred from empirical data.]
- Shalizi, Discrete Mathematics and Theoretical Computer Science, AB(DMCS) (2003): 11-30. [Dynamical systems on random networks]

Applications and Extensions of Causal States, Continued

- Görnerup and Crutchfield. SFI 04-06-020. [Self-assembling evolutionary systems]
- Ray. *Signal Processing*. **84**:1114. 2004.
- Shalizi, et al. *Physical Review Letters*. **93**:118701. 2004. [Cellular automata in more than one dimension]
- Padro and Padro, in *Proceedings of the Fifth International Workshop on Finite-State Methods and Natural Language Processing.* 2005.
- Young, et al. *Physical Review Letters.* **94**:098701. 2005. [Two-dimensional brain slices. Applications to Alzheimer's disease.]
- Park, et al. *Physica A*. **379**:179. 2007. [Financial time series. Stock market.]
- Klinkner, et al. `arXiv:q-bio/0506009v2`. [Shared information in neural networks.]
- Shalizi, et al. *Phys. Rev.E*. **73**: 036104. 2006. [2D cellular automata. Automatic order-parameter finding!]

Computational Mechanics Conclusions:

Questions:

- What are patterns and how can we discover them?
- What does it mean to say a system is organized?

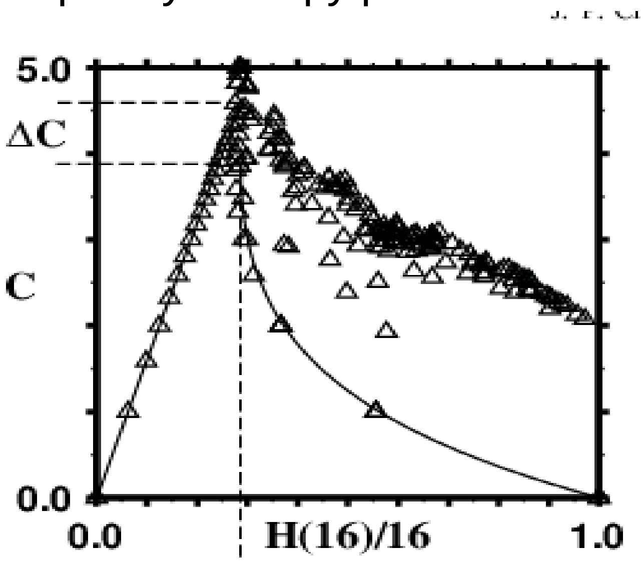
Summary:

- Computation theory classifies sets of sequences by considering how difficult it is to recognize them.
- Causal states and ϵ-machines adapt computation theory for use in a probabilistic setting.
- The ϵ-machine provides an answer to the question: What patterns are present in a system?
- The ϵ-machine can be inferred directly from observed data.
- The ϵ-machine reconstruction pattern can discover patterns—even patterns that we haven't seen before.

An Example and Some Thoughts on Emergence

The results and figures on next few slides are from Crutchfield and Young, in *Complexity, Entropy and the Physics of Information*, Addison-Wesley, 1990.

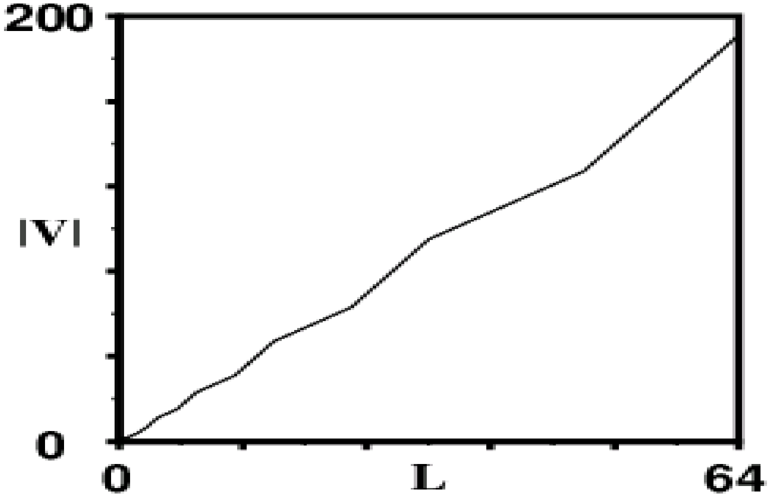
- Consider the symbolic dynamical system generated by the logistic equation $f(x) = rx(1-x)$.
- Figure shows 193 complexity-entropy pairs for different r values for the logistic map.



- The linear region on the left corresponds to periodic behavior.
- To the right of the linear region, the behavior is chaotic.

Logistic Equation: Critical Machine

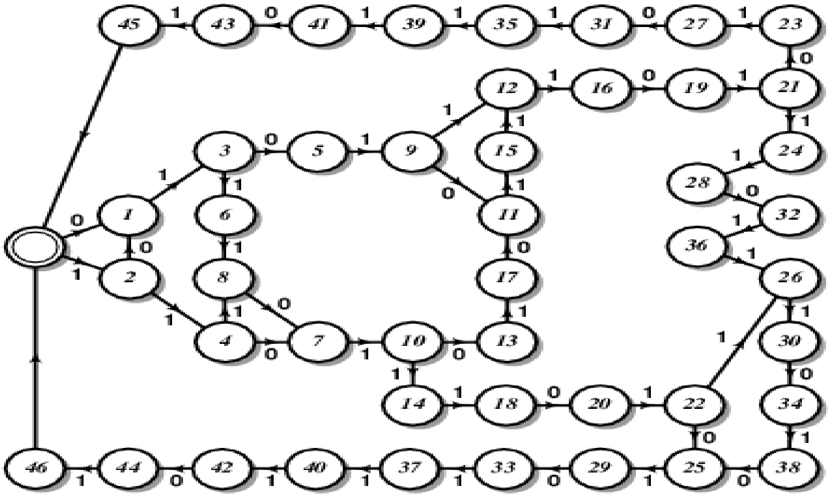
- What happens as the periods get larger and we approach the phase transition?
- At the period-doubling accumulation point the number of states $\mathbf{V}$ in the ϵ-machine diverges.



- This suggests that there is no longer a finite representation at the lowest level of the Chomsky Hierarchy.
- The nature of the divergence leads one to a higher-level computational model for the system.

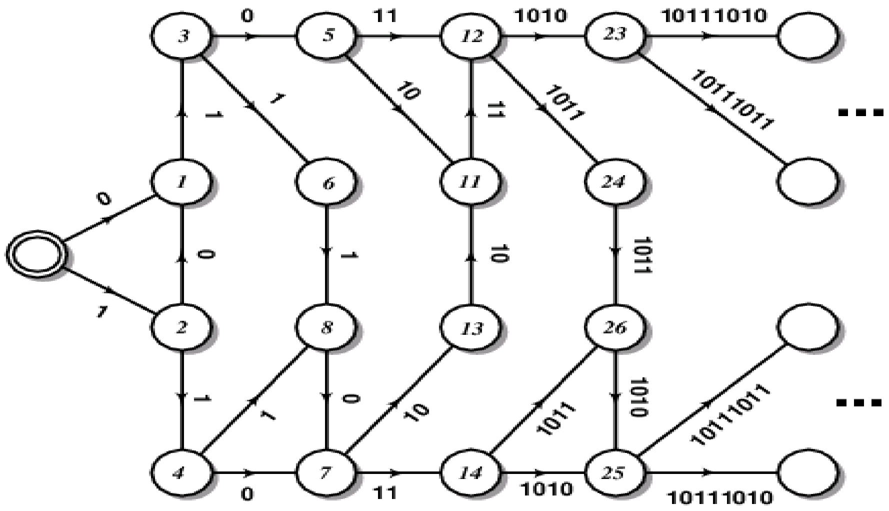
Critical Machine, continued

- ϵ-machine estimated with a window size of $L = 16$:



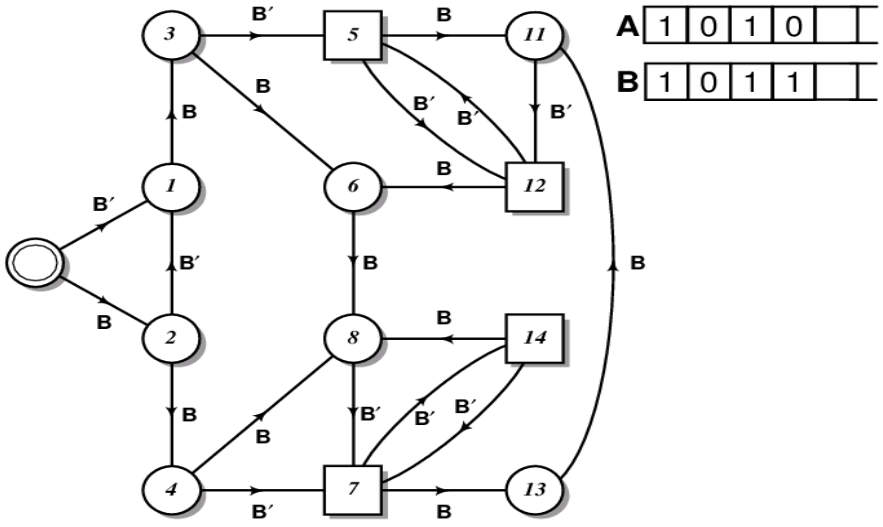
Critical Machine, continued

- Deterministic chains replaced by equivalent strings:



- Looking at the regularities in the machine leads one to a finite representation at a higher computational level.

Critical Machine: Finite Representation



- Two string registers, A and B. Start with A containing 0 and B containing 1.
- Squares are new type of state. When making a transition from a square, update string registers: $A \to BB$, and $B \to BA$.
- B' is B with the last bit flipped. (E.g., if $B = 1010$, $B' = 1011$.)
- **Main Point:** Diverging model size at lower level of Chomsky hierarchy necessitated a new model at a higher level (context sensitive languages).
- Innovation in response to emergence.

Emergence?

- Computational Mechanics and related approaches let us speak about complexity or organization in a precise way.
- So, in a given situation we can tell if complexity increases.
- One might view emergence to be a large, qualitative increase in complexity. Something genuinely new emerges, as in the above example.
- It is sometimes said that emergent phenomena are those which cannot be predicted or calculated from a lower-level description.
- But predicted by whom? Why should emergence depend on my calculational skills?

Emergence?

- Another notion, known sometimes as *strong emergence* suggests that for large systems there are new fundamental laws that come into play.
- I don't think there are many scientists who believe this.
- I have not yet found or devised a definition of emergence that I find compelling.
- But I still find the notion of emergence to be compelling.
- I don't think this invalidates the notion of emergence. But it should be treated with care.

Of Exactitude in Science

...In that Empire, the craft of Cartography attained such Perfection that the Map of a Single province covered the space of an entire City, and the Map of the Empire itself an entire Province. In the course of Time, these Extensive maps were found somehow wanting, and so the College of Cartographers evolved a Map of the Empire that was of the same Scale as the Empire and that coincided with it point for point. Less attentive to the Study of Cartography, succeeding Generations came to judge a map of such Magnitude cumbersome, and, not without Irreverence, they abandoned it to the Rigours of sun and Rain. In the western Deserts, tattered Fragments of the Map are still to be found, Sheltering an occasional Beast or beggar; in the whole Nation, no other relic is left of the Discipline of Geography.

From Travels of Praiseworthy Men (1658) by J. A. Suarez Miranda

The piece was written by Jorge Luis Borges and Adolfo Bioy Casares. English translation quoted from J. L. Borges, A Universal History of Infamy, Penguin Books, London, 1975. `http://www.kyb.tuebingen.mpg.de/bu/people/bs/borges.html`

Some thoughts on Reduction

- Reduction is sometimes seen as the opposite of the study of emergence.
- I don't believe this. In a sense, all science is reductive.
- What alternative is there? We can't study the whole world at the same time or use a map that is full size.
- What is important, I think, is to not pretend one isn't being reductive.
- I think that reduction is fine, but reductionism isn't.
- Similarly, I think that studying emergence is great, but I am a little suspicious of holism.
- In general, I think that reductionism vs. holism is a false dichotomy. These approaches need not be in opposition to each other.
- Perhaps Complex Systems is a synthesis of reductionism and holism.

Part X

A Critical Survey of Measures of Complexity

Some Thoughts on Complexity Measures

The term *complexity* has many different meanings. At least one adjective is needed to help distinguish between different uses of the word:

- Kolmogorov-Chaitin Complexity
- Computational Complexity
- Stochastic Complexity
- Statistical Complexity
- Structural Complexity
-
-

Note: Some portions of the first half of this presentation were prepared jointly with Jim Crutchfield.

Deterministic Complexity

The *Kolmogorov-Chaitin* complexity $K(x)$ of an object x is the length, in bits, of the smallest program (in bits) that when run on a *Universal Turing Machine* outputs x and then halts.

References:

- Kolmogorov, *Problems of Information Transmission*, 1:4-7. (1965)
- Kolmogorov, *IEEE Trans. Inform. Theory*, IT-14:662-664. (1968).
- Solomonoff. *Inform. Contr.*, 7:1-22, 224-254. (1964).
- Chaitin, *J. Assoc. Comp. Mach.*, 13:547-569. (1966).
- Martin-Löf, *Inform. Contr.*, 9:602-619. (1966).
- **Books:**
  - Chpt. 7 of: Cover and Thomas, “Elements of Information Theory,” Wiley, 1991.
  - Chaitin, “Information, Randomness and Incompleteness,” World Scientific, 1987.

