Introduction to Dynamical Systems

and

Measuring Randomness and Complexity: Information, Computation, and Emergence

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

© David P. Feldman

Overview

- First Talk: A quick overview of dynamical systems and chaos.
- Second Talk: Different techniques for measuring and quantifying different sorts of randomness, unpredictability, structure, organization, complexity, and so on.
- I will have to skip many details, but will provide references and advice so you can dig deeper if you wish.
- Along the way, I will offer some thoughts about what makes the study of complex systems similar to, and different from, other types of science.
- These notes are based on a week-long series of lectures I have given at SFI's Complex Systems Summer School since 2004.
- These slides, and the full set of CSSS slides are posted on my website.

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.

© David P. Feldman

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.

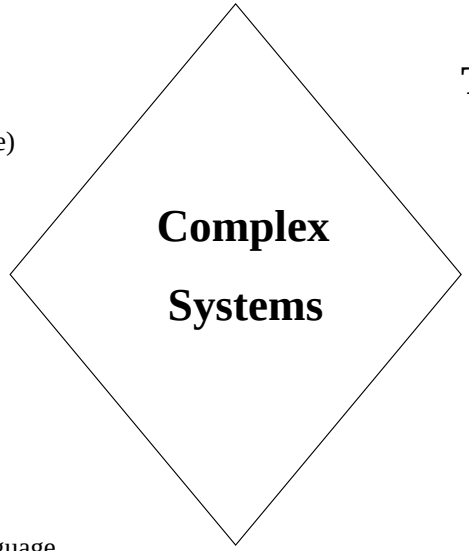
Tools

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...

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 talks will focus on some topics from the bottom and the right.
- I will also offer some thoughts on what isn't (and what might be) on the top of the quadrangle.

Complexity and Emergence: 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.
- And what is emergence? A new or hard-to-predict phenomena or structure?
- New to whom? Hard to predict to whom?

Chaos (and Complexity): The *Longue Durée*

Aubin and Dahan Delmedico write about Chaos and Dynamical Systems:

“We take the emergence of ”chaos” as a science of nonlinear phenomena... as a vast process of sociodisciplinary convergence and conceptual reconfiguration.... In order to come up with an exhaustive historical analysis of these origins [of “nonlinear science”] one needs to be able to deal at once with domains as varied as fluid mechanics, parts of engineering, and population dynamics.”

They refer to chaos as having an “**ample and bushy genealogy**.”

Aubin and Dahan Dalmedico, Writing the History of Dynamical Systems and Chaos: *Longue Durée* and Revolution, Disciplines and Cultures. *Historia Mathematica* 29 (2002), 1-67.

doi:10.1006/hmat.2002.2351

Complexity: The *Longue Durée*?

- I believe much the same can be said about Complex Systems.
- There are many different streams of thought that flow together to form the study of Complex Systems: Chaos/Dynamical Systems, Genetic Algorithms/A-life, Economics, and so on.
- The confluence of these streams is not a unitary discipline or a coherent theory, but a “sociodisciplinary convergence and conceptual reconfiguration.”
- Complex Systems has a tangled genealogy. But one of the deepest roots is the study of chaos and dynamical systems.

Part II

Introduction to Chaos: Basic Definitions, SDIC

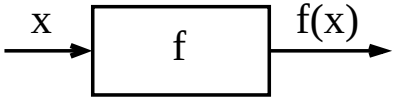
© David P. Feldman

A Brief, Introductory Overview of Dynamical Systems and Chaos

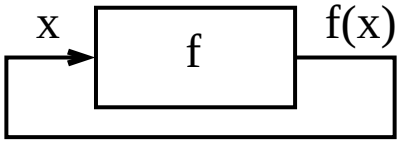
- A **Dynamical System** is any system that changes over time
  - A differential equation
  - A system of differential equations
  - Iterated functions
  - Cellular automata
- The goal of this brief introduction is to define a handful of terms, define chaos and sensitive dependence on initial conditions, 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 $f(x)$:



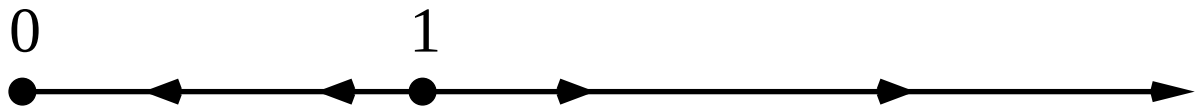
- Iterating the function is then achieved by feeding the output back to the function, making a feedback loop:



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.

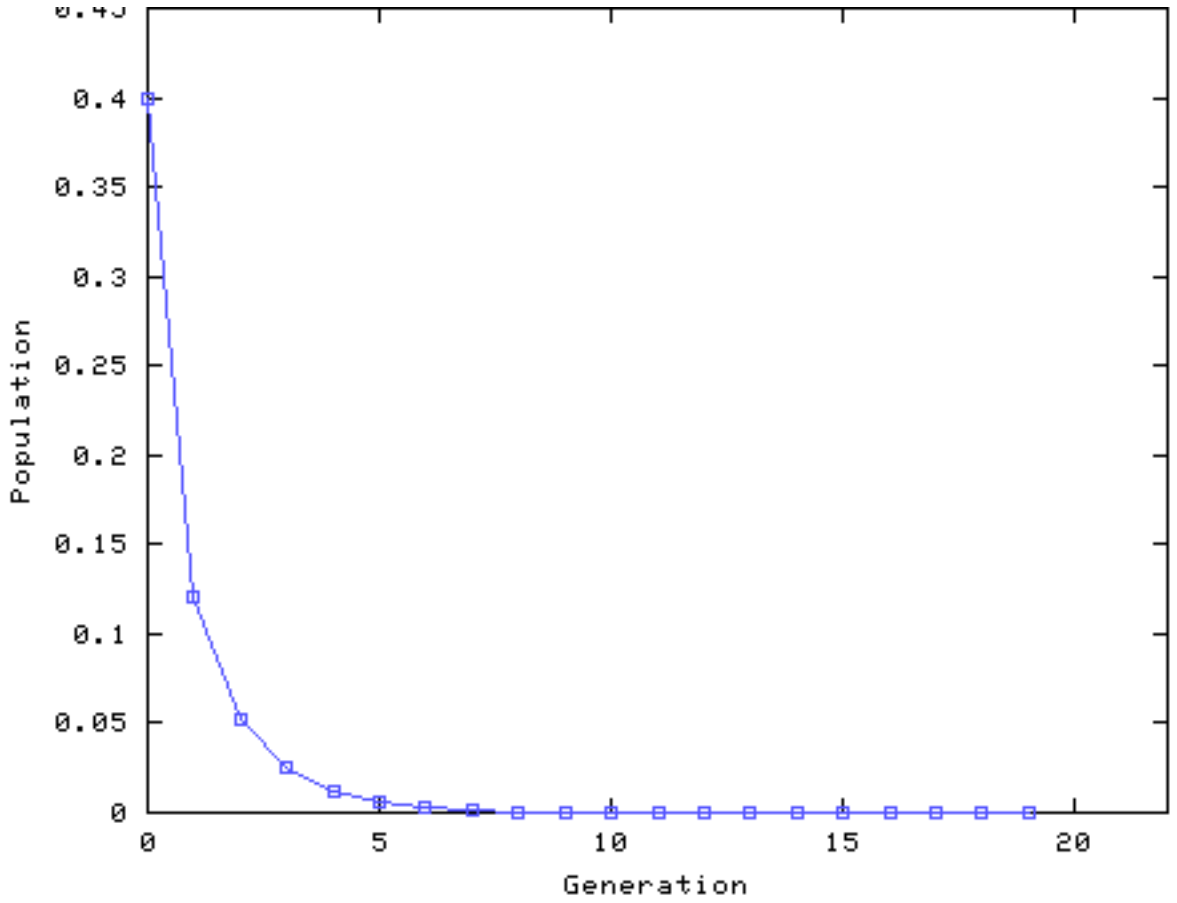
- The seed 3 tends toward infinity—it gets bigger and bigger.
- Any $x_0 > 1$ will tend toward infinity.
- If $x_0 = 1$ or $x_0 = 0$, then the point never changes. These are fixed points.
- If $0 \leq x_0 < 1$, then x_0 approaches 0.
- We can summarize this with the following diagram:



- 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

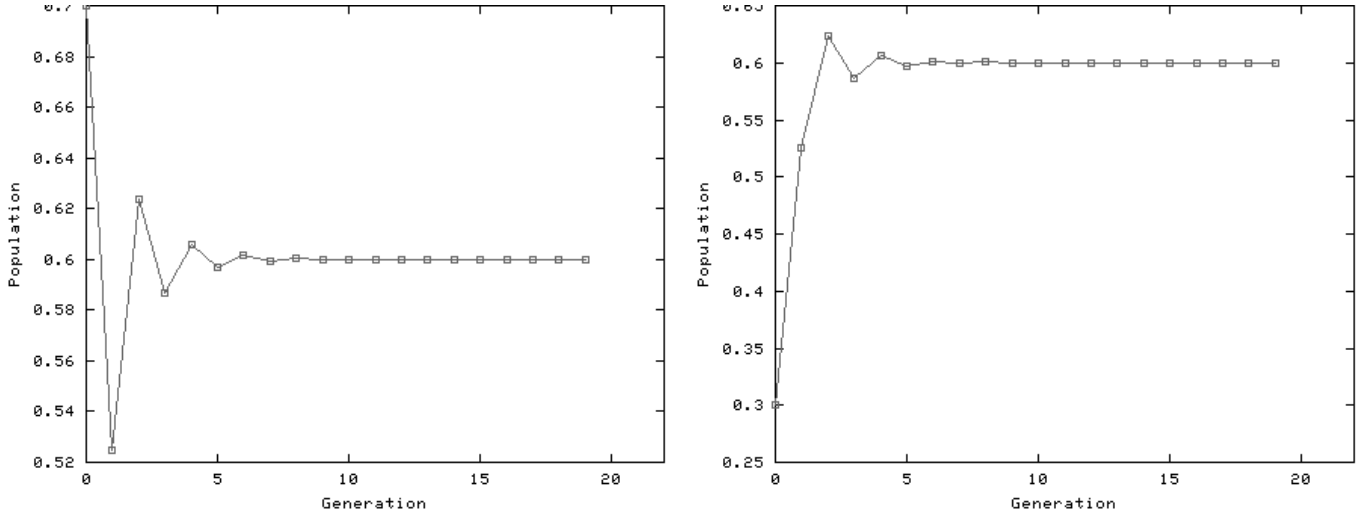
- Logistic equation: $f(x) = rx(1 - x)$.
- A simple model of resource-limited population growth.
- The population x is expressed as a fraction of the carrying capacity. $0 \leq x \leq 1$.
- r is a parameter—the growth rate—that we will vary.
- Let's first see what happens if $r = 0.5$.