Kolmogorov Complexity $\approx$ Randomness

- The Kolmogorov complexity $K(x)$ is maximized for random strings, since it requires a deterministic accounting of all symbols in the string.
- The average growth rate of $K(x)$ is equal to the entropy rate h_μ.
- If x = trajectory of a chaotic dynamical system f:

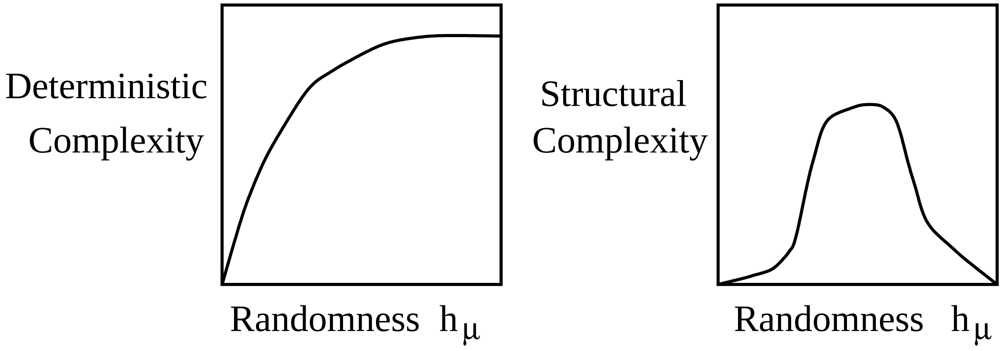
  $$K(x(t)) \;=\; h_\mu(f) \quad \text{for typical } x(0)\ .$$

  (Brudno, *Trans. Moscow Math. Soc.*, 44:127. (1983).)
- If a string x is random, then it possesses no regularities. Thus,

  $$K(x) \;=\; |\text{Print}(x)|\ .$$
- That is, the shortest program to get a UTM to produce x is to just hand the computer a copy of x and say “print this.”
- The Kolmogorov complexity provides a powerful lens with which to consider what it means to be random.

Measures of “Complexity” that Capture a Property Distinct from Randomness

- The entropy rate h_μ and the Kolmogorov Complexity $K(x)$ do not measure pattern or structure or correlation or organization.
- h_μ and $K(x)$ are maximized for random strings.



- Structural complexity or statistical complexity measures are not maximized by random strings.
- The excess entropy $\mathbf{E}$ and the statistical complexity C_μ are examples of structural complexity measures.
- The following slides review a few other statistical complexity measures.

Other Approaches to Structural Complexity

Logical Depth

- The **Logical Depth** of x is the **run time** of the shortest program that will cause a UTM to produce x and then halt.
- Logical depth is not a measure of randomness; it is small for both trivially ordered and random strings.
- Recall that if a string x is random, then it possesses no regularities. Thus,

  $$K(x) = |\mathrm{Print}(x)| \ .$$

  is the shortest UTM program that will reproduce it.
- This is a long program, because x is long. But presumably it would take very little compute time to run.
- References: Bennett, *Found. Phys.*, 16:585-592, 1986. Bennett, in *Complexity, Entropy and the Physics of Information*, Addison-Wesley, 1990.

Other Approaches to Statistical Complexity, Continued

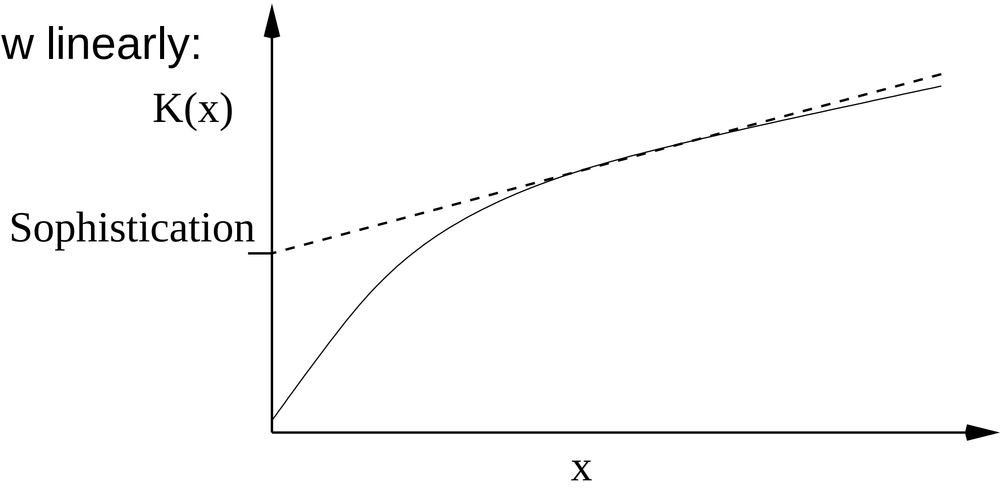
Thermodynamic Depth

- Thermodynamic depth of an object is the total entropy generated in the production of that object. (Lloyd and Pagels, *Annals of Physics*, 188:186-213).
- However, Shalizi and Crutchfield (*Phys. Rev. E*. **59**:275. 1999), show that:
  - Thermodynamic depth depends crucially on the notion of state.
  - Lloyd and Pagels give no general prescription for how states should be chosen.
  - Once states are chosen, thermodynamic depth is equivalent to the reverse time entropy rate.

Other Approaches to Structural Complexity, continued

Sophistication:

- Koppel, *Complex Systems*, 1:1087-91, 1987.
- The Kolmogorov complexity $K(x)$ of an object will grow linearly in x. The sophistication is essentially the size of the model—the part of $K(x)$ that doesn't grow linearly:



- The linear growth of $K(x)$ is due to the length of the string.
- The constant part $K(x)$ describes the regularities of the string.
- Roughly speaking, the sophistication can be thought of as the Kolmogorov version of the excess entropy.

Other Approaches to Structural Complexity, continued

Non-Linear Modeling

- Wallace and Boulton, 1968.
- Crutchfield and McNamara, *Complex Systems* 1: 417-452, 1987.
- Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, 1989.
- Crutchfield and Young, in *Complexity, Entropy and the Physics of Information*, Addison-Wesley, 1990.

Model Convergence and Hierarchical Grammatical Complexities

- Badii and Politi, *Complexity: Hierarchical Structures and Scaling in Physics*, Cambridge, 1997.
- Badii and Politi, *Phys. Rev. Lett.*, 78:444-447, 1997.

Note: Badii and Politi's book contains a solid discussion of many different structural complexity measures.

Early Uses of Mutual Information

Probably only of historical interest. (?)

- Rothstein, in *The Maximum Entropy Formalism*, MIT Press, 1979.
- Chaitin, in *Information, Randomness, and Incompleteness*, World Scientific, 1987.
- Gatlin, *Information Theory and the Living System*, Columbia University Press. 1972.
- Watanabe, *Knowing and Guessing: A Quantitative Study of Inference and Information*, Wiley, 1969.

Other Approaches to Structural Complexity, continued

Miscellaneous References:

- Kolmogorov, *Russ. Math. Surveys*, 38:29, 1983.
- Wolfram, *Comm. Math. Phys.*, 96:15-57, 1984.
- Wolfram, *Physica D*, 10:1-35, 1984. [Very influential paper classifying the complexity of cellular automata. In my view not a rigorous or compelling classification, but interesting nevertheless.]
- Hubermann and Hogg, *Physica D*, 22:376-384, 1986.
- Bachas and Hubermann, *Phys. Rev. Lett.*, 57:1965, 1986
- Peliti and Vulpiani, eds., *Measures of Complexity*, Springer-Verlag, 1988.
- Wackerbauer, et. al., *Chaos, Solitons & Fractals*, 4:133-173, 1994. [Very interesting review article and overview of different approaches to complexity.]
- Atmanspacher, et al., *Physica A*, 243:819. 1997.
- Bar-Yam, *Dynamics of Complex Systems*, Addison-Wesley, 1997.

Other Approaches to Structural Complexity, continued

More Miscellaneous References:

- Gowin et al., *Acta Astronautica*, **49**:3, 2001. [Complexity measures applied to bone architecture.]
- Gramms, *Phys. Rev. E*, **50**:2616, 1994. [Entropy of critical circle maps.]
- Ebeling, *Physica D*, **109**:42, 1997. [Nice review of entropy and entropy convergence of a number of systems, including large texts.]
- Freund, *Phys. Rev. E*, **53**:5793, 1996. [Scaling of entropies for intermittent processes.]
- Li, *Complex Systems*, **5**:381, 1991.
- Newth and Finnigan, *Aust. J. Chem.*, **59**:841. [Review article about emergence and self-organization. Somewhat dismissive of computational mechanics.]
- Boffetta, et al., *Physics Reports*, **356**:367 2002. [Major review article about a predictability approach to dynamical systems and complexity.]

This is not a comprehensive list.

Non-constructive Complexity: The Road Untakable

All Universal Turing Based complexity measures suffer from several drawbacks:

1. They are uncomputable.
2. By adopting a UTM, the most powerful discrete computation model, one loses the ability to distinguish between systems that can be described by computational models less powerful than a UTM.

UTM-based “complexity” measures include:

- **Logical Depth:** Bennett, *Found. Phys.*, 16:585-592, 1986.
- **Sophistication:** Koppel, *Complex Systems*, 1:1087-91, 1987.
- **Effective Complexity:** Gell-Mann and Lloyd, *Complexity*, 2:44-52, 1996.

On the other hand, (some) UTM-based arguments may be useful for providing a clear framework for expressing notions of complexity.

Complexity = Order × Disorder?

- There are a number of complexity measures of the form:

$$\text{Complexity} = \text{Order} \times \text{Disorder}$$

- Disorder is usually some form of entropy.
- Sometimes “order” is simply $(1 - h_\mu)$.
- Often, “order” is taken to be some measure of “distance from equilibrium,” where equilibrium and equiprobability are sometimes considered to be synonymous.

In my view these sorts of complexity measures have some serious shortcomings:

- Lack a clear interpretation and direct accounting of structure.
- Unclear that distance from equilibrium is equivalent to order.
- Assign a value of zero complexity to all systems with vanishing entropy.

Complexity = Order × Disorder?, continued

But, you can read the papers and decide for yourself. See, e.g.,

- Shiner, et al. *Phys. Rev. E*, 59:1459. 1999.
- Lopez-Ruiz, et al., *Phys. Lett. A*, 209:321. 1995.
- Piasecki, et al., *Physica A*, 307:157. 2002.

For some critiques, see:

- Feldman and Crutchfield, *Phys. Lett. A*, 238:244. 1998.
- Crutchfield, et al., *Phys. Rev. E*, 62:2996. 2000.
- Binder. *Phys. Rev. E*, 62:2998. 2000.

Part XI

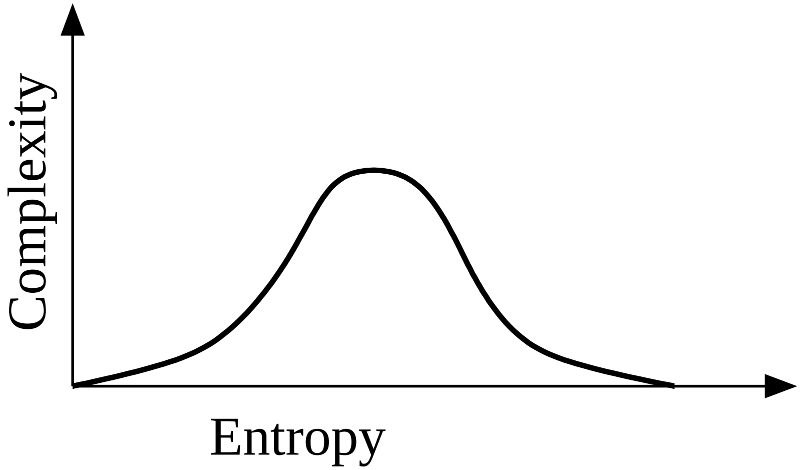
Complexity vs. Entropy and Edges of Chaos

Complexity vs. Entropy

- What is the relationship between complexity and entropy?
- Are they completely unrelated? Is complexity the opposite of entropy?
- Is complexity an *absence* of unpredictability, or the *presence* of something else?

One approach: Prescribing Complexity vs. Entropy Behavior

- Zero Entropy ⟶ Predictable ⟶ simple and not complex.
- Maximum Entropy ⟶ Perfectly Unpredictable ⟶ simple and not complex.
- Complex phenomena combine order and disorder.
- Thus, it must be that complexity is related to entropy as shown:

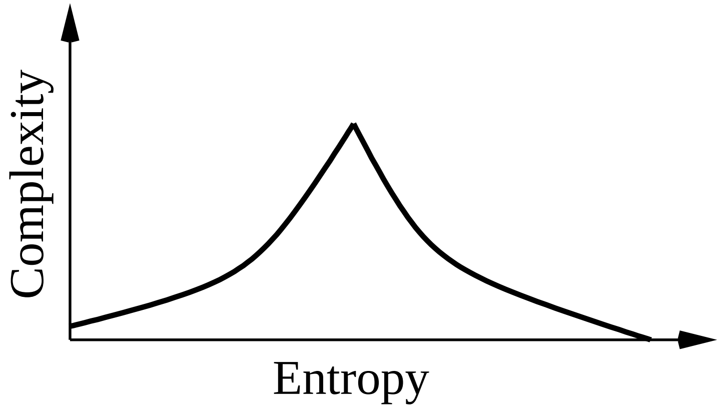


- This plot is often used as the central criteria for defining complexity.

Complexity-Entropy Phase Transition?

Edge of Chaos?

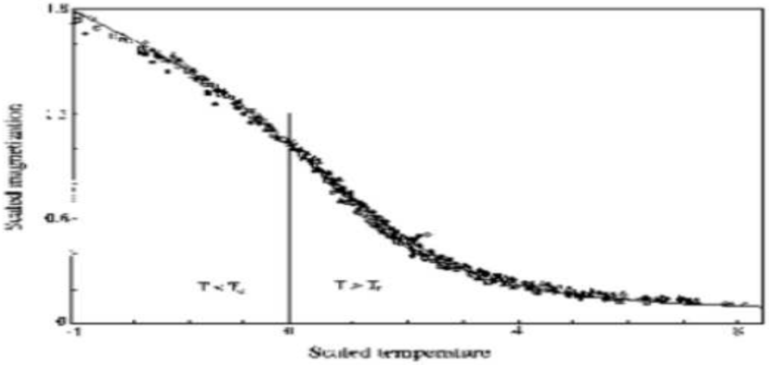
- Additionally, it has been conjectured that there is a sharp transition in complexity as a function of entropy:



- Perhaps this complexity-entropy curve is *universal*—it is the same for a broad class of apparently different systems.
- Part of the motivation for this is the remarkable success of universality in critical phenomena and condensed matter physics.

Data Collapse

- Scaled magnetization vs. scaled temperature for five different magnetic materials: EuO, Ni, YIG, $CrBr_3$, and Pd_3Fe.

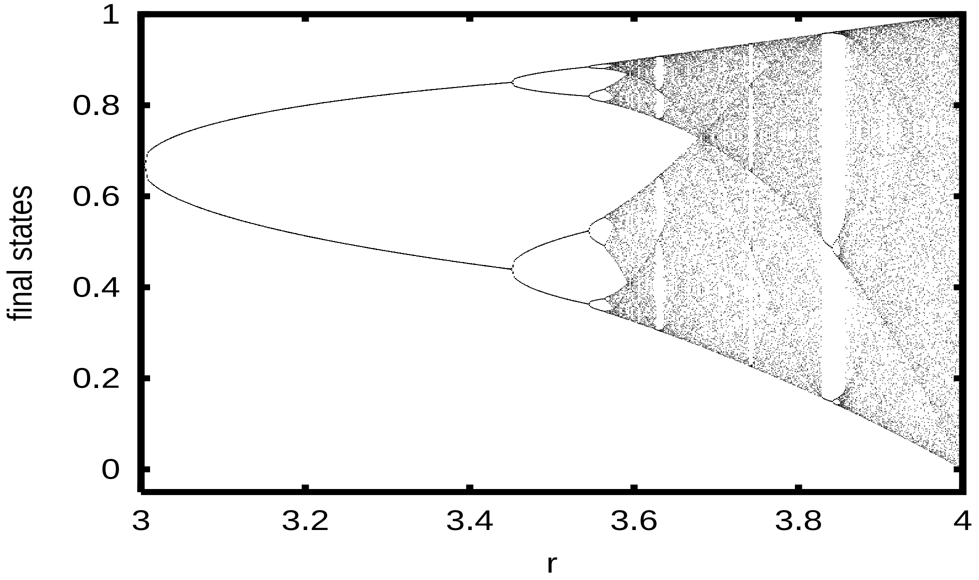


- These materials are very different, but clearly possess some deep similarities.
- Figure source: H.E. Stanley, *Rev. Mod. Phys*. **71**:S358. 1999.
- Perhaps there is a similar data collapse for some appropriate definitions of complexity and entropy.
- Note: One could trivially obtain this by simply defining complexity to be a single-valued function of the entropy.

Complexity vs. Entropy: A Different Approach
Define Complexity on its own Terms

- Do not prescribe a particular complexity-entropy behavior.
- To be useful, a complexity measure must have a clear interpretation that accounts in a direct way for the correlations and organization in a system.
- Consider a well known complexity measures: excess entropy
- Calculate complexity and entropy for a range of model systems.
- Plot complexity vs. entropy. This will directly reveal how complexity is related to entropy.
- Is there a universal complexity-entropy curve?

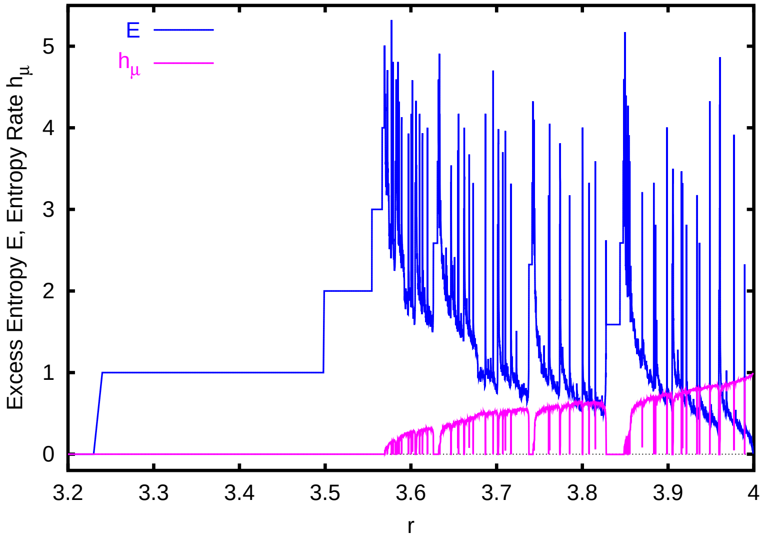
Logistic Equation: Bifurcation Diagram



- For a given r (horizontal axis), the “final states” are shown.
- Chaotic behavior appears as a solid vertical line.
- Examples:
  - $r = 3.2$: Period 2.
  - $r = 3.5$: Period 5.
  - $r = 3.7$: Chaotic.

Complexity vs. Entropy: Logistic Equation

Plot the excess entropy $\mathbf{E}$ and the entropy rate h_μ for the logistic equation as a function of the parameter r.



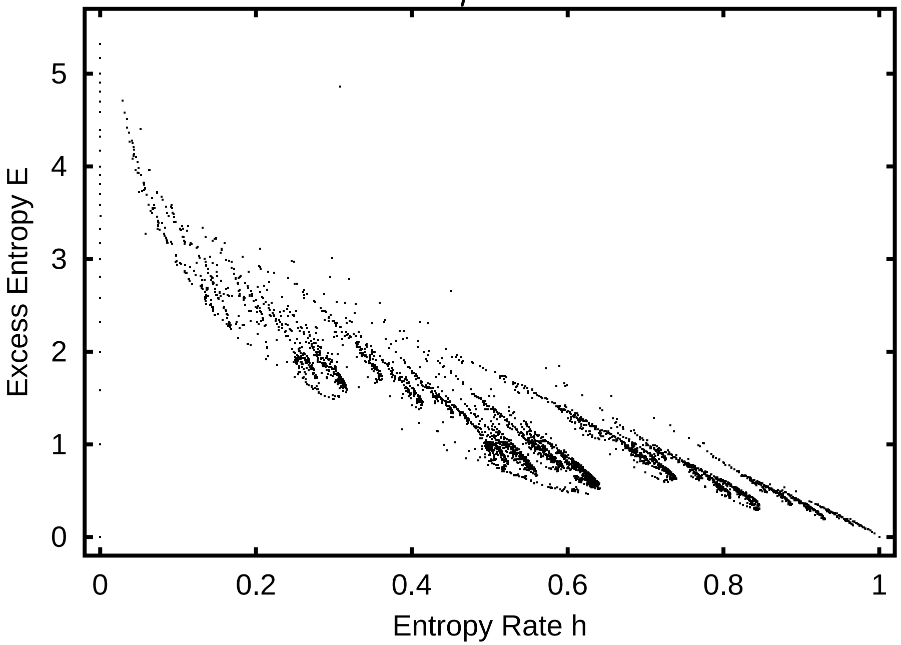
- Note that $\mathbf{E}$ and h_μ depend on a complicated way on r.
- Hard to see how complexity and entropy are related.
- Numerical results. For each r, 1×10^7 symbols were generated. The largest L was 30 for low entropy sequences. r was varied by increments of 0.0001.

Complexity-Entropy Diagrams

- Plot complexity vs. entropy. This will directly reveal how complexity is related to entropy.
- This is similar to the idea behind phase portraits in differential equations: plot two variables against each other instead of as a function of time. This shows how the two variables are related.
- It provides a parameter-free way to look at the intrinsic information processing of a system.
- Complexity-entropy plots allow comparisons across a broad class of systems.

Complexity-Entropy Diagram for Logistic Equation

- Excess entropy $\mathbf{E}$ vs. entropy rate h_μ from two slides ago.



- Structure is apparent in this plot that isn't visible in the previous one.
- Not all complexity-entropy values can occur; there is a forbidden region.
- Maximum complexity occurs at zero entropy.
- Note the self-similar structure. This isn't surprising, since the bifurcation diagram is self-similar.

Ising Models

Consider a one- or two-dimensional Ising system with nearest and next nearest neighbor interactions:

- This system is a one- or two-dimensional lattice of variables $s_i \in \{\pm 1\}$.
- The energy of a configuration is given by:

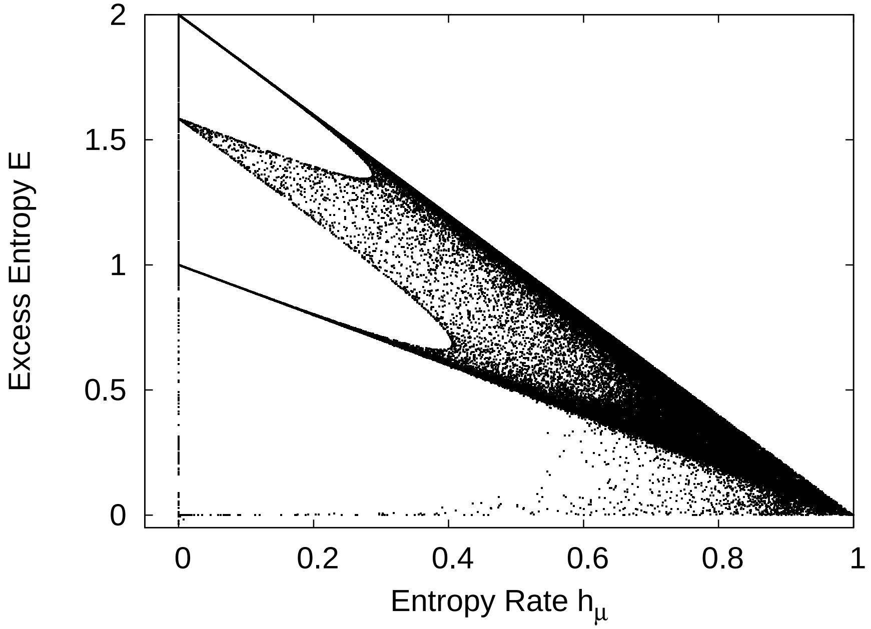
$$\mathcal{H} \equiv -J_1 \sum_i s_i s_{i+1} - J_2 \sum_i s_i s_{i+2} - B \sum s_i \; .$$

- The probability of observing a configuration $\mathcal{C}$ is given by the Boltzmann distribution:

$$\Pr(\mathcal{C}) \propto e^{-\frac{1}{T}\mathcal{H}(\mathcal{C})} \; .$$

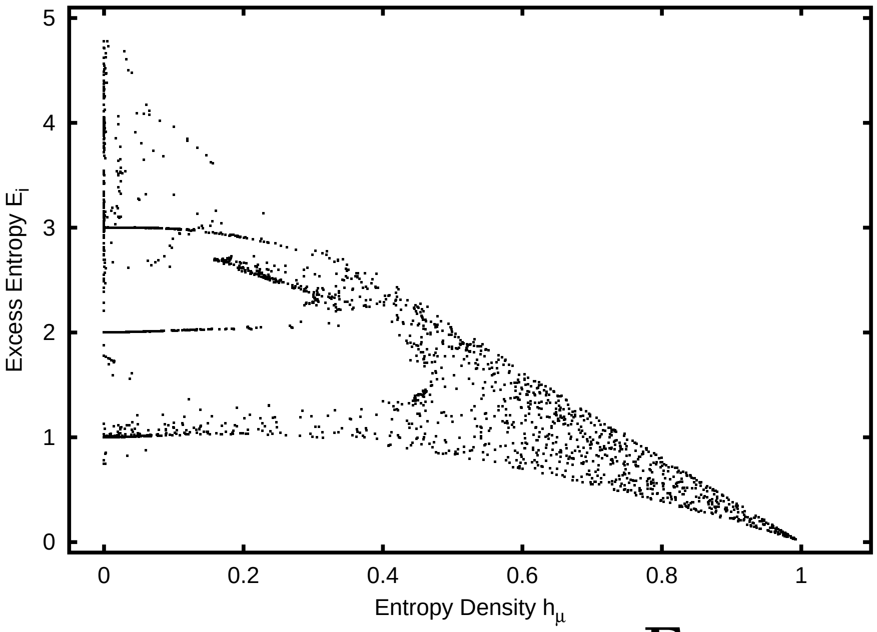
- Ising models are very generic models of spatially extended, discrete degrees of freedom that have some interaction that makes them want to either do the same or the opposite thing.