- You can make your own plots at `http://hornacek.coa.edu/dave/Chaos/`.
- 0 is an attracting fixed point.

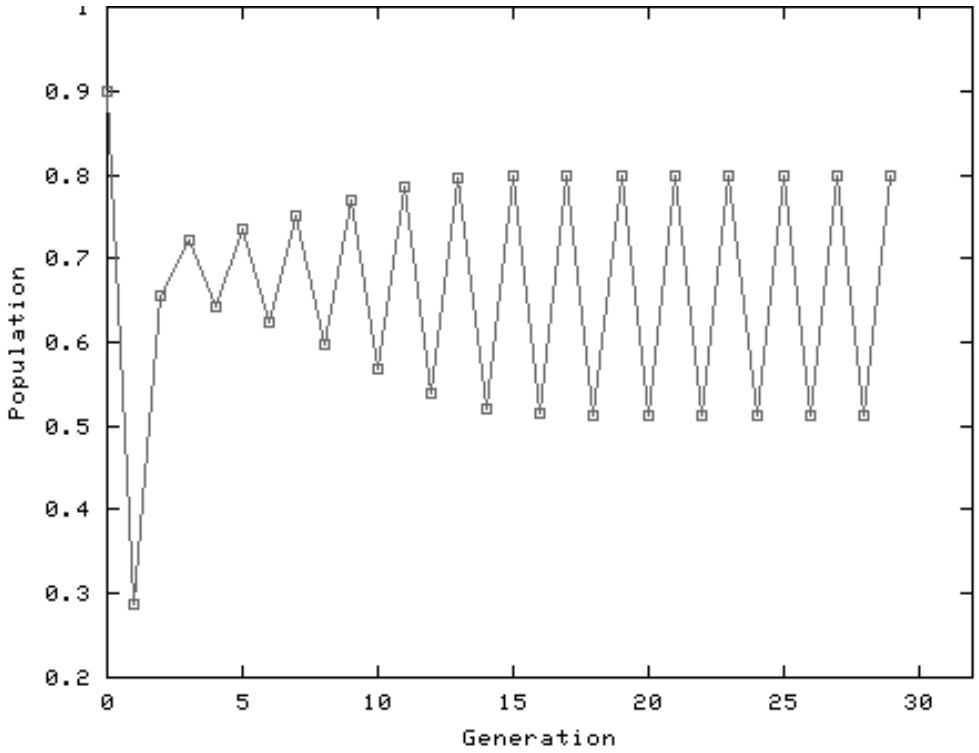
Logistic Equation, $r = 2.5$

- Logistic equation, $r = 2.5$.



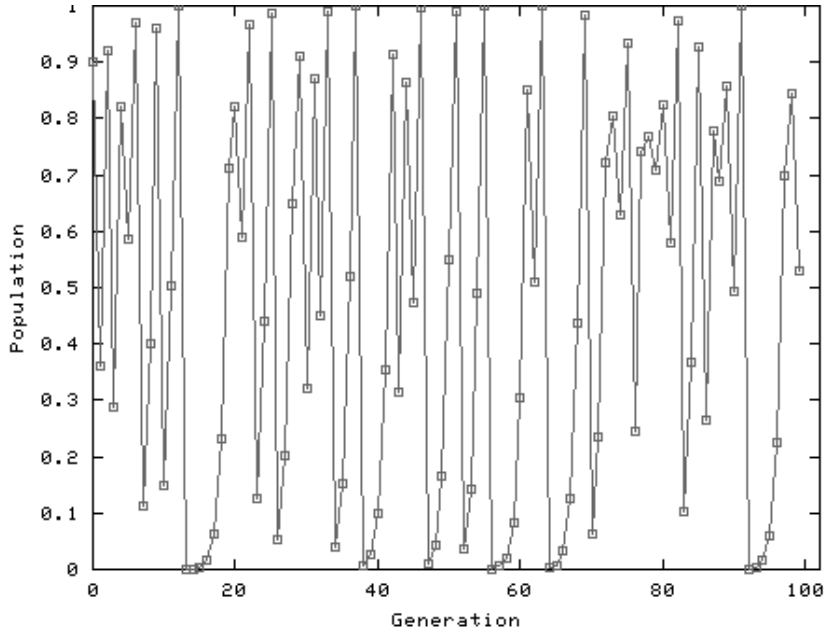
- All initial conditions are pulled toward 0.6.
- (Note that there are different vertical scales on the two plots.)
- 0.6 is an attracting fixed point.

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.

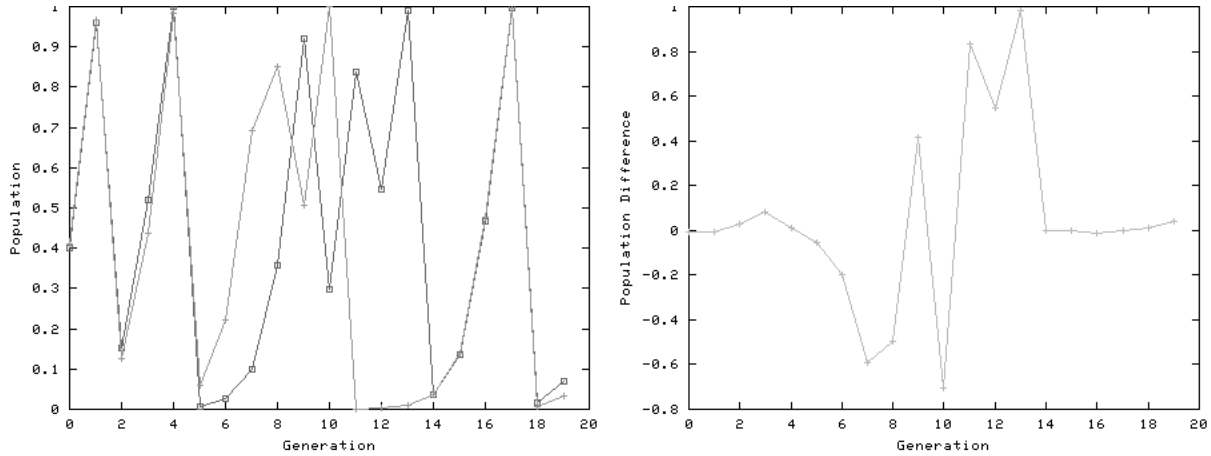
Logistic Equation, $r = 4.0$



- 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 $x_0 = 0.41$.

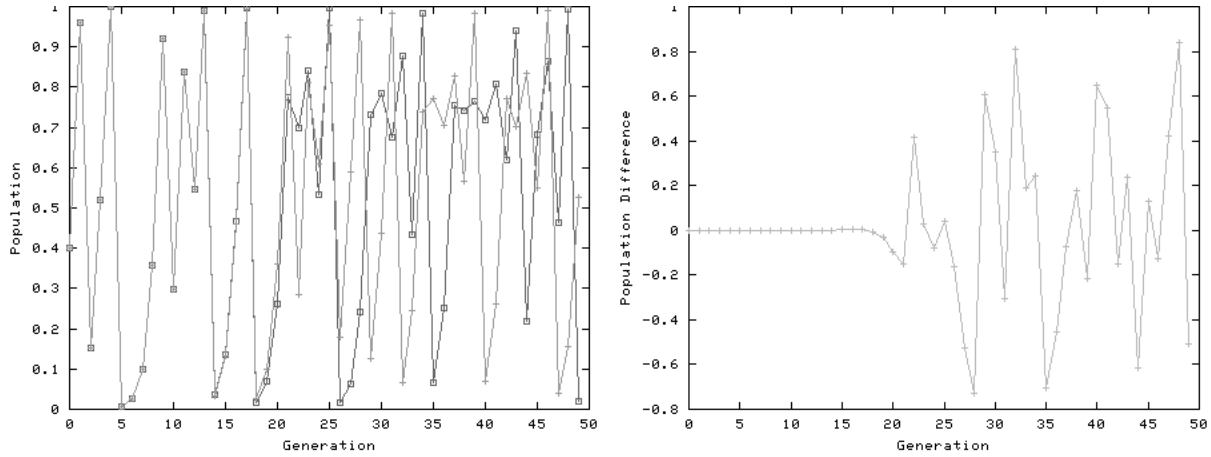


- The right graph plots the difference between the two orbits on the left with slightly different initial conditions.
- Note that the difference between the two orbits grows.
- Can think of one initial condition as the true one, and the other as the measured one.

- 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.

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.
- Feldman. *Chaos and Fractals: An Elementary Introduction*. Oxford. 2011.