Complexity-Entropy Diagram for 1D Ising Models



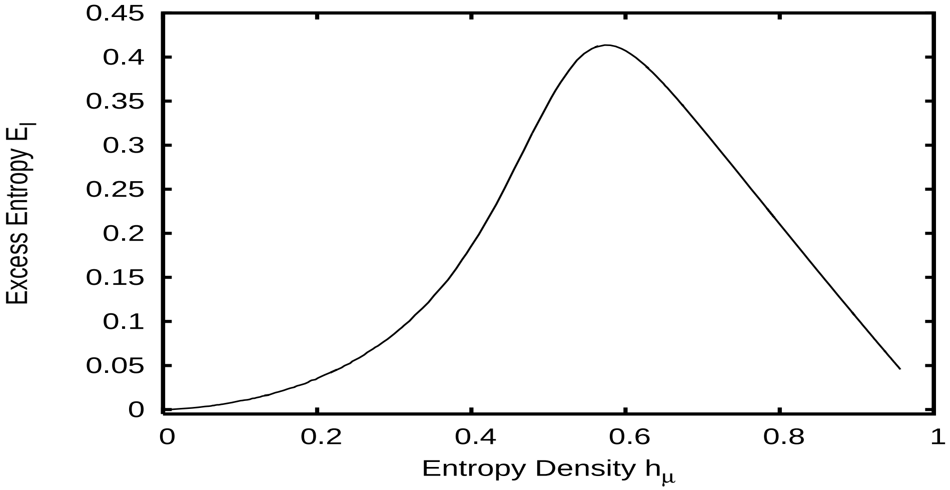
- Excess entropy $\mathbf{E}$ vs. entropy rate h_μ for the one-dimensional Ising model with anti-ferromagnetic couplings.
- Model parameters are chosen uniformly from the following ranges: $J_1 \in [-8, 0]$, $J_2 \in [-8, 0]$, $T \in [0.05, 6.05]$, and $B \in [0, 3]$.
- Note how different this is from the logistic equation.
- These are exact transfer-matrix results.

Complexity-Entropy Diagram for 2D Ising Models



- Mutual information form of the excess entropy $\mathbf{E}_\mathrm{i}$ vs. entropy density h_μ for the two-dimensional Ising model with AFM couplings
- Model parameters are chosen uniformly from the following ranges: $J_1 \in [-3, 0]$, $J_2 \in [-3, 0]$, $T \in [0.05, 4.05]$, and $B = 0$.
- Surprisingly similar to the one-dimensional Ising model.
- Results via Monte Carlo simulation of 100x100 lattices.

Complexity-Entropy Diagram for 2D Ising Model Phase Transition



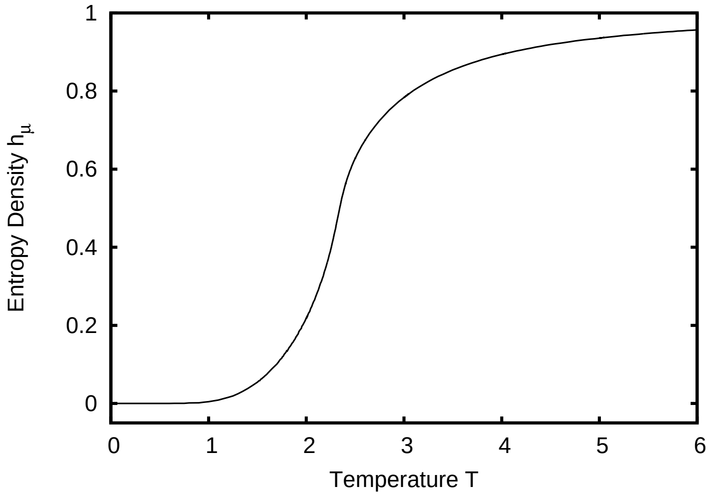
- Convergence form of the excess entropy $\mathbf{E}_{\rm c}$ vs. entropy density h_μ for the two-dimensional Ising model with NN couplings and no external field.

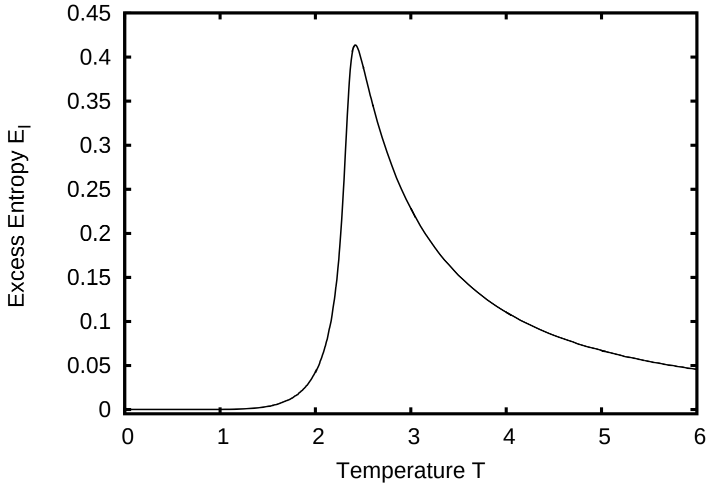
- Model undergoes phase transition as T is varied at $T \approx 2.269$.

- There is a peak in the excess entropy, but it is somewhat broad.

- Results via Monte Carlo simulation of 100x100 lattice.

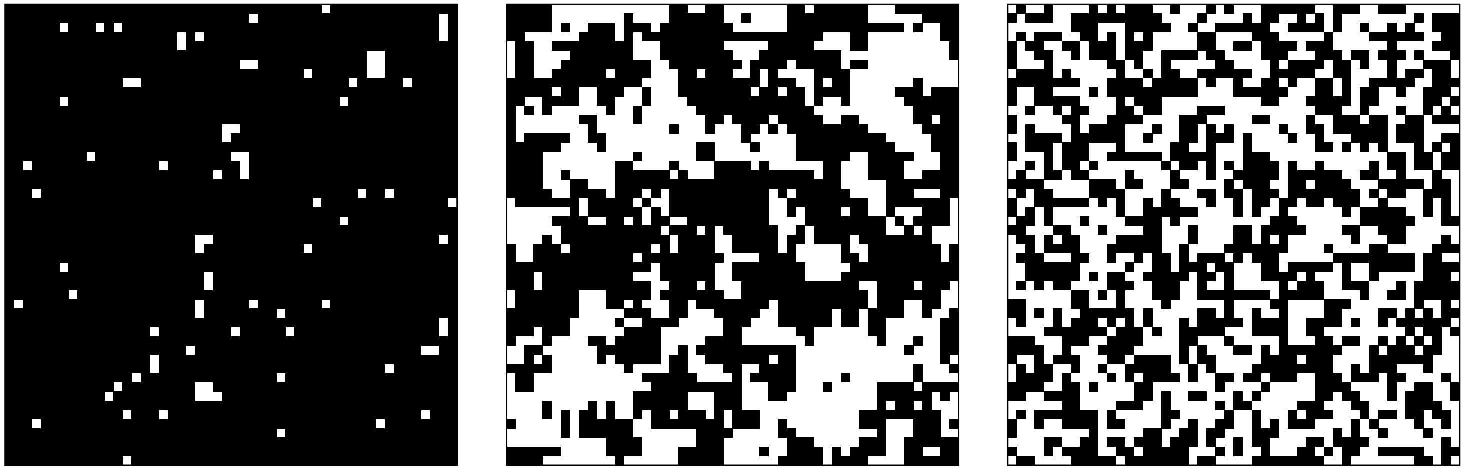
Complexity-Entropy Diagram for 2D Ising Model Phase Transition, continued





- Convergence form of the excess entropy $\mathbf{E}_{\rm c}$ vs. entropy density h_μ versus temperature T for the two-dimensional Ising model with NN couplings and no external field.
- Model undergoes phase transition as T is varied at $T \approx 2.269$.
- There is a peak in the excess entropy is broader if plotted as a function of T than when plotted against h_μ as on the previous slide.
- Results via Monte Carlo simulation of 100x100 lattice.

Ising Model Configurations

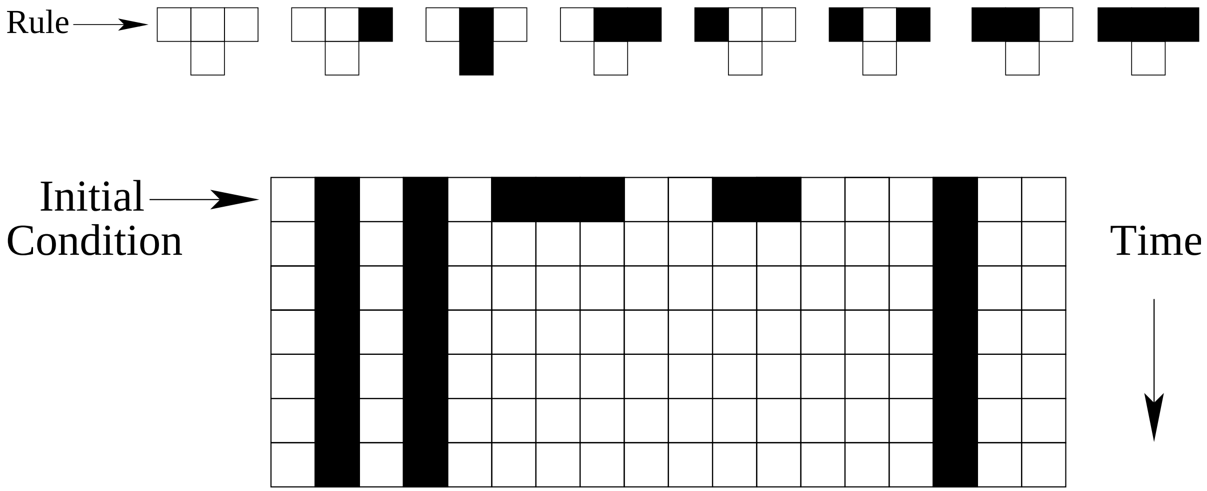


- Typical configurations for the 2D Ising model below, at, and above the critical temperature.

Cellular Automata

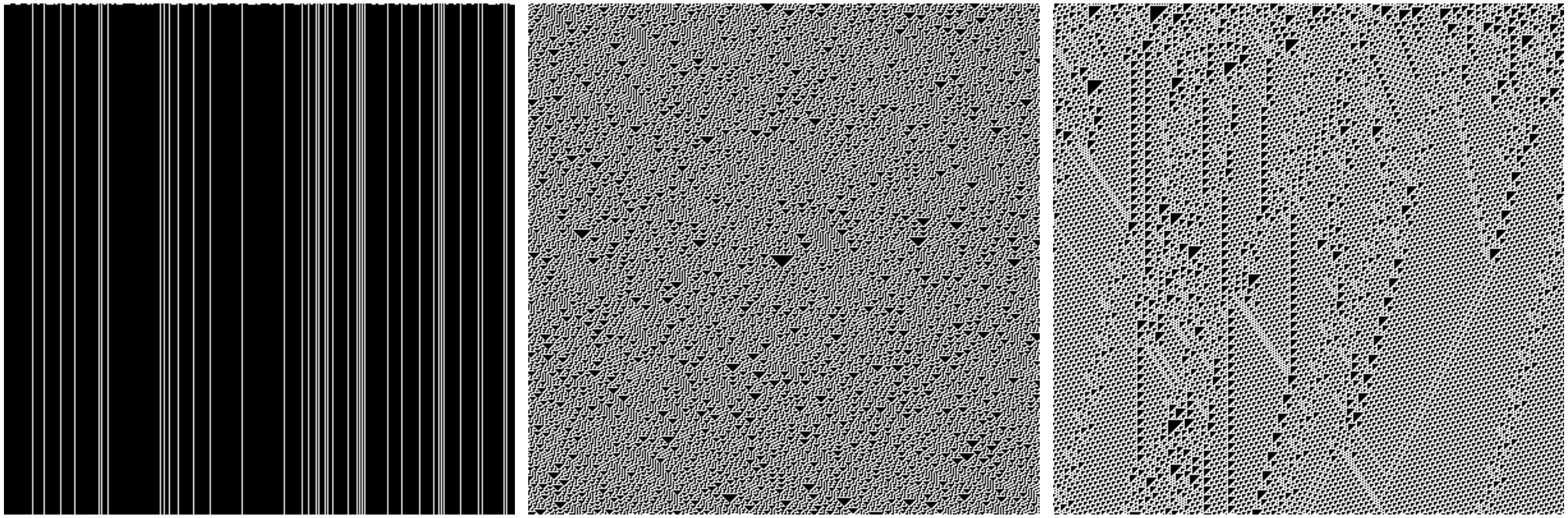
- The next row in the grid is determined by the row directly above it according to a given rule
- Start with a random initial condition

Example:



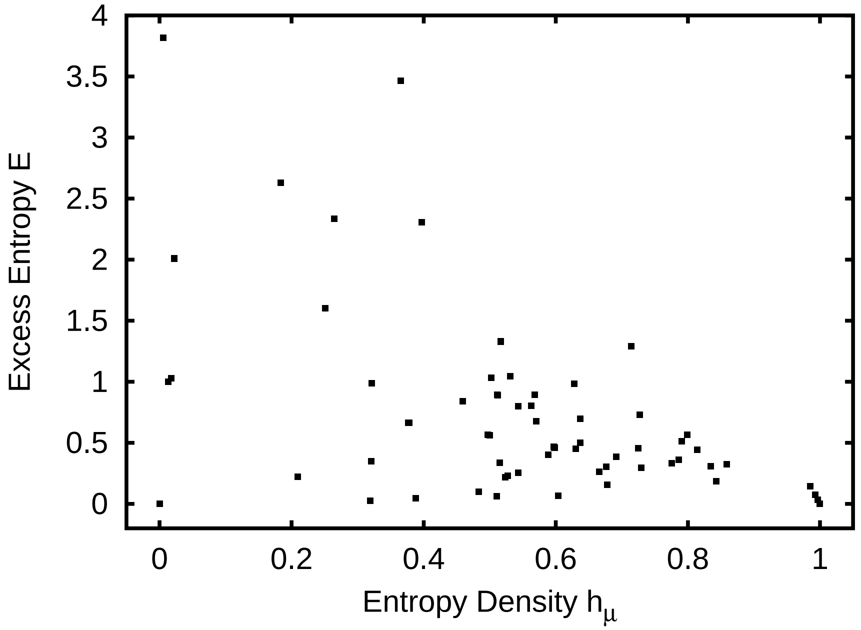
- The number of cells away from the center cell that the rule considers is known as the radius of the CA.