© David P. Feldman

Part III

More Chaos: Period Doubling, Universality, Lyapunov Exponents

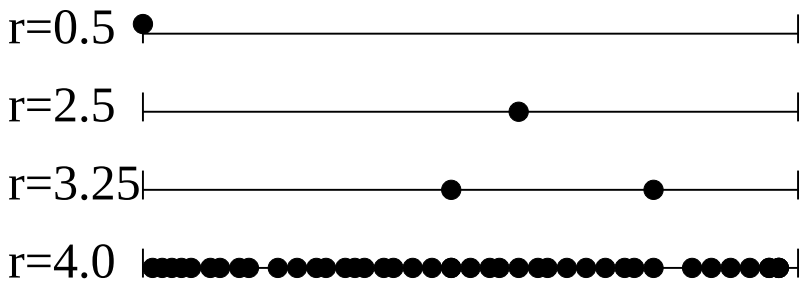
© David P. Feldman

Introduction to Chaos Part II

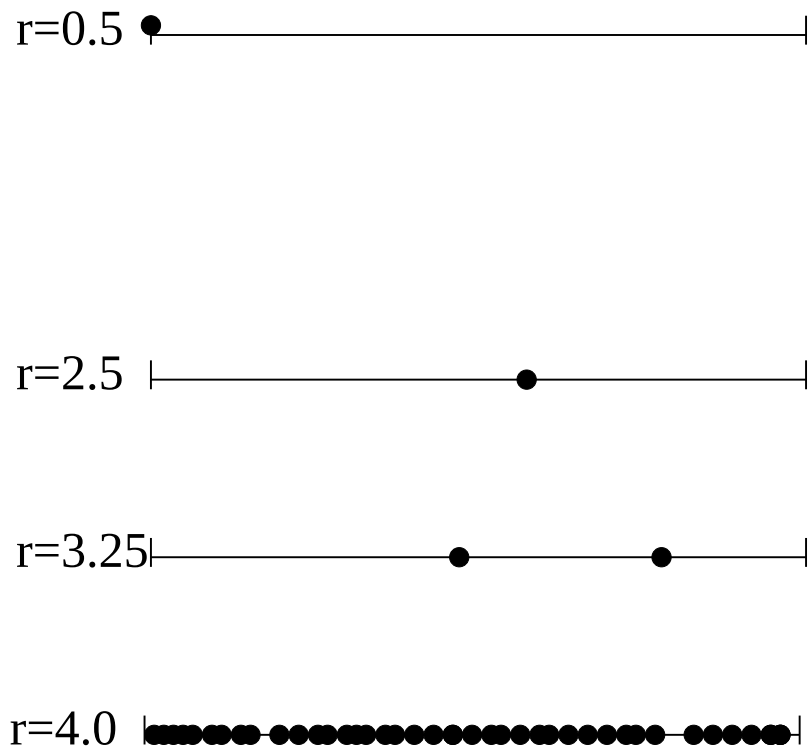
We have seen several possible long-term behaviors for the logistic equation:

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:

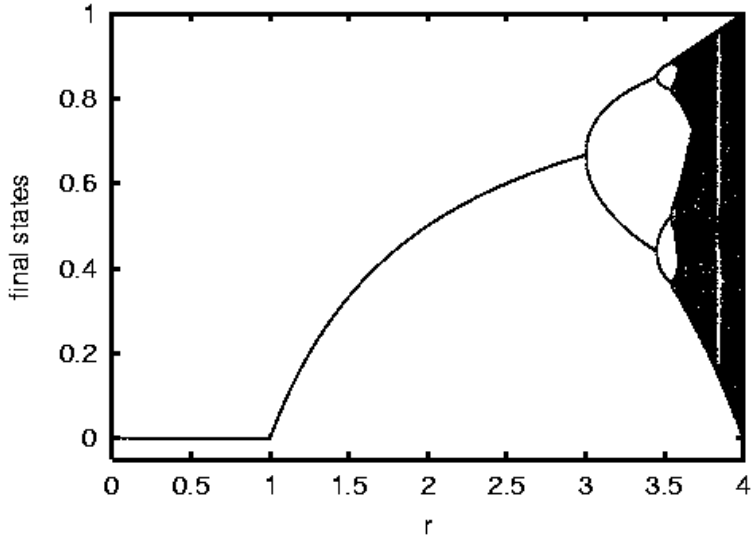


- I.e., for each r, iterate and plot the final x values as dots on the number line.
- What else can the logistic equation do??



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

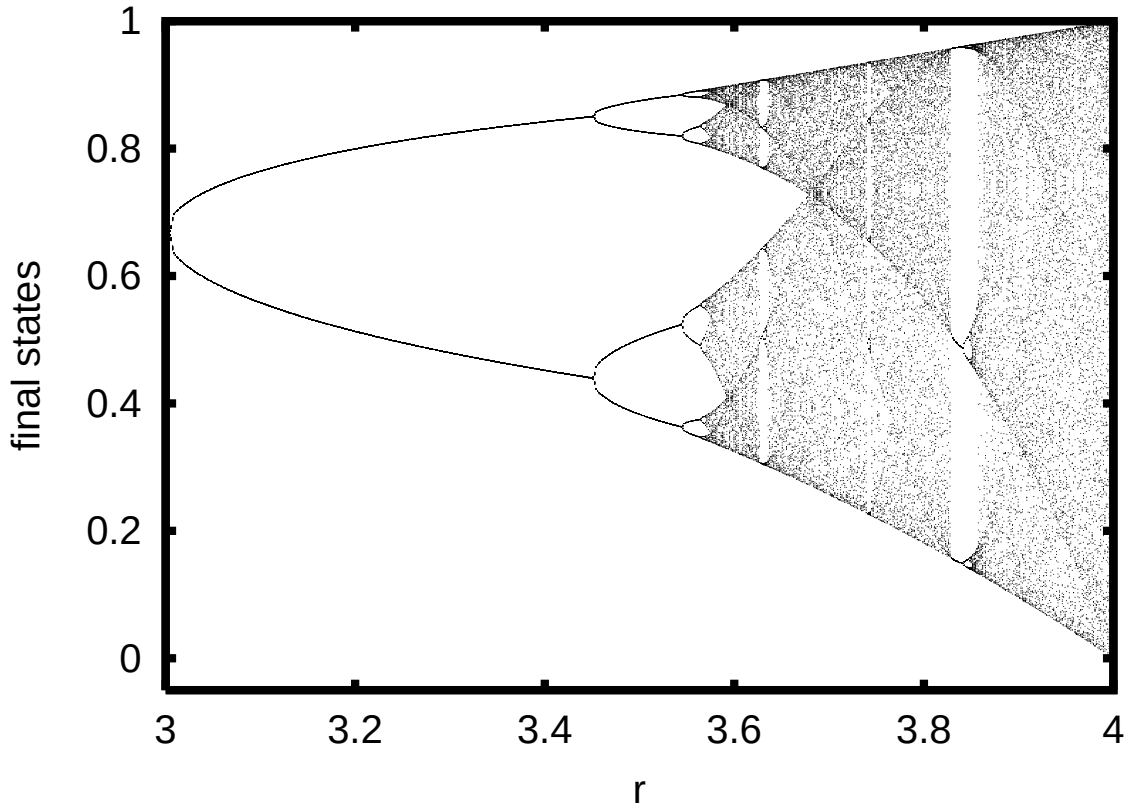
Bifurcation Diagram



- 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

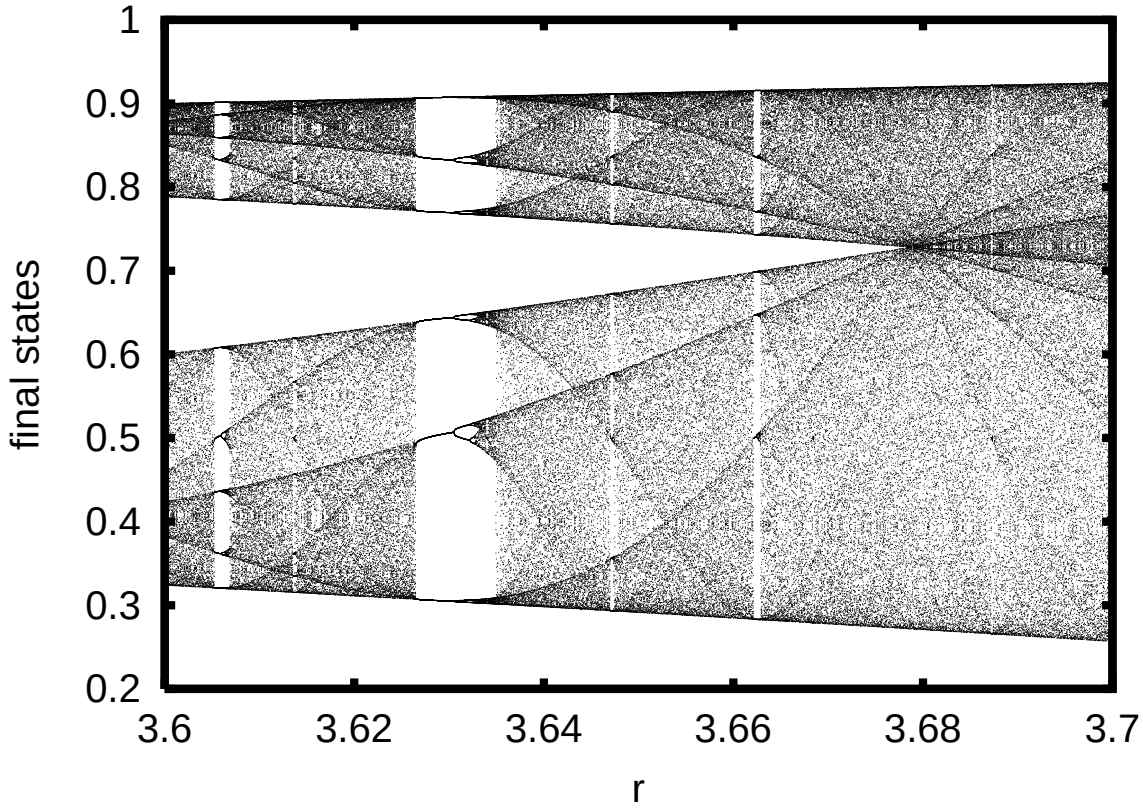
Let's zoom in on a region of the bifurcation diagram:



- 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

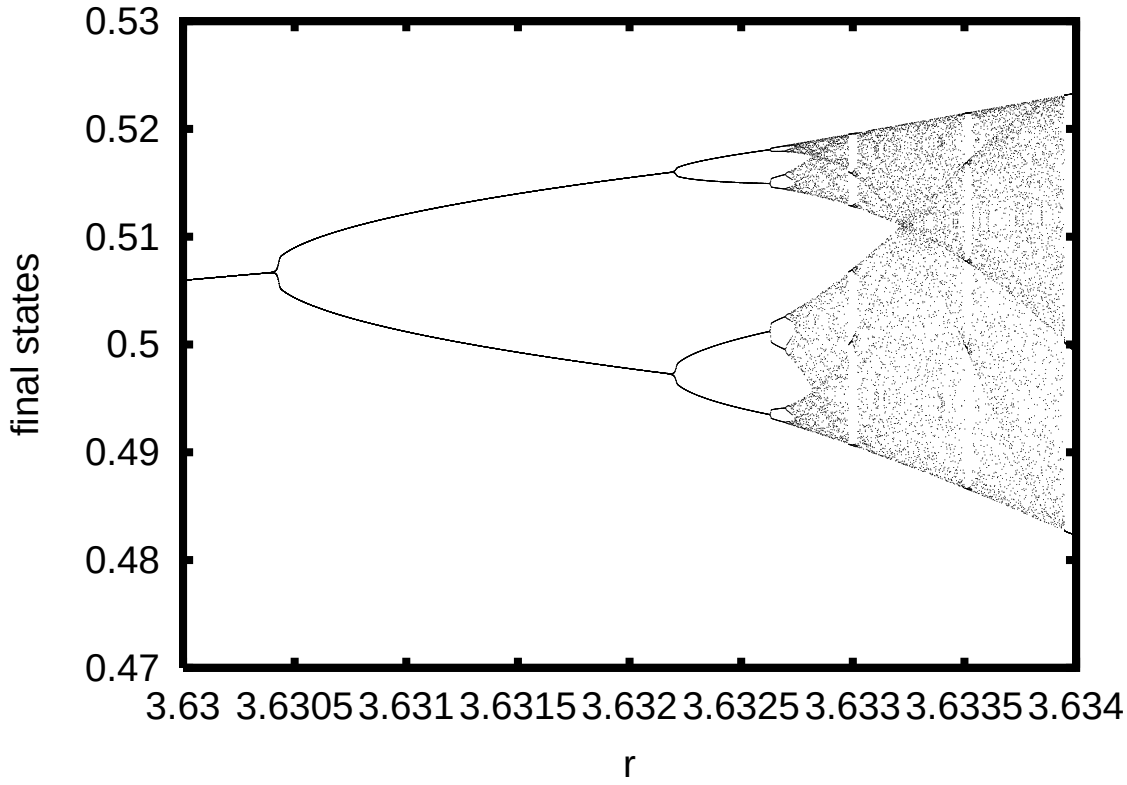
Let's zoom in again:



- Note the sudden changes from chaotic to periodic behavior.

Bifurcation diagram, continued

Let's zoom in once more:



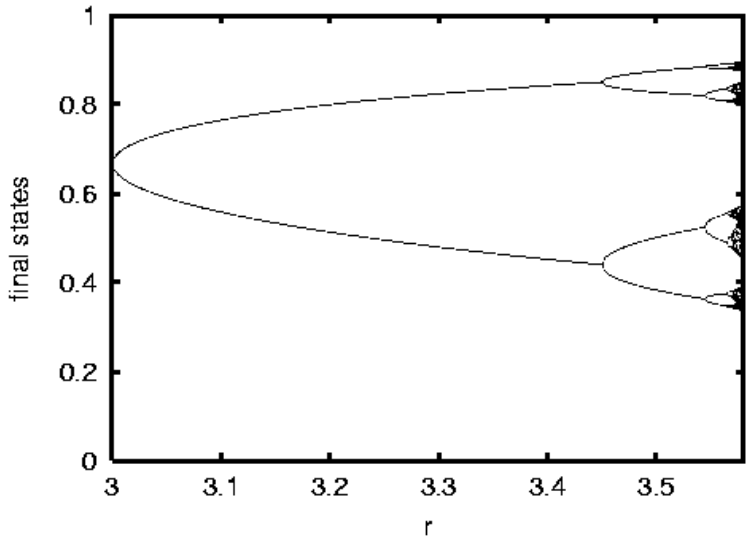
- Note the small scales on the vertical axis, and the tiny scale on the y axis.
- Note the self-similar structure. As we zoom in we keep seeing sideways pitchforks.

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.
- This is an example of a sort of "order within chaos."
- Bifurcations—a sudden, qualitative change in behavior as a parameter is continuously varied—is a generic feature of non-linear systems.
- 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

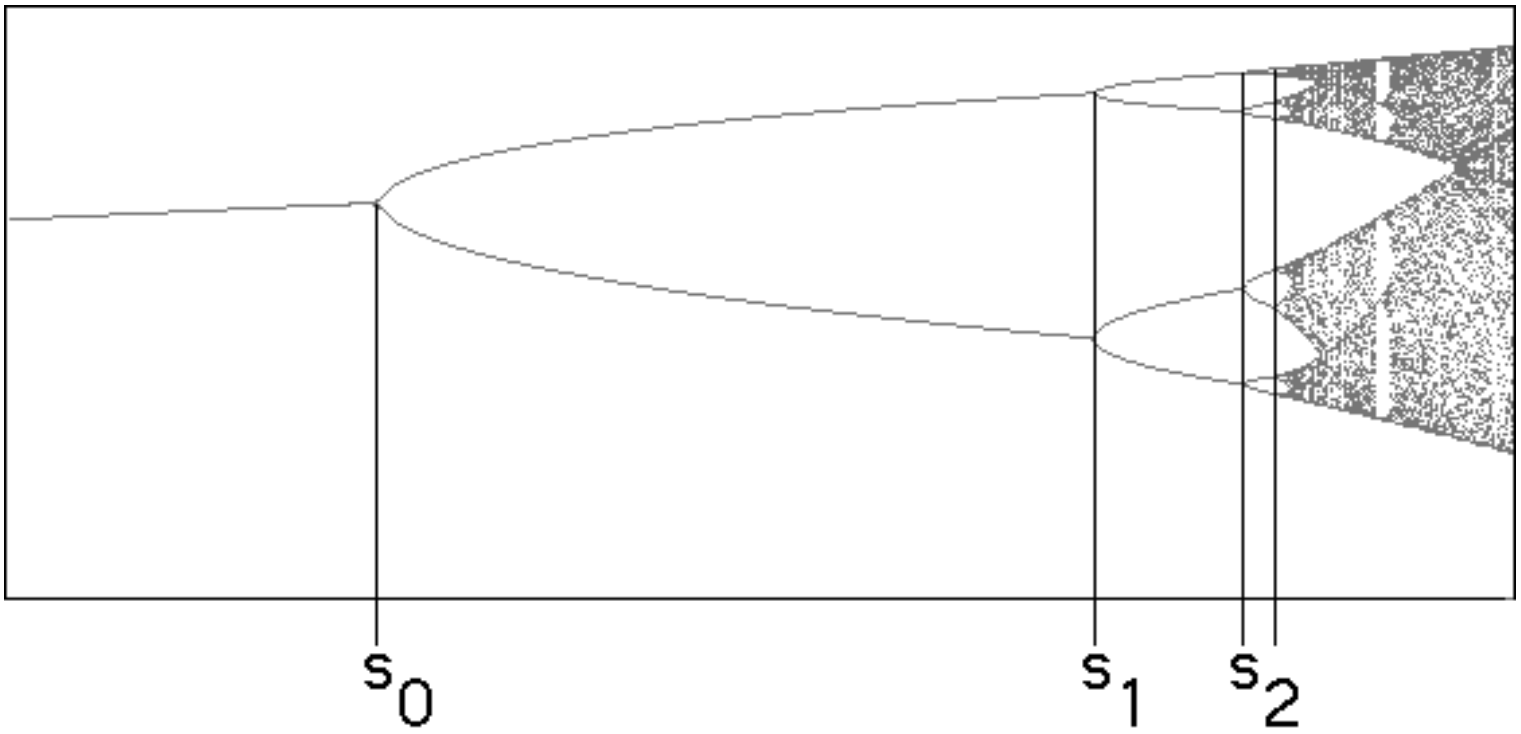
- As r is increased from 3, a sequence of period doubling bifurcations occur.



- 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.

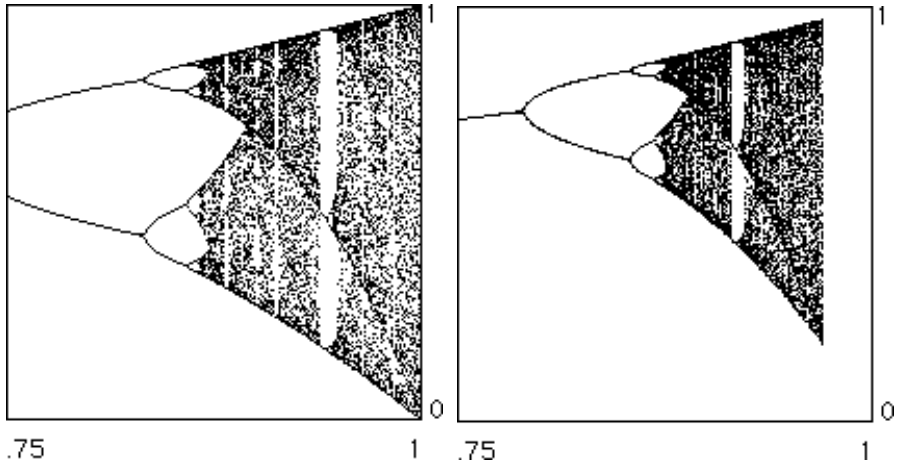


- The first ratio is: $\delta_1 = \frac{s_1 - s_0}{s_2 - s_1}$.
- The n$^{\text{th}}$ ratio is: $\delta_n = \frac{s_n - s_{n-1}}{s_{n+1} - s_n}$.

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 $f(x) = \frac{27}{4}rx^2(1-x)$.
- 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.

Measuring Sensitive Dependence: Lyapunov Exponent

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

- The Lyapunov exponent λ is a measure of the average exponential rate at which near-by trajectories are pulled apart.

$$\text{Difference between trajectories} \approx 2^{\lambda n} \ , \tag{1}$$

  where n is the iterate number.

- If λ is positive, then the system has sensitive dependence on initial conditions.
- The larger the value of λ, the greater the sensitivity.
- The Lyapunov exponent is a widely used and well understood way to characterize different chaotic systems.