Different Rules Yield Different Patterns



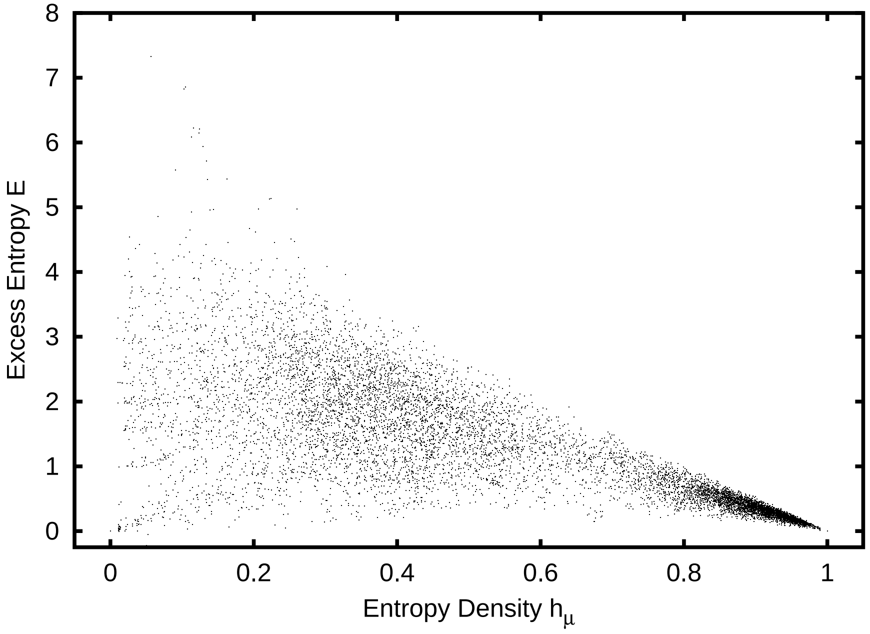
- Each pattern is for a different rule.

Complexity-Entropy Diagram for Radius-1, 1D CAs (aka Elementary CAs, or ECAs)



- Excess entropy $\mathbf{E}$ and entropy density h_μ for all distinct (88) one-dimensional elementary cellular automata.
- $\mathbf{E}$ and h_μ from the spatial strings produced by the CAs.
- Since there are so few ECAs, it's hard to discern a pattern. What if we try radius-2 CAs?

Complexity-Entropy Diagram for Radius-2, 1D CAs



- Excess entropy $\mathbf{E}$ vs. entropy rate h_μ for $10,000$ radius-2, binary CAs.
- $\mathbf{E}$ and h_μ from the spatial strings produced by the CAs.
- The CAs were chosen uniformly from the space of all such CAs.
- There are around 4.3×10^9 such CAs, so it is impossible to sample the entire space.

What is Typical?

- It is hard to know what it means to sample a model class with 4.3 billion members, especially when there's not a clear notion of what it means for particular CAs to be "close" to each other.
- We can sample uniformly. But if the real world can be described by CAs there's no reason to believe that it sampled the model space uniformly.
- What if we want to look for structure in the model space? We could parametrize the space in some way and then vary the parameter.

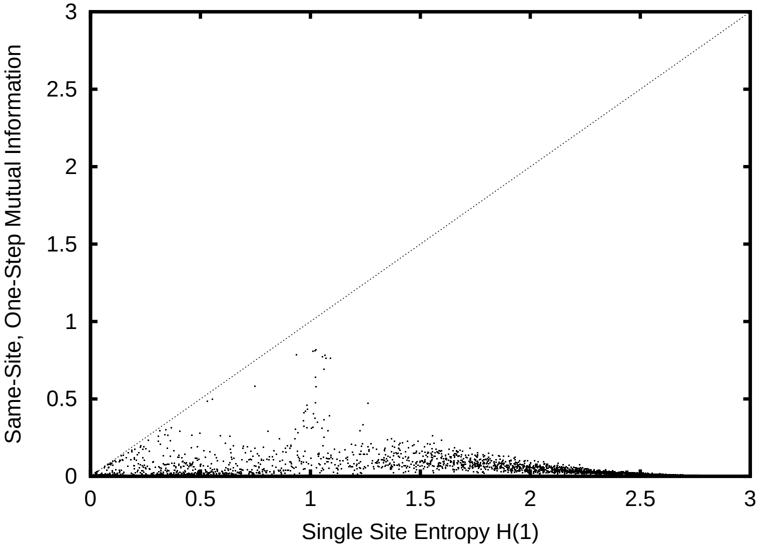
Langton's Lambda

- One such parametrization is Langton's λ. (Langton, *Physica D*, **42**:12, 1990.)
- Let N denote the number of sites in the neighborhood, K the alphabet size, and n the number of particular neighborhoods in a particular CA rule that map to 0.
- Then λ is defined as the fraction of nonzero transitions:

$$\lambda \equiv \frac{K^N - n}{K^N} \ .$$

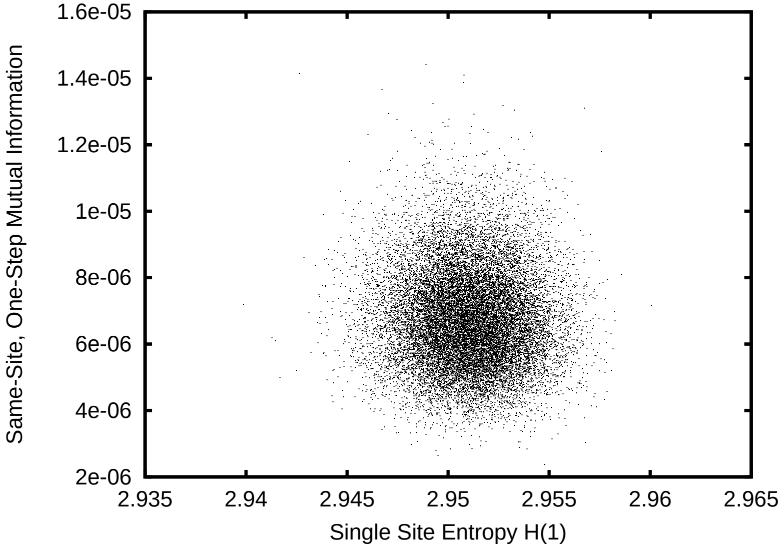
- Langton considered two-dimensional, 8-states, radius-1 cellular automata.
- There are around $10^{30,000}$ such CAs!
- Sweep λ from 0 to 1 in increments of 0.01. Randomly generate 50 different CA rules for each λ.
- For each CA, calculate single-site entropy and one-step mutual information.

“Complexity” vs. “Entropy” for 2D 8-state CAs: λ Sampling



- Single-site entropy $H(1)$ and same-site, one-step mutual information I_2^t for $r = 1$, $K = 8$ two-dimensional cellular automata.
- The straight-line is an exact upper bound for I_2^t as a function of $H(1)$.
- Plots of this form were sometimes taken to indicate a complexity-entropy transition in CA rule space.
- However, this doesn’t look like a sharp transition to me.
- What happens if we sample the CAs uniformly instead of by sweeping λ?

“Complexity” vs. “Entropy” for 2D 8-state CAs: Uniform Sampling



- Single-site entropy $H(1)$ and same-site, one-step mutual information I_2^t for $r = 1$, $K = 8$ two-dimensional cellular automata.

- $20,000$ rules were sampled randomly.

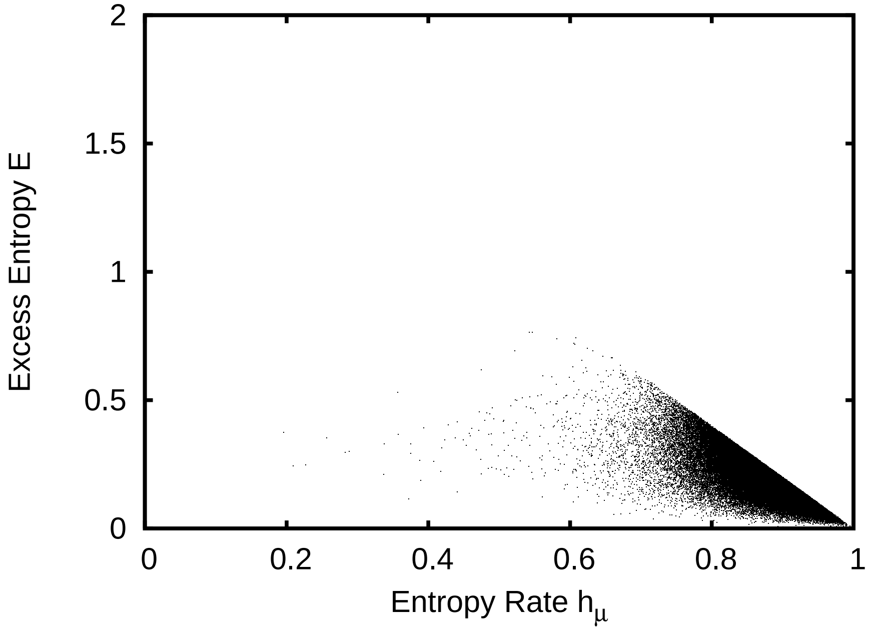
- Note the very small variation in $H(1)$ and I_2^t. All data in this figure would appear as a single dot in the lower right hand corner of the previous figure.

- Conclusion: If there is a complexity-entropy transition in CAs, it is, in a sense, a “projection” arising from the λ parametrization.

On CA Phase Transitions, or not

- I have not found any evidence of anything approaching a complexity-entropy phase transition for CAs.
- There is, however, evidence that there is a sharp transition in single-site entropy $H(1)$ as a function of λ. (Li, et al, *Physica D*, **45**:77, 1990. Wooters and Langton, *Physica D*, **45**:95, 1990.
- A mean-field (infinite-radius), infinite-K-limit argument suggests that the λ for this transition is $\lambda_c \approx 0.27$. Below this critical value $H(1)$ vanishes; above, it is nonzero (Wooters and Langton).
- However, taking the mean-field limit as described above results in a class of models that is quite far removed from CAs.
- Also, a transition with respect to λ is a transition as the rule is varied. This is very different than a transition in terms of h_μ and $\mathbf{E}$, which are functions of the configurations themselves.

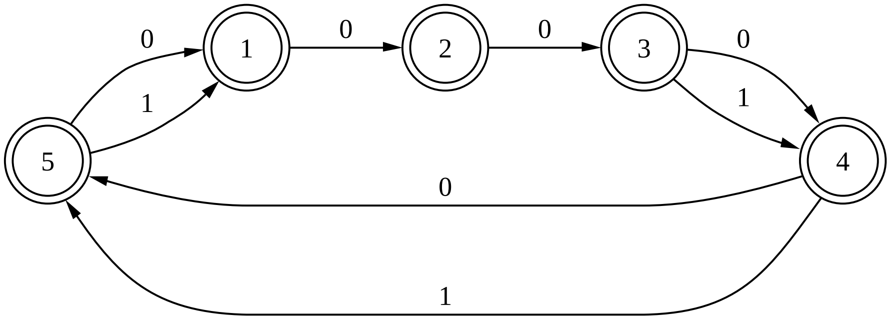
Complexity-Entropy Diagram for Markov Models

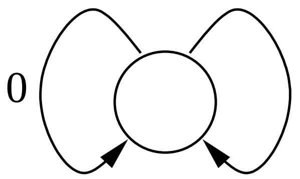


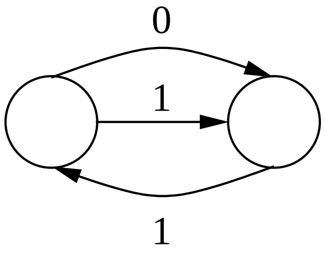
- Excess entropy $\mathbf{E}$ vs. entropy rate h_μ for $100,000$ random Markov models.
- The Markov models here have four states, corresponding to dependence on the previous two symbols, as in the 1D NNN Ising model.
- Transition probabilities chosen uniformly on $[0,1]$ and then normalized.
- Note that these systems have no forbidden sequences.