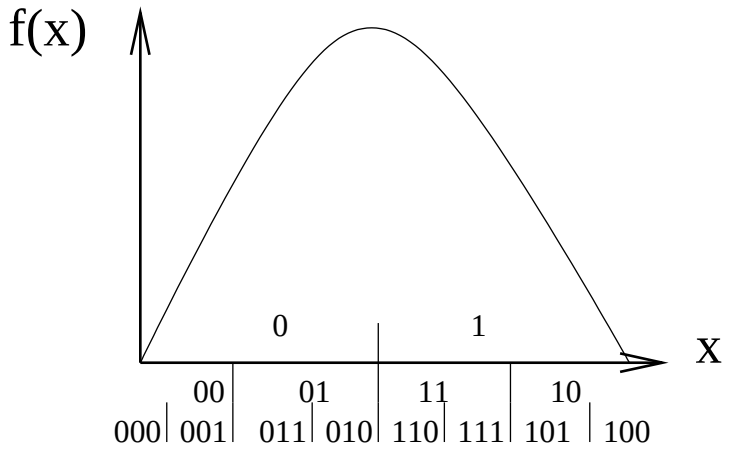
Symbolic Dynamics

- 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.
- Chaos and other structures can be stable.
- Aubin and Dahan Dalmedico: [C]haos has definitely blurred a number of old epistemological boundaries and conceptual oppositions hitherto seemingly irreducible such as order/disorder, random/nonrandom, simple/complex, local/global, stable/unstable, and microscopic/macroscopic.

  Aubin and Dahan Dalmedico, *Historia Mathematica* 29 (2002), 167. doi:10.1006/hmat.2002.2351

Chaos $\Rightarrow$ 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?
- What is emergence?

Part IV

Information Theory: Motivation, Basic Definitions, Noiseless Coding Theorem

© David P. Feldman

Structure?

- Complex Systems is concerned with order, pattern, structure, regularity, emergence.
- How can we quantify or make sense of these notions?
- Useful tools can be found from a number of fields, including information theory and computation theory.
- In this portion of my presentation I'll survey some of these tools and show how they can be adapted for use in complex systems.

© David P. Feldman

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:

$$H[X] \equiv -\sum_{x \in \mathcal{X}} \Pr(x) \log_2(\Pr(x)) \ . \tag{2}$$

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:

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

- 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

© David P. Feldman

Information Theory: Part II

Applications to Stochastic Processes

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

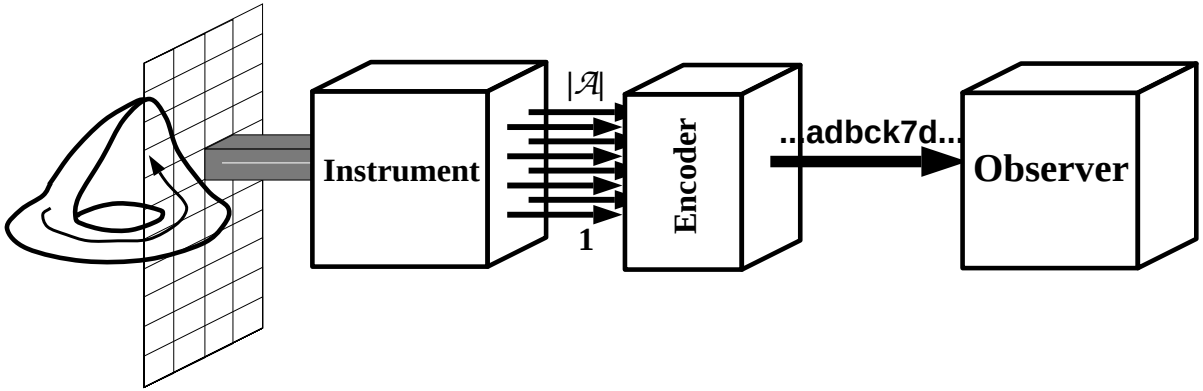
$$\cdots 0011001001010110100110011101011 0 \cdots$$

- 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"): $\overrightarrow{s} \equiv S_0 \, S_1 \, S_2 \cdots$

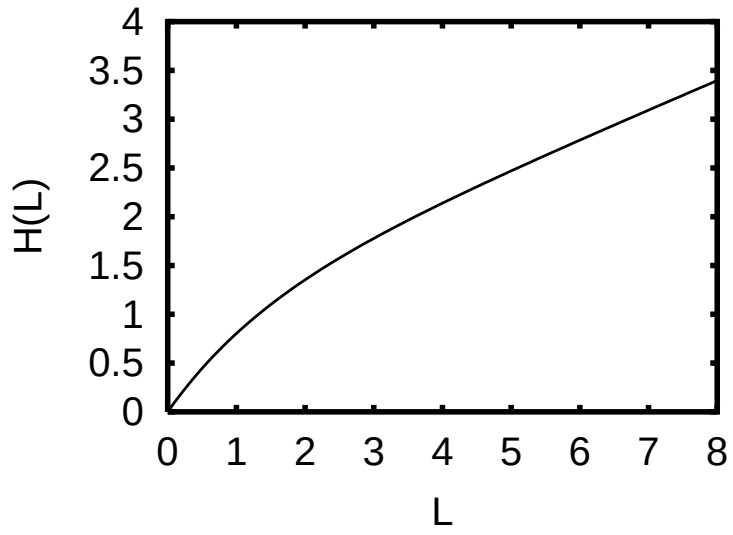
$$\cdots 1101010010110101010010010010 \cdots$$

Entropy Growth

- Entropy of L-block:

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

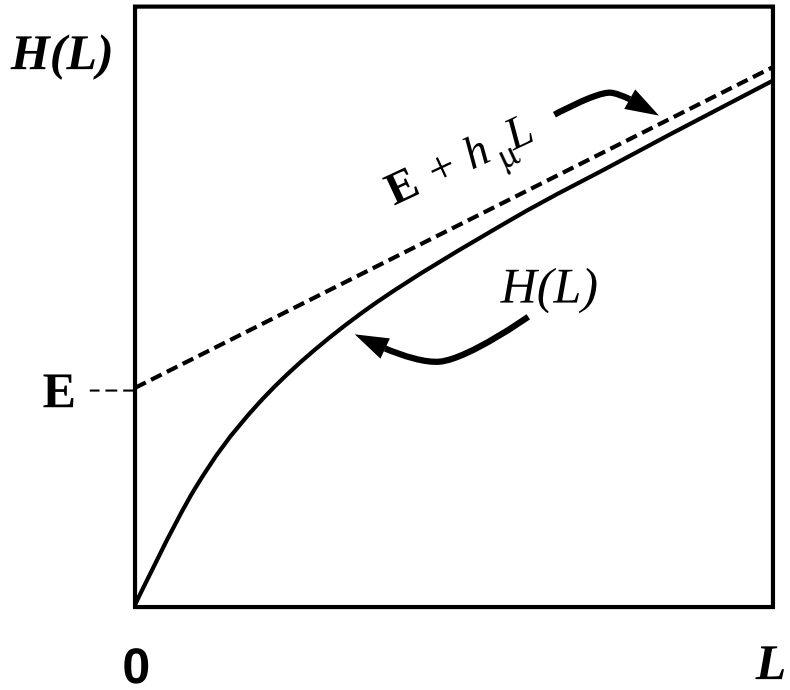
- $H(L) =$ average uncertainty about the outcome of L consecutive variables.



- $H(L)$ 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:*

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

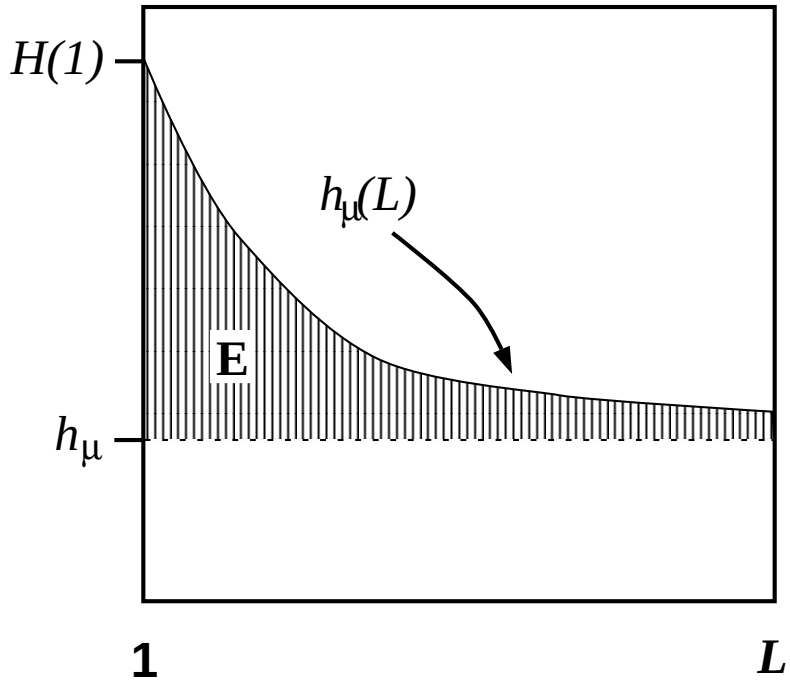
- $h_\mu(L) = H[S_L | S_1 S_1 \ldots S_{L-1}]$
- I.e., $h_\mu(L)$ is the average uncertainty of the next symbol, given that the previous L symbols have been observed.

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.

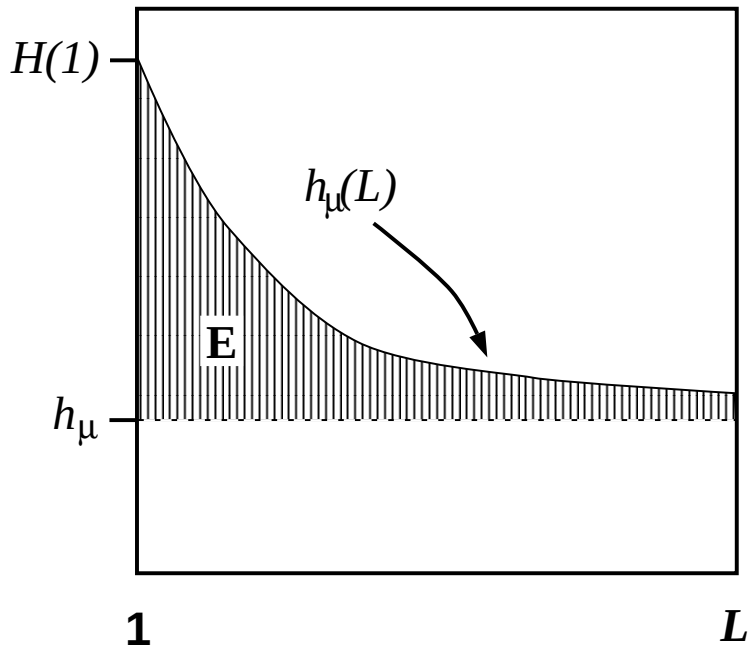
How does $h_\mu(L)$ approach h_μ?

- 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":

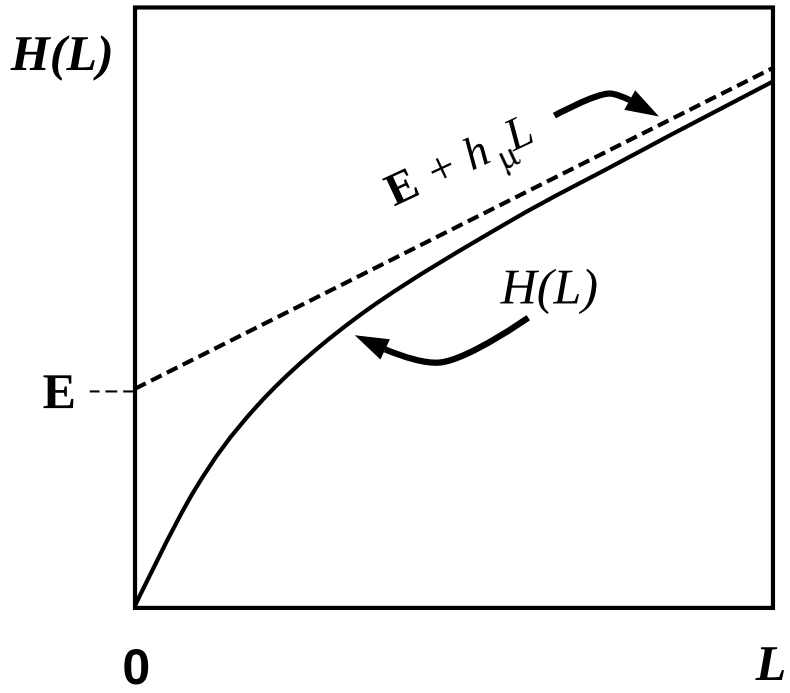
$$\mathbf{E} = I(\overleftarrow{S}; \overrightarrow{S}) \equiv \sum_{\{\overleftrightarrow{s}\}} \Pr(\overleftrightarrow{s}) \log_2 \left[\frac{\Pr(\overleftrightarrow{s})}{\Pr(\overleftarrow{s})\Pr(\overrightarrow{s})} \right] .$$

- $\mathbf{E}$ 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

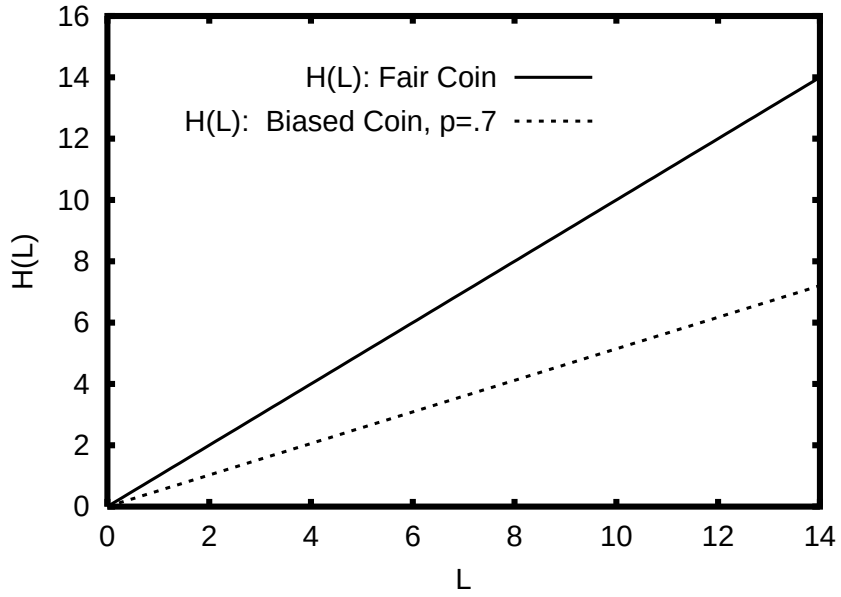
- $\mathbf{E}$ is the y-intercept of the straight line to which $H(L)$ asymptotes.
- $\mathbf{E} = \lim_{L\to\infty} \left[H(L) - h_\mu L\right]$.



Excess Entropy Summary

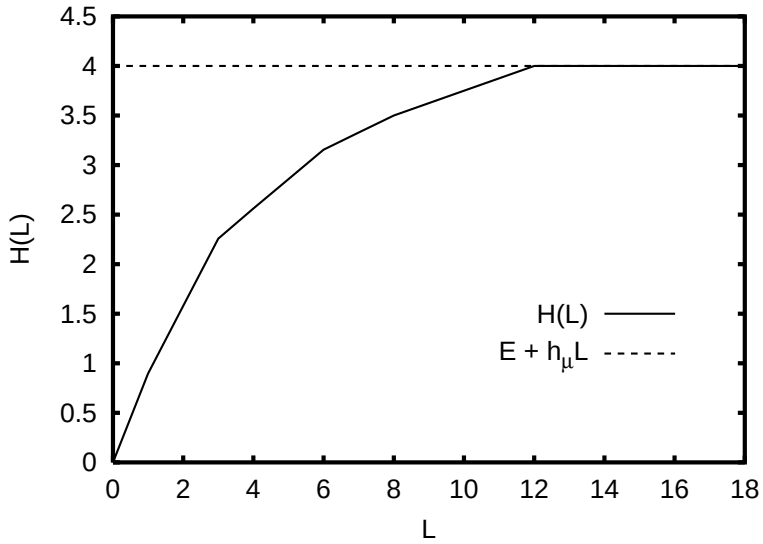
- Is a structural property of the system — measures a feature complementary to entropy.
- Measures memory or spatial structure.
- Lower bound for statistical complexity, minimum amount of information needed for minimal stochastic model of system

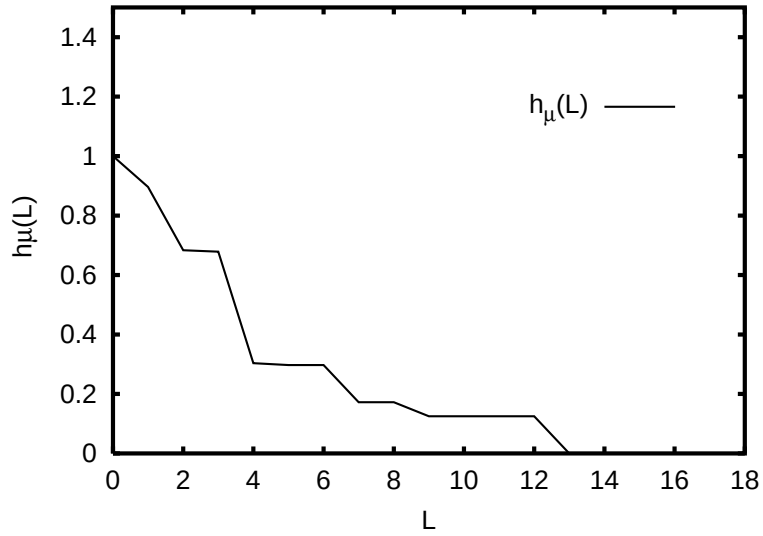
Example I: Fair Coin



- 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





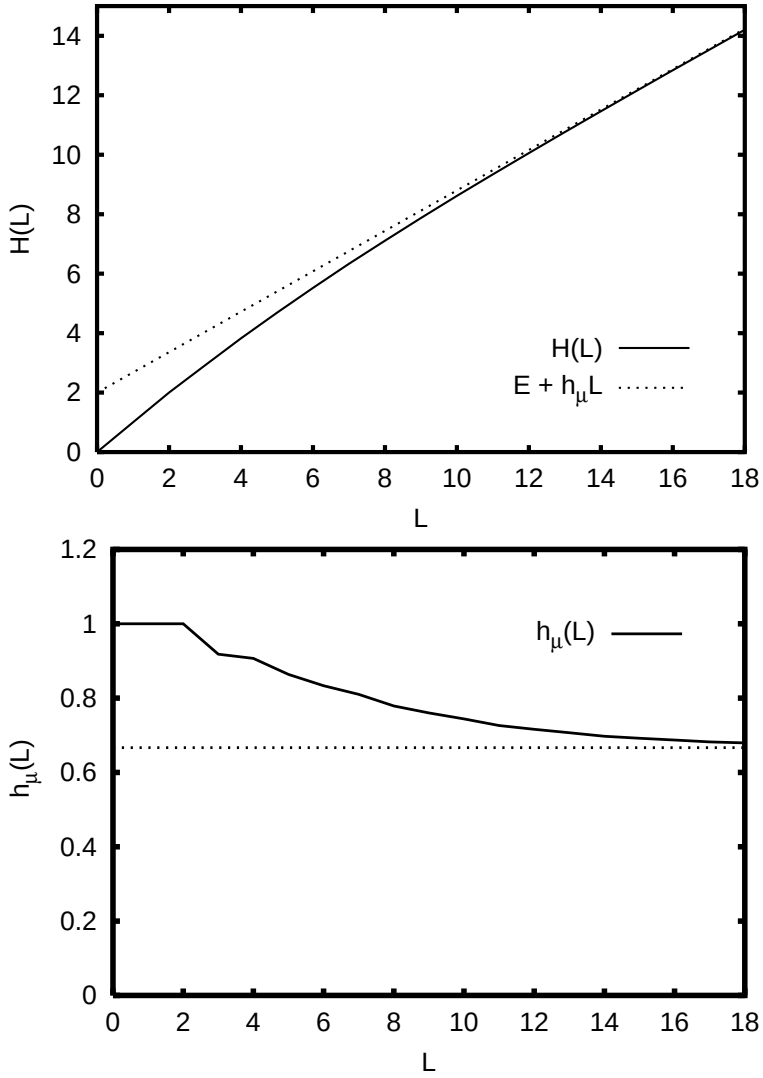
- Sequence: . . . 1010111011101110 . . .