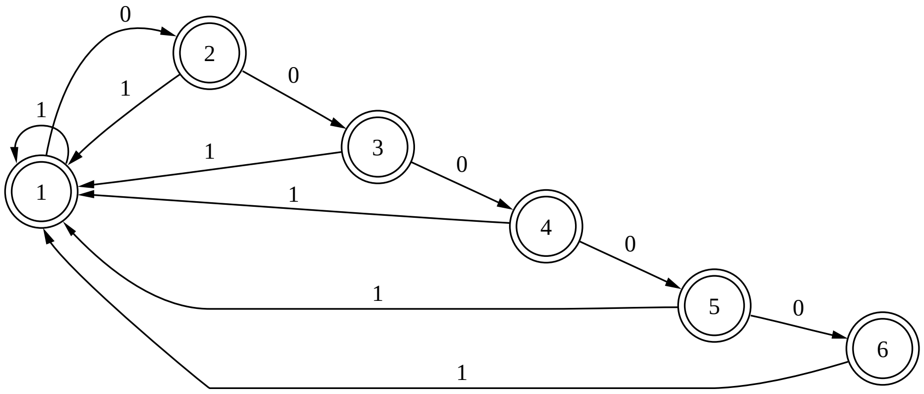
Topological Markov Chain Processes

- Consider finite-state machines that produce 0's and 1's.
- Assume all branching transitions are equally probable
- Examples:

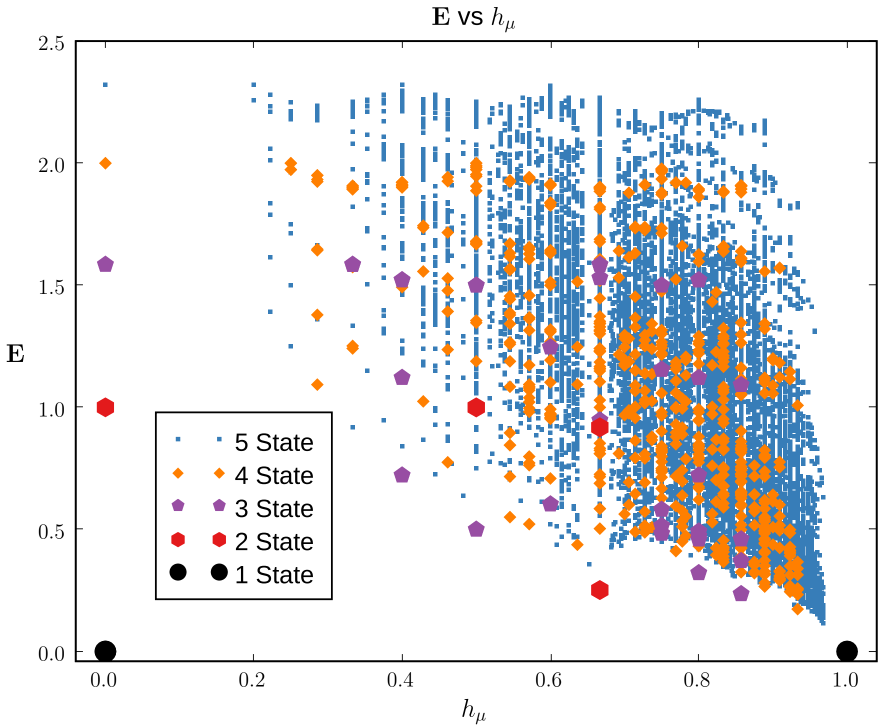








Complexity-Entropy Diagram for Topological Processes



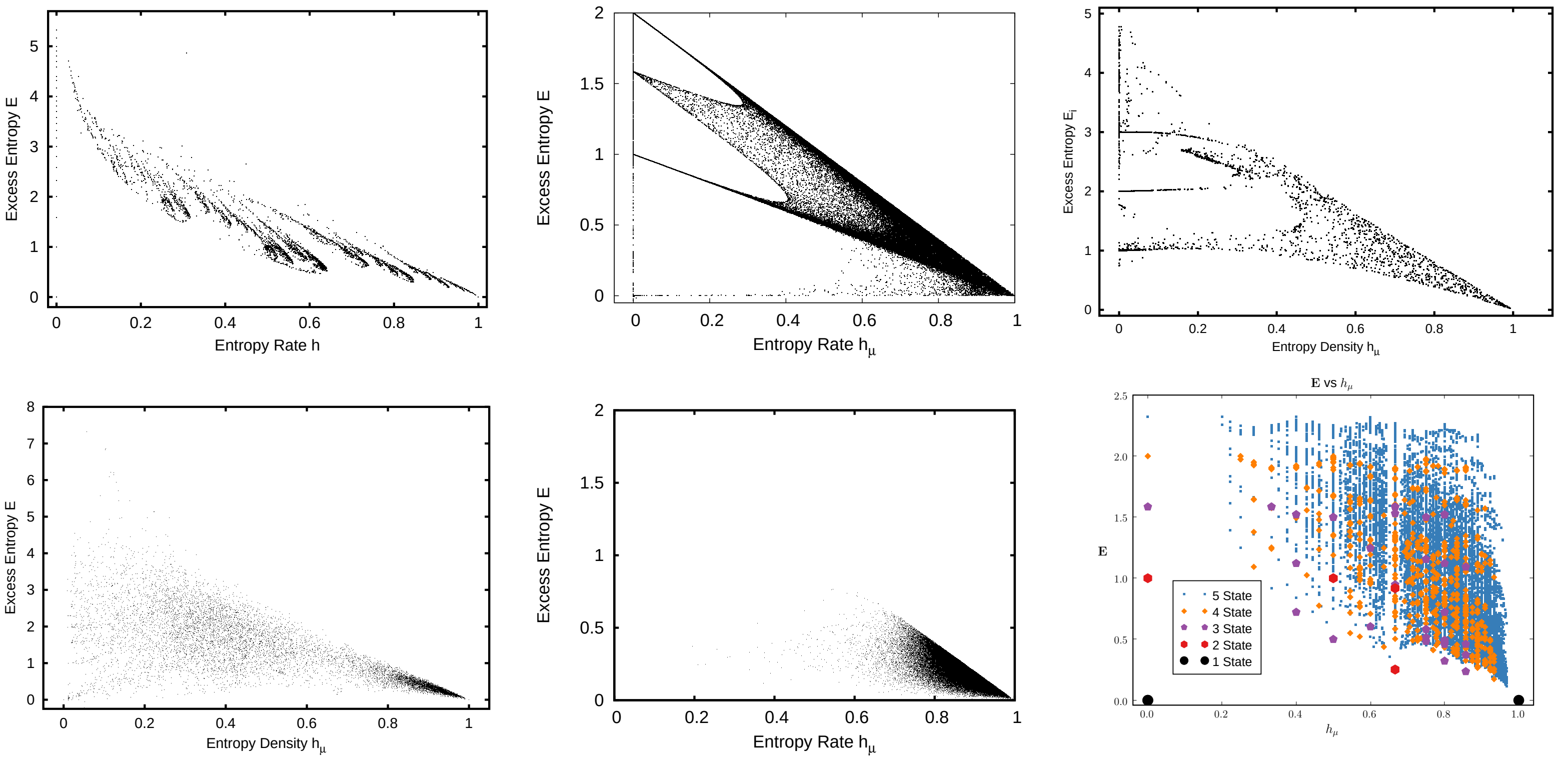
- $h_\mu, \mathbf{E}$ pairs for all $14,694$ distinct topological processes of $n = 1$ to $n = 6$ states.
- Enumeration algorithm by Carl McTague, **E** calculation by Chris Ellison.
- Note the prevalence of high-entropy, high-complexity processes.

A Gallery of Complexity-Entropy Diagrams

The next slide shows, left to right, top to bottom, complexity-entropy diagrams for:

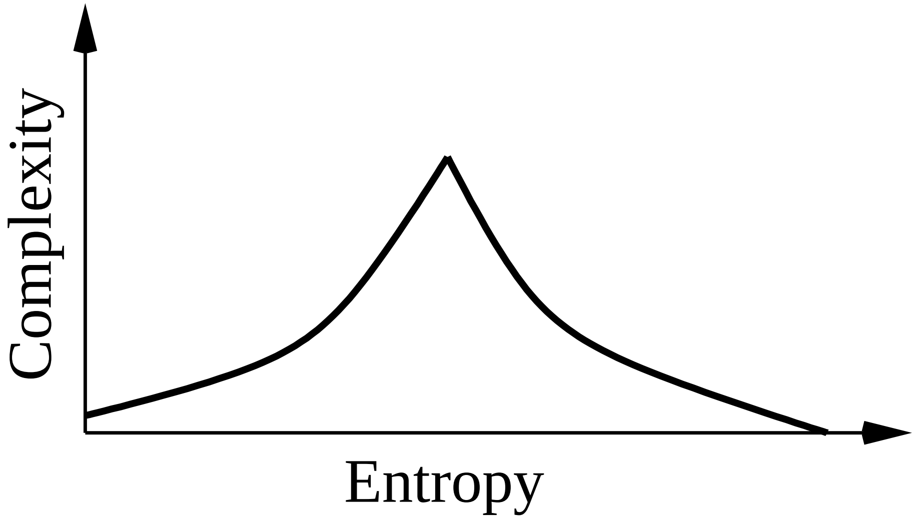
1. Logistic Equation
2. One-Dimensional Ising model with nearest- and next-nearest-neighbor interactions
3. Two-Dimensional Ising model with nearest- and next-nearest-neighbor interactions
4. One-Dimensional radius-2 cellular automata
5. Random Markov chains
6. All 6-state topological processes

A Mosaic of Complexity-Entropy Diagrams



Complexity-Entropy Diagrams: Summary

- Is it the case that there is a universal complexity-entropy diagram?



- No!
- However, because of this non-universality, complexity-entropy diagrams provide a useful way to compare the information processing abilities of different systems.
- Complexity-entropy plots allow comparisons across a broad class of systems.

Complexity-Entropy Diagrams: Conclusions

- There is not a universal complexity-entropy curve.
- Complexity is not necessarily maximized at intermediate entropy values.
- It is not always the case that there is a sharp complexity-entropy transition.
- Complexity-entropy diagrams provide a way of comparing the information processing abilities of different systems in a parameter-free way.
- Complexity-entropy diagrams allow one to compare the information processing abilities of very different model classes on similar terms.
- There is a considerable diversity of complexity-entropy behaviors.

Edge of Chaos?

Is there an edge of chaos to which systems naturally evolve? My very strong hunch is no, not in general. See the following pair of papers.

- Packard, “Adaptation to the Edge of Chaos” in *Dynamic Patterns in Complex Systems*, Kelso et.al, eds., World Scientific, 1988
- Mitchell, Hraber, and Crutchfield “Revisiting the ‘Edge of Chaos’ ” *Complex Systems*, 7:89-130, 1993. (Response to Packard, 1988).

Transitions in CA Rule Space?

- Is there a sharp complexity transition in CA rule space? No, unless you parametrize the space of CAs in a very particular way. The “transition,” then, is a result of the parametrization and not the space itself.

Transitions in CA Rule Space References

- Langton. “Computation at the Edge of Chaos,” *Physica* D (1990).
- Li, Packard and Langton, “Transition Phenomena in Cellular Automata Rule Space” *Physica D* 45 (1990) 77.
- Wooters and Langton, “Is there a Sharp Phase Transition for Deterministic Cellular Automata?”, *Physica D* 45 (1990) 95.
- Crutchfield, “Unreconstructible at Any Radius”, *Phys. Lett.* A 171: 52-60, 1992.
- Feldman, et al, “Organization of Intrinsic Computation.” In preparation.

Part XII

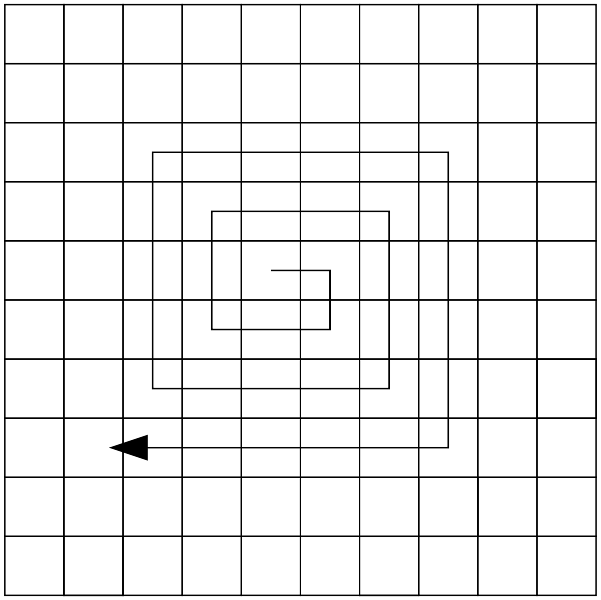
Toward Two-dimensional Complexity Measures: Some Answers, More Questions

Moving to Two Dimensions

- There are a number of measures of complexity/pattern/organization for one-dimensional systems that are fairly well understood and have found numerous applications.
- The same is not true for two dimensions.
- Initial Premises:
  1. 2D patterns are fundamentally different than 1D patterns.
  2. We do not seek a single measure of complexity or structure; there will be many complexity measures.
  3. Important to specify what you're trying to measure and why.
- For a full discussion with many references, please see Feldman and Crutchfield, *Phys. Rev. E*, **67**:051104. 2003.
- In what follows, I'll mainly focus on two-dimensional extensions of the excess entropy.

Two Dimensions = One dimension?

- One approach is to parse the 2D pattern by following a 1D path through it
- One could then apply 1D measures of structure such as $\mathbf{E}$ or C_μ.



- However, this is certainly not the way we usually understand or process a 2D pattern.

Two Dimensions = One dimension?