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

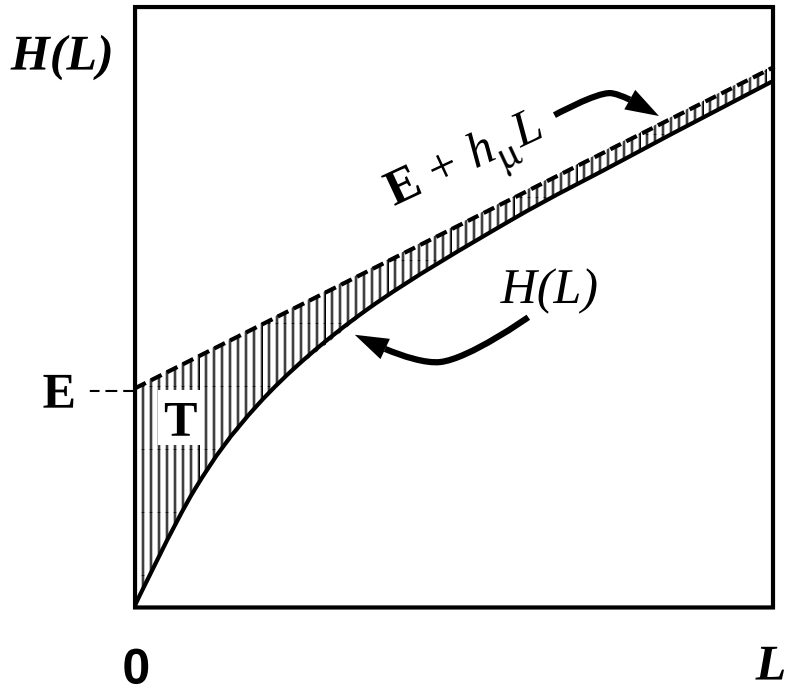
- Crutchfield and Packard, *Intl. J. Theo. Phys*, 21:433-466. (1982); *Physica D*, 7:201-223, 1983. [Dynamical systems]
- Shaw, “The Dripping Faucet ..., ” Aerial Press, 1984. [A dripping faucet]
- Grassberger, *Intl. J. Theo. Phys*, 25:907-938, 1986. [Cellular automata (CAs), dynamical systems]
- Szépfalusy 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 $\mathbf{T}$.
- $\mathbf{T}$ 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.

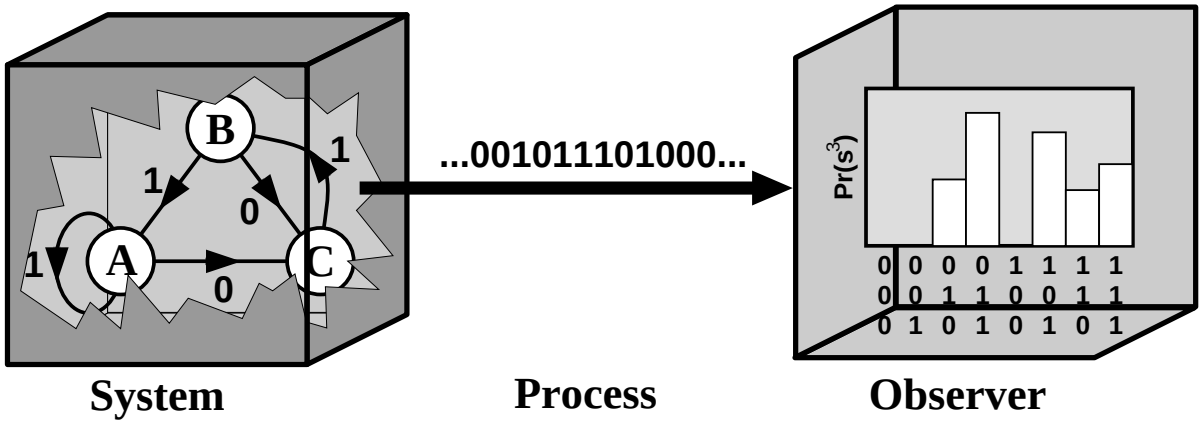
   More on this in Part IX.

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.



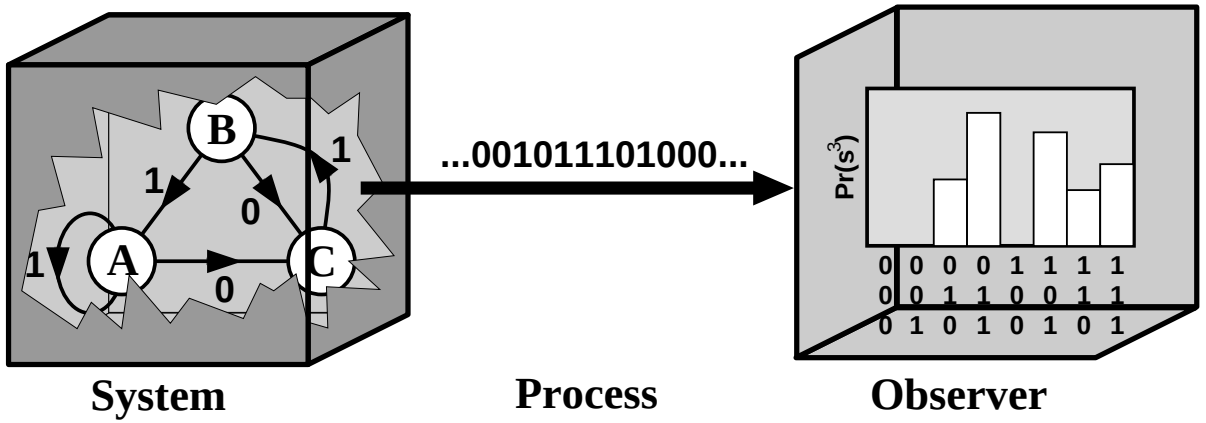
- One simply forms histograms of occurrences of particular sequences and uses these to estimate $\Pr(s^L)$, from which $\mathbf{E}$ and h_μ may be readily calculated.