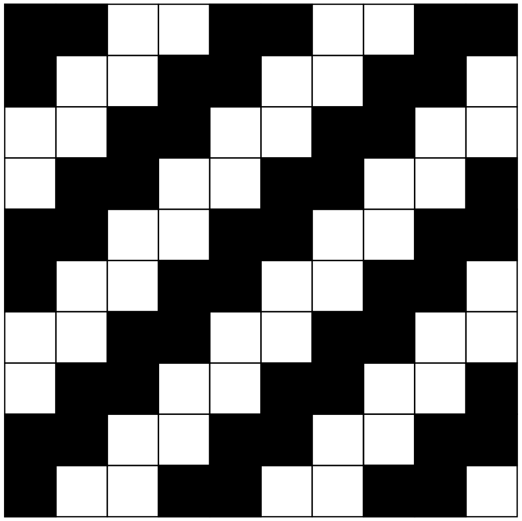
- As a formal matter, a space-filling 1D path will give the correct entropy density.
- However, the 1D path will introduce spurious long-range correlations.
- The measured complexity is thus a property of both the configuration and the path you choose through the configuration.
- **Q1: Is there a general way to "subtract off" the complexity associated with the path, leaving one with a quantity that measures a property of the configuration alone?**
- Probably not in general, but maybe for particular cases? However, this doesn't mean that 1D paths aren't interesting.

Walking vs. Flying: Worms and Birds

- **Birds** see entire pattern at once, from above.
- **Worms** can only “see” current cell. Worms can move around from square to square.
- The worms and birds experience the pattern very differently.
- Do the worms and the birds agree on any aspects of the complexity of the configuration?
- A dilemma: it would take a worm a great deal of memory to keep track of where they have been.
- **Q2: What are the relationships between the spatial complexity properties of the configuration and the complexity as perceived by an agent free to roam the lattice?** Perhaps there are some interesting analogies with intrinsic and extrinsic quantities in differential geometry.

2D Synchronization

- Suppose a worm-like agent knows a 2D periodic pattern. It is then dropped onto a grid and it can move around. It initially doesn't know where it is.
- **Q3: How difficult is it for such an agent to synchronize to a periodic 2D pattern, and how is this reflected in the pattern's information theoretic properties?**
- This strikes me as a well posed, tractable problem.



Toward a Two-dimensional Entropic Analysis: Entropy Rate

- Goal: extend 1D info theory—entropy rate h_μ and excess entropy $\mathbf{E}$—to two spatial dimensions.
- In so doing, we will need to think carefully about the definitions of these quantities in 1D.
- We'll start by reconsidering the entropy rate h_μ in one dimension.
- The definition of h_μ is the entropy per symbol:

$$h_\mu \equiv \frac{H(L)}{L} \; .$$

- Pictographically, this may be written:

$$h_\mu \equiv \lim_{L \to \infty} \frac{H\left[\,\overset{\longleftarrow L \longrightarrow}{\square\square\square\square\square}\,\right]}{L} \; . \tag{23}$$

Toward a 2D Entropic Analysis: 2D Entropy Density

- In analogy with Eq. (23), let's define the 2D entropy density as the entropy of a rectangle divided the number of sites in the rectangle, in the limit that the rectangle size goes to infinity: **Entropy density:**

$$h_\mu \equiv \lim_{M,N\to\infty} \frac{H\left[\begin{array}{c} \longleftarrow M \longrightarrow \\ \begin{array}{|c|c|c|c|c|}\hline \; & \; & \; & \; & \; \\ \hline \; & \; & \; & \; & \; \\ \hline \; & \; & \; & \; & \; \\ \hline \; & \; & \; & \; & \; \\ \hline \; & \; & \; & \; & \; \\ \hline \end{array} \end{array} \quad \begin{array}{c} \uparrow \\ N \\ \downarrow \end{array}\right]}{MN} \; . \tag{24}$$

- This limit exists for a translationally invariant system, provided that the ratio N/M remains finite and non-zero when taking the limit.

- This definition seems to me to be the only sensible starting point for 2D entropy.

Toward a 2D Entropic Analysis: Conditional Entropy

- In 1D, the entropy density can be expressed as the limit of the entropy of a single spin conditioned on a block of adjacent spins.
- $h_\mu(L) \, = \, H[S_L|S_1 S_1 \ldots S_{L-1}]$
- Or, pictographically:

$$h_\mu(L) \, = \, H[\, \boxtimes \, | \, \overset{(L-1)\longrightarrow}{\boxtimes\square\square\square\square\square} \,] \, . \qquad (25)$$

- The entropy rate h_μ is then obtained by taking the $L \to \infty$ limit:

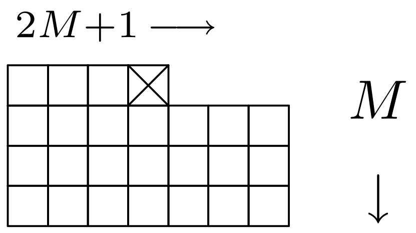
$$h_\mu \, = \, \lim_{L\to\infty} h_\mu(L) \qquad (26)$$

- Can we do the same thing in 2D? That is, can we come up with a single-spin entropy conditioned on an appropriate block of spins such that in some limit it is equal to the entropy density of Eq. (23)?

2D Conditional Entropy

- Yes.
- Define the following conditional entropy:

$$h_\mu(M) \equiv H\Big[\ \boxtimes\ \Big|$$



$$\Big] . \qquad (27)$$

- Then,

$$h_\mu = \lim_{M \to \infty} h_\mu(M) . \qquad (28)$$

where h_μ is that which was defined in Eq. (23).

- This form of conditional entropy has been discovered independently several times.
- Earliest reference: Alexandrowicz, *J. Chem. Phys*, **55**:2265, 1971.
- This result has been used to analyze a variety of lattice systems. See Feldman and Crutchfield (2003) for references.

1D ⟶ 2D Excess Entropy?

- In 1D there were three different expressions for the excess entropy:

1. **$\mathbf{E}$ as sum of entropy density over-estimates:**

$$\mathbf{E}_{\rm c} \equiv \sum_{L=1}^{\infty} h_\mu(L) - h_\mu \ . \qquad (29)$$

2. **$\mathbf{E}$ as sub-extensive part of $H(L)$:**

$$H(L) = H\Big[\overset{\longleftarrow L \longrightarrow}{\square\square\square\square\square}\Big] \sim \mathbf{E}_{\rm a} + h_\mu L \ , \ \text{as } L \to \infty \ . \qquad (30)$$

3. **$\mathbf{E}$ as mutual information:**

$$\mathbf{E}_{\rm i} = \lim_{L\to\infty} I\Big[\overset{\longleftarrow L}{\square\square\square\square\square}\ ;\ \overset{L \longrightarrow}{\square\square\square\square\square}\Big] \ . \qquad (31)$$

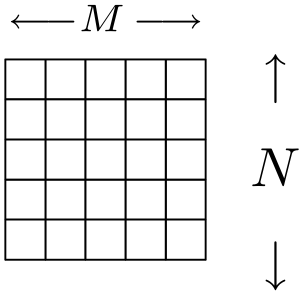
- In 1D each of these expressions yields the same value for $\mathbf{E}$.

Several 2D Excess Entropies

- In 2D generalizations of these different expressesion for the excess entropy yield different results.
- 2D Excess Entropy as a sum of entropy density over-estimates:

$$\mathbf{E}_{\mathrm{c}} \equiv \sum_{M=1}^{\infty} h_\mu(M) - h_\mu \ . \tag{32}$$

- 2D Excess Entropies as sub-extensive parts of the total entropy growth:

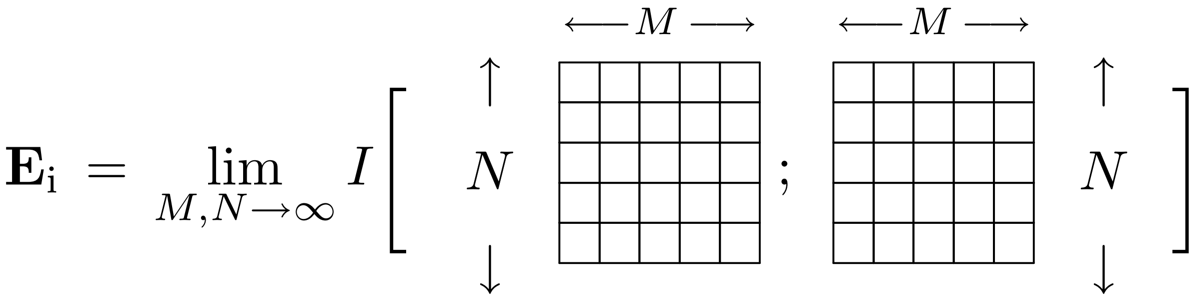


$$H(M,N) = H\Big[\qquad \Big] \sim \mathbf{E}_{\mathrm{a}} + \mathbf{E}_{\mathrm{a}}^{(x)} M + \mathbf{E}_{\mathrm{a}}^{(y)} N + h_\mu M N \tag{33}$$

- **Q4: What do these different sub-extensive terms mean?**

Several 2D Excess Entropies, continued

- 2D Excess Entropy as mutual information:


$\mathbf{E}_{\mathrm{i}} = \lim_{M,N\to\infty} I$
$\longleftarrow M \longrightarrow$
$\longleftarrow M \longrightarrow$
$\uparrow N \downarrow$
;
$\uparrow N \downarrow$

(34)

- One could also define an excess entropy between two blocks that are joined by a horizontal boundary.
- For that matter, one could use a diagonal boundary or even a nonlinear boundary.

Too Many 2D Excess Entropies?

- I do not think it's a problem that there are multiple two-dimensional excess entropies.
- 2D patterns are sufficiently more subtle than 1D patterns that it seems natural to have multiple excess entropies.
- A similar thing happens with derivatives:
  - At a point, there is one derivative of a function of one variable.
  - There are, in a sense, an infinite number of derivatives of a two-dimensional function.
- **Q5: What does 2D discrete calculus suggest about 2D information theory? Are there useful analogs of the divergence or Green's Theorem?**
- I suspect that thinking of entropy with 2D discrete calculus will be very fruitful, as was the case in 1D.

Other 2D Entropy Questions

- **Q6: What are the relationships among the 3 different 2D excess entropy forms?**
  - Preliminary work shows that $\mathbf{E}_\mathrm{i}$ and $\mathbf{E}_\mathrm{c}$ are very similar, but not identical.
- In 1D there is a natural ordering of blocks of adjacent spins.
- There is no such natural ordering in 2D; lattices are partially ordered sets.
- **Q7: Are there mathematical results or intuitions from the study of lattices and posets that could help us?**

Another 2D Approach

- For a 2D system that evolves in time, consider a particular site.
- The site’s “past” is the set of states that could have influenced it.
- For a radius-1 CA, this will be a cone, opening upward into the past.
- One can then apply computational mechanics using this cone as the pasts, and a future-cone as the futures.
- References:
  - Shalizi, et al., *Phys. Rev. Lett.* **93**:118701, 2004.
  - Shalizi, et al. *Phys. Rev. E.* **73**: 036104. 2006.
- The light-cone approach works well for the causal (temporal) structure of lattices, but it does not speak directly to the two-dimensional pattern.
- For another, related approach, see Young, et al. *Physical Review Letters.* **94**:098701. 2005.

Two-Dimensional Complexity Summary

- In my view, the above considerations are primarly of mathematical interest.
- Although there have been some important successes, this work is still essentially one-dimensional in character.
- My guess is that a very different approach is needed. Humans understand 2D patterns in a way that is fundamentally different than 1D.
- I suspect that there will not be a single, comprehensive theory of 2D patterns.
- I think in 2D it is especially important to think carefully about why one wants to measure complexity, the assumptions behind a given complexity measure, and the context in which it will be used.
- I think this area is wide open, challenging, potentially rewarding, and quite fun.

Part XIII

Thoughts on the Subjectivity of Complexity

Thoughts on the Subjectivity of Complexity

- There is not a general, all-purpose, objective measure of complexity.
- Objective knowledge is, in a sense, knowledge without a knower.
- Subjective knowledge depends on the knower. In a sense, it is an opinion.
- Complexity, at least as I've been using the term, is a measure of the difficulty of describing or modeling a system.
- This will depend on who is doing the observing and what assumptions they make.
- Depending on the observer a system may appear more or less complex.
- Entropy and complexity are often related in interesting ways.
- I'll illustrate this with four examples.

Example I: Disorder as the Price of Ignorance

- Let us suppose that an observer seeks to estimate the entropy rate.
- To do so, it considers statistics over sequences of length L and then estimates h_μ using an estimator that assumes $\mathbf{E} = 0$.
- Call this estimated entropy ${h_\mu}'(L)$. Then, the difference between the estimate and the true h_μ is (Prop. 13, Crutchfield and Feldman, 2003):

$$h'_\mu(L) - h_\mu \; = \; \frac{\mathbf{E}}{L} \; .$$

- In words: The system appears more random than it really is by an amount that is directly proportional to the the complexity $\mathbf{E}$.
- In other words: regularities ($\mathbf{E}$) that are missed are converted into apparent randomness ($h'_\mu(L) - h_\mu$).
- Crutchfield and Feldman, “Regularities Unseen, Randomness Observed.” *Chaos*. 15:23-54. 2003.

Example II: Effects of Bad Discretization

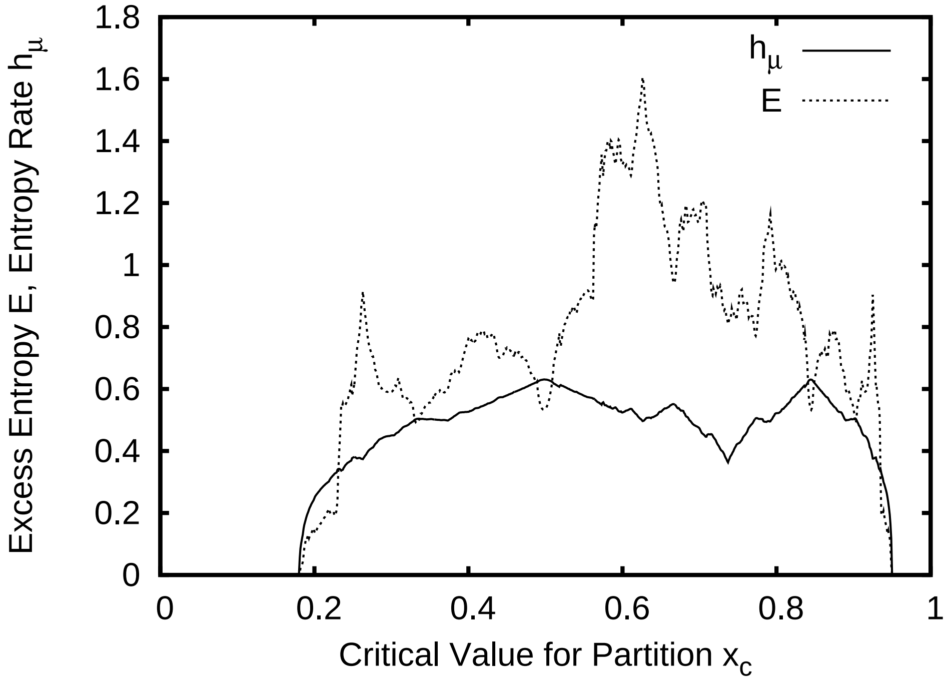
- Iterate the logistic equation: $x_{n+1} = f(x_n)$, where $f(x) = rx(1-x)$.
- Result is a sequence of numbers. E.g., $0.445, 0.894, 0.22, 0.344, \ldots$.
- Generate symbol sequence via:

$$s_i = \begin{cases} 0 & x \leq x_c \\ \\ 1 & x > x_c \end{cases} .$$

- As we've seen, for many values of r this system is chaotic.
- It is well-known that if $x_c = 0.5$, then the entropy of the symbol sequence is equal to the entropy of the original sequence of numbers.
- Moreover, it is well known that h_μ is maximized for $x_c = 0.5$.

Example II: Effects of Bad Discretization (continued)

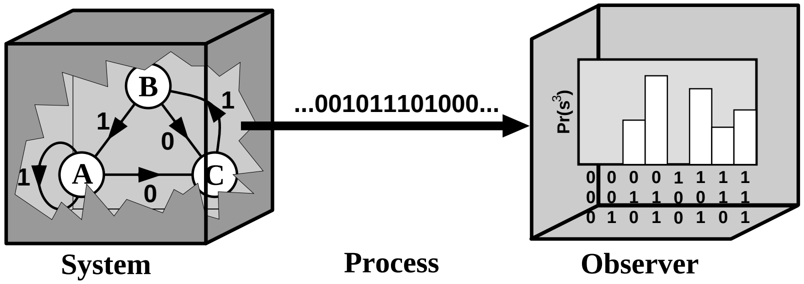
- Our estimates for h_μ and $\mathbf{E}$ depend strongly on x_c.
- Using an $x_c \neq 0.5$ leads to an h_μ is always lower than the true value.
- Using an $x_c \neq 0.5$ can lead to an over- or an under-estimate of $\mathbf{E}$.



- Note: $r = 3.8$ in this figure.

Example III: A Randomness Puzzle

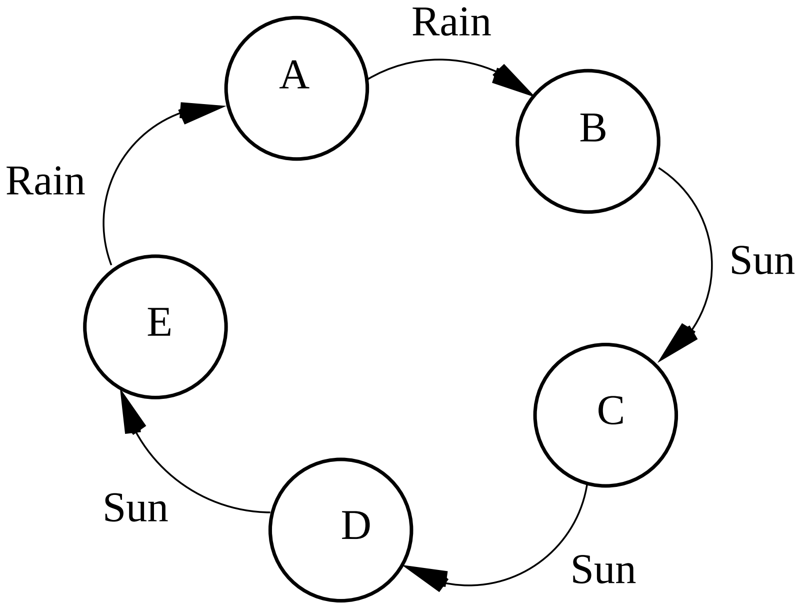
- Suppose we consider the binary expansion of π. Calculate its entropy rate h_μ and we'll find that it's 1.
- How can π be random? Isn't there a simple, deterministic algorithm to calculate digits of π?
- It is not random if one uses Kolmogorov complexity, since there is a short algorithm to produce the digits of π.
- It is random if one uses histograms and builds up probabilities over sequences.
- This points out the *model-sensitivity* of both randomness and complexity.



- Histograms are a type of model. See, e.g., Knuth. arxiv.org/physics/0605197. 2006.

Example IV: Unpredictability due to Asynchrony

- Imagine a strange island where the weather repeats itself every 5 days. It's rainy for two days, then sunny for three days.



- You arrive on this deserted island, ready to begin your vacation. But, you don't know what day it is: $\{A, B, C, D, E\}$.
- Eventually, however, you will figure it out.

Example IV: Unpredictability due to Asynchrony

- Once you are synchronized—you know what day it is—the process is perfectly predictable; $h_\mu = 0$.
- However, before you are synchronized, you are uncertain about the internal state. This uncertainty decreases, until reaching zero at synchronization.
- Denote by $\mathcal{H}(L)$ the average state uncertainty after L observations are made.
- The total state uncertainty experienced while synchronizing is the **Transient Information** $\mathbf{T}$:

$$\mathbf{T} \equiv \sum_{L=0}^{\infty} \mathcal{H}(L) \ . \tag{35}$$

Example IV: Unpredictability due to Asynchrony

- It turns out that different periodic sequences with the same P can have very different $\mathbf{T}$'s.
- For a given period P:

$$\mathbf{T}_{\max} \sim \frac{P}{2} \log_2 P \ , \tag{36}$$

and

$$\mathbf{T}_{\min} \sim \frac{1}{2} \log_2^2 P \ , \tag{37}$$

- E.g., if $P = 256$, then

$$\mathbf{T}_{\max} \approx 1024 \ , \ \text{and} \ \ \mathbf{T}_{\min} \approx 32 \ . \tag{38}$$

- For disturbingly more detail, see Feldman and Crutchfield, “Synchronizing to Periodicity.” *Advances in Complex Systems.* 7:329-355. 2004.

Summary of Examples

- In all cases choice of representation and the state of knowledge of the observer influence the measurement of entropy or complexity.

  1. Ignored complexity is converted to entropy.
  2. Measurement choice can lead to an underestimate of h_μ and an over- or under-estimate of $\mathbf{E}$.
  3. π appears random.
  4. A periodic sequence is unpredictable and, in a sense, complex.

- Hence, statements about unpredictability or complexity are necessarily a statement about the observer, the observed, and the relationship between the two.

- So complexity and entropy are relative, but in an objective, clearly specified way.

Modeling Modeling

- Much of what I have presented in the last several lectures can be viewed as an abstraction of the modeling process itself.
- These examples provide a crisp setting in which one can explore trade-offs between, say, the complexity of a model and the observed unpredictability of the object under study.
- The choice of model can strongly influence the result yielded by the model. This influence can be understood.
- The hope is these models of modeling can give us some general, qualitative insight into modeling.

Model Dependence

- There is no (computable), all-purpose measure of randomness or complexity.
- This isn't cause for despair. Just be as clear as you can about your modeling assumptions.
- Sometimes modeling assumptions can be hidden.
- I don't think will ever be a 100% objective measure of complexity. A statement about complexity will always be, to some extent, a statement about both the observer and the observed.

Part XIV

Some Open Questions and Thoughts on Areas for Further Work

Other Open Questions and Areas to Explore

- Grammatical complexity of unimodal maps:
  - It has been observed that the symbolic dynamics of unimodal maps produce sequences that are regular languages or context-sensitive languages, but not context-free languages.
  - Xie has conjectured that this is a general result.
  - This conjecture has not yet been proved.
  - Xie, *Grammatical Complexity and One-Dimensional Dynamical Systems*, World Scientific, 1996. Hao and Xie, in *Annual Reviews in Nonlinear Science and Complexity*, 2007, to appear.
- To what extent can these techniques be extended to non-stationary systems?
  - I think there has been a little, but not much, work on this question.

Other Open Questions and Areas to Explore, Continued

- The excess entropy diverges for many (all?) sequences above regular languages in the Chomsky hierarchy.
  - Is this true for all such sequences?
  - Can anything be learned from the nature of this divergence?
  - Is the nature of the divergence related to computation theoretic quantities, e.g., the number of states in a push-down automata?
- Complexity of networks
  - What does it mean to say a network is topologically complex?
  - Might measures of complexity of the sort discussed here capture important properties of network structure or dynamics on networks?
- There are many empirical data sets that one could apply these techniques to: weather, stock market, commodity prices, neuronal signals, screaming monkeys, etc.

Other Open Questions and Areas to Explore, Continued

- Applying these measures to inhomogeneous or disordered systems such as spin glasses
  - Some preliminary work suggests that this could be a very fruitful, although somewhat difficult area.
- More closely relating complexity measures to the theory of critical phenomena.
- In general, I believe that these tools are a useful framework for considering questions of complexity, organization, and emergence?
- Is there a (semi)-rigorous, possibly objective measure of emergence?

Other Open Questions and Areas to Explore, Continued

- Is the behavior of the excess entropy or statistical complexity universal near the critical points of continuous phase transitions?
- There are some stochastic 1D CAs and interacting particle systems that I think would be very interesting to calculate the excess entropy and statistical complexity of.
- Some of these systems exhibit continuous phase transitions, and thus they would provide a nice setting for studying critical properties of the excess entropy and related quantities.
- The properties and applications of the transient information are only partially understood.
- Can one make more explicit the connection between measures of complexity and the difficulty of learning a pattern?

Part XV

Conclusion: Summary and Thoughts on “Principles of Complexity,”

Conclusion

Goals Revisited

1. Present some tools, models, paradigms that are useful in complex systems.
2. Discuss the applicability and un-applicability of these various tools.
3. Provide references and advice so you can learn more about these topics if you wish.
4. Present some thoughts about what makes the study of complex systems similar to, and different from, other types of science.
5. Provide some background which may help you get more out of other lectures.
6. Have fun.

Tools

In the introduction to my lectures, I said that:

Most tools and techniques for complex systems will need to:

1. Measure unpredictability, distinguish between different sorts of unpredictability, work with probabilities
2. Be able to measure and discover pattern, complexity, structure, emergence, etc.
3. Be inferential; be inductive as well as deductive. Must infer from the system itself how it should be represented.

I hope to have given you some tools that you can use, apply, modify, and/or reject as you see fit.

Complex Systems Science?

Is there a science or theory of complex systems? Can there be one? My hunch is that the answer is no, at least not in the usual sense of theory.

- Perhaps looking for a unifying theory of complex systems is to forget the message of emergence: that the whole is the greater than the sum of its parts, and that innovation and novelty is the norm.
- On the other hand, I don't think it's the case that every complex system is different. There may be some unifying tools, principles and ideas.
- My strong hunch is that a theory of complex systems will be primarily concerned with **methods** and **tools** as opposed to universal governing principles or equations.

What Good are Complex Systems?

- Complex systems provide a new set of paradigms or exemplars: e.g., logistic equation, random graphs, CAs, Schelling's tipping model, etc.) These serve as stories we tell about what the world is like, and provide an important counterbalance to linear, reductive, "rational" models that still are predominant in many fields.
- The model systems of the sort I've focused on here may have little to say directly about complicated, real-world phenomena. However, these systems provide a very clear setting in which to explore the discovery of pattern, and fundamental tradeoffs between randomness and order. This can hone intuition when considering other, real-world complex systems.

What Good are Complex Systems?, continued

- I believe that there is an aesthetic and perhaps even normative component to the study of complex systems. Part of what the field has in common is a group of people with similar tastes and concerns and a sense of what is interesting:
  - How the world is put together, rather than how it's taken apart.
  - A fascination with patterns and their formation.
  - A fascination with diversity.
  - A willingness to take risks.
  - A recognition of interrelationships and complexity.

Thanks and Acknowledgments

- Hao Bai-lin, Ling-Lie Chau, Jim Crutchfield, Erica Jen, Kristian Lindgren, Susan McKay, Carl McTague, Cris Moore, Richard Scalettar, Cosma Shalizi, Dan Upper, Dowman Varn, Jon Wilkins, Karl Young,