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

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

© David P. Feldman

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

© David P. Feldman

1101101001

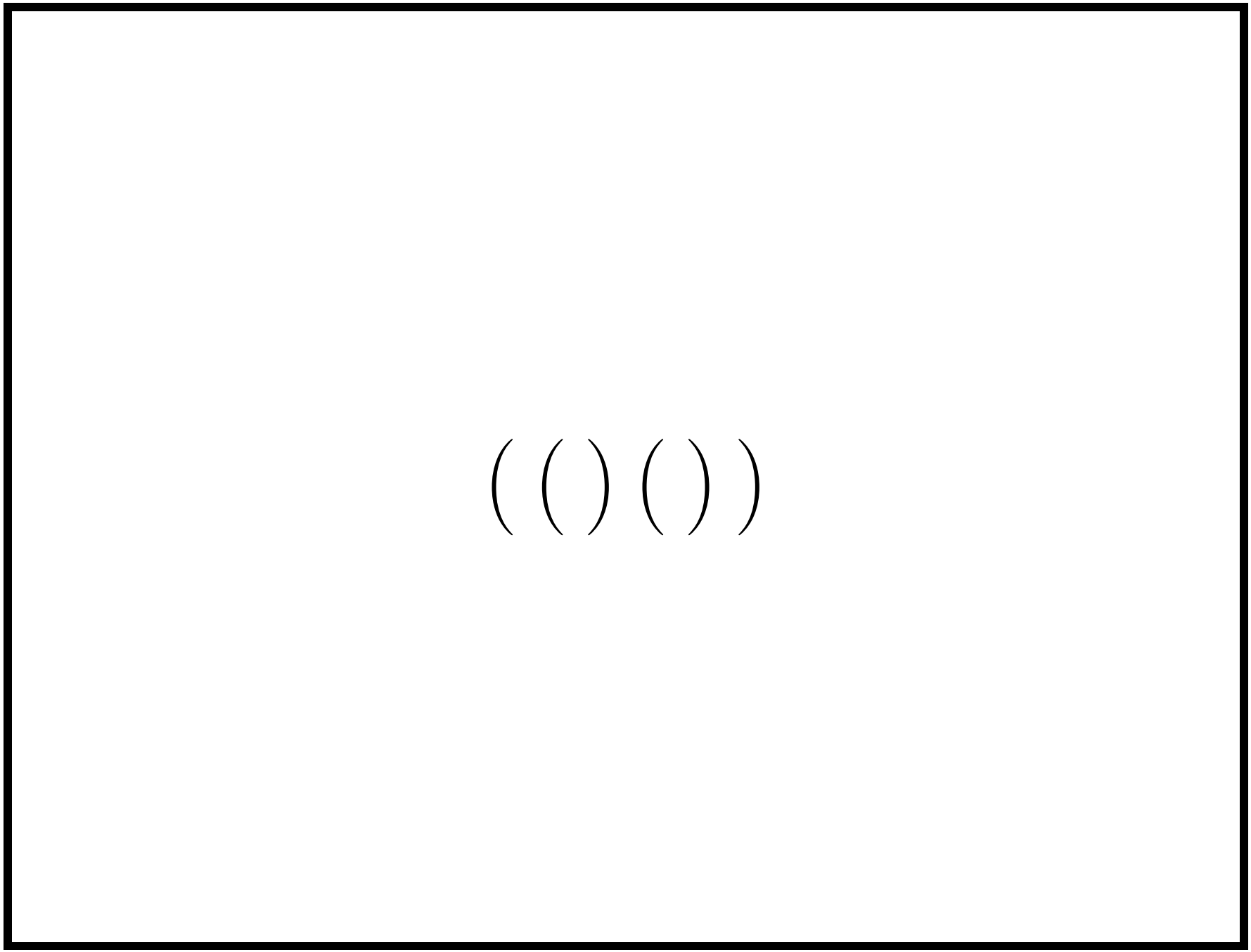
© David P. Feldman

110110101011

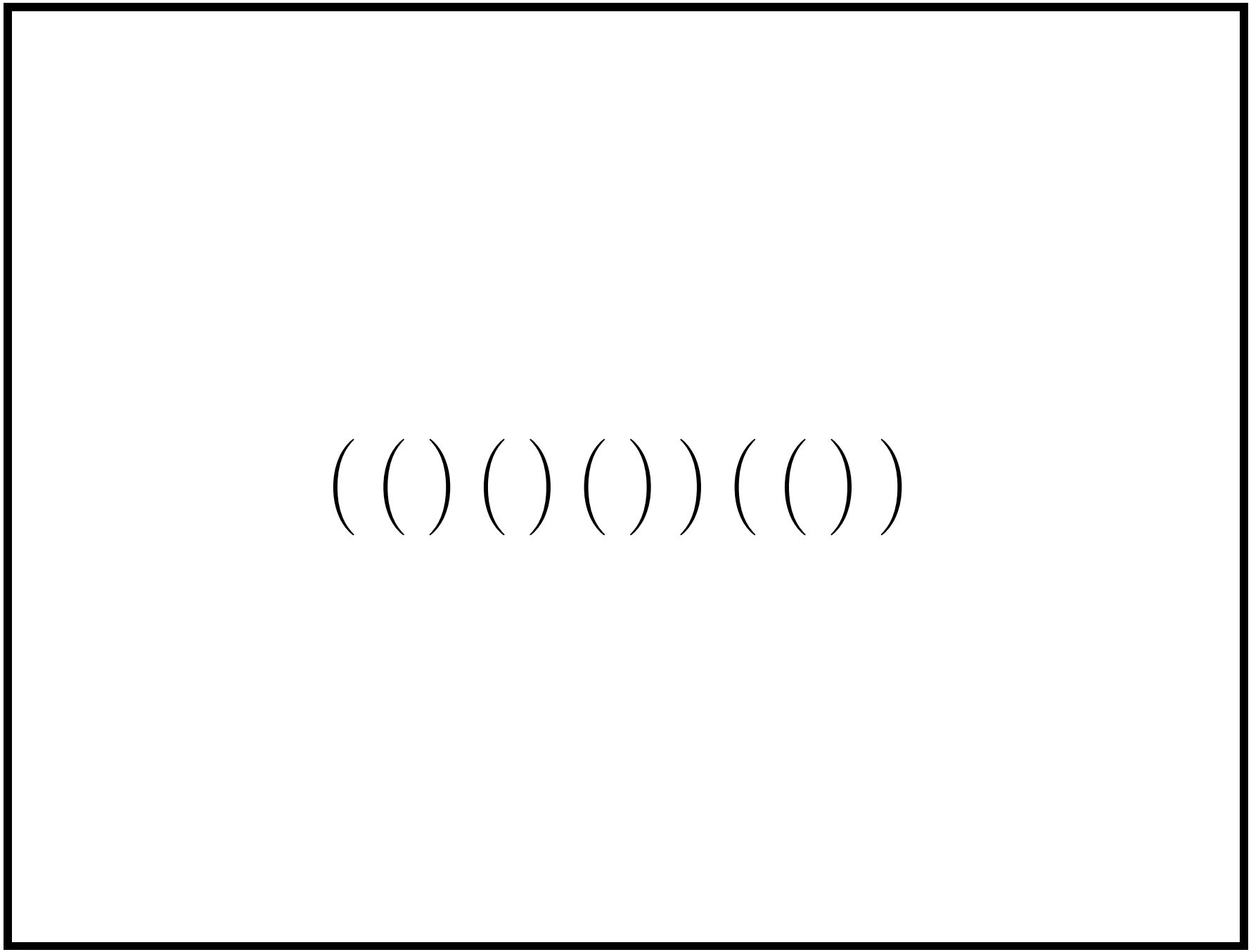
© David P. Feldman

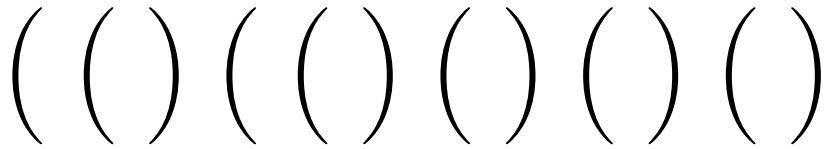
Example 2:

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



© David P. Feldman





© David P. Feldman

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!

110001100001

© David P. Feldman

11000011

© David P. Feldman

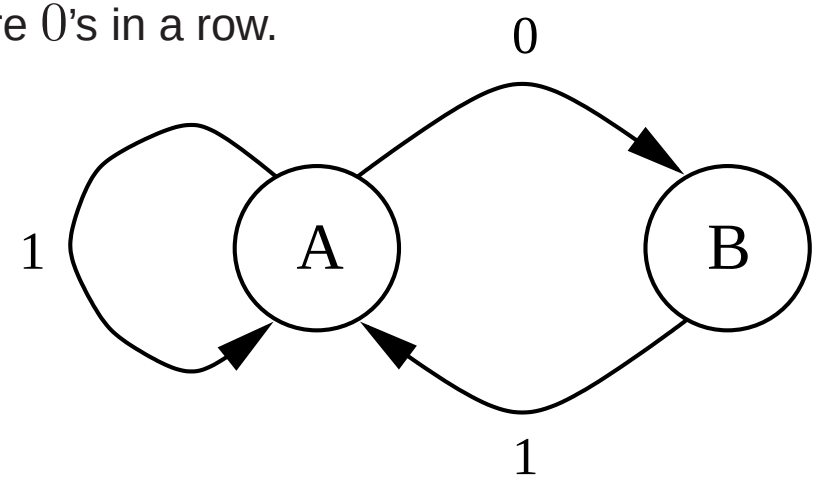
$$11\overbrace{00000000000}^{11\text{ elements}}11$$

© David P. Feldman

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

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:

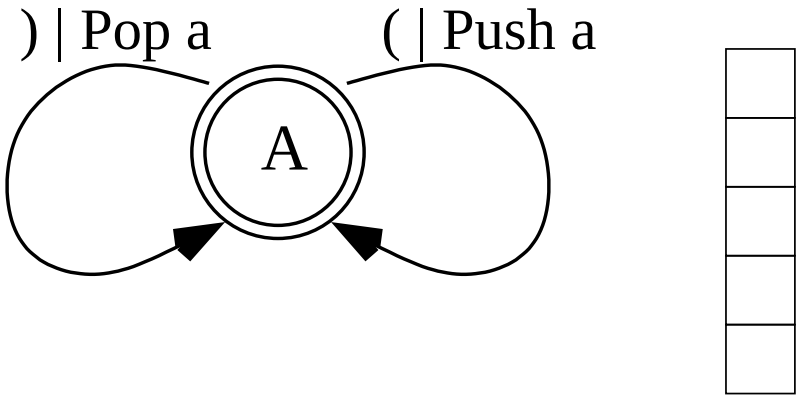
- A regular expression is a **rule**.
- A regular language is a **set**.
- A FSM is a **machine**.

Regular languages $\leftrightarrow$ 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

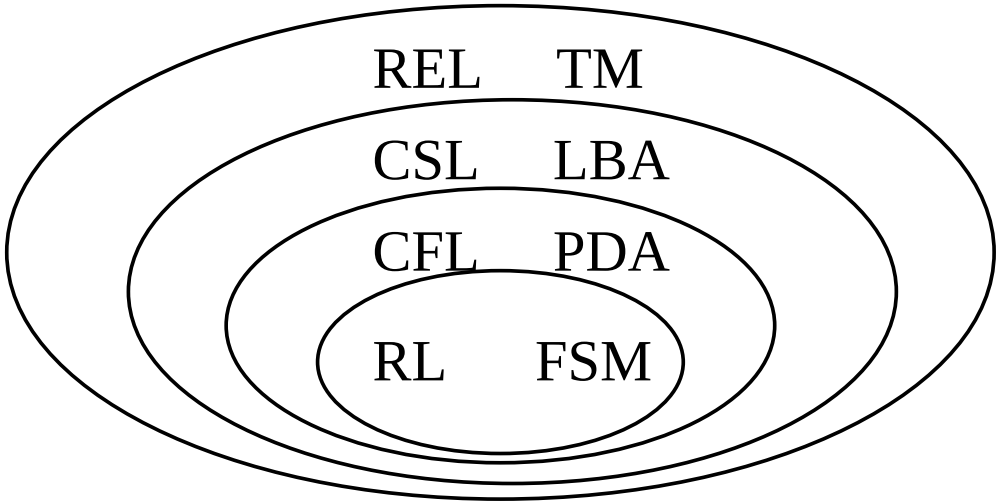
- $(,)$ are **terminals**, symbols in the alphabet $\mathcal{A}$.
- W, V are **variables**, symbols not in $\mathcal{A}$, to be eventually replaced by terminals.
- CFL's are context free in the sense that the production rule depends only on the variable, not on where the variable is in the string.

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.

© David P. Feldman

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.

© David P. Feldman

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 VII

An Introduction to Computational Mechanics

© David P. Feldman

An Introduction to Computational Mechanics

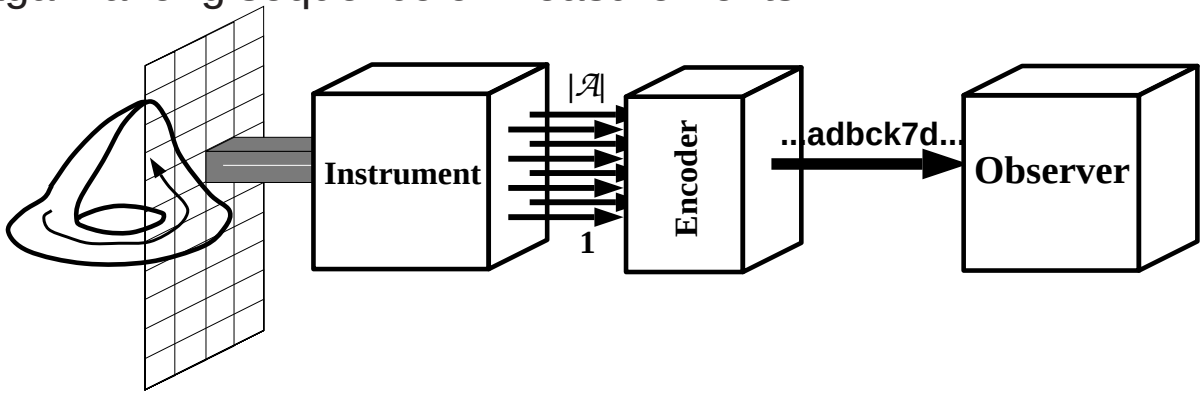
1. Computational Mechanics provides another way of measuring an object's complexity or regularities.
2. Unlike the excess entropy, computational mechanics makes use of the the models of formal computation to provide a direct, structural accounting of a system's intrinsic information processing.
3. Computational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

$$\ldots 10111110101110111010111 \ldots$$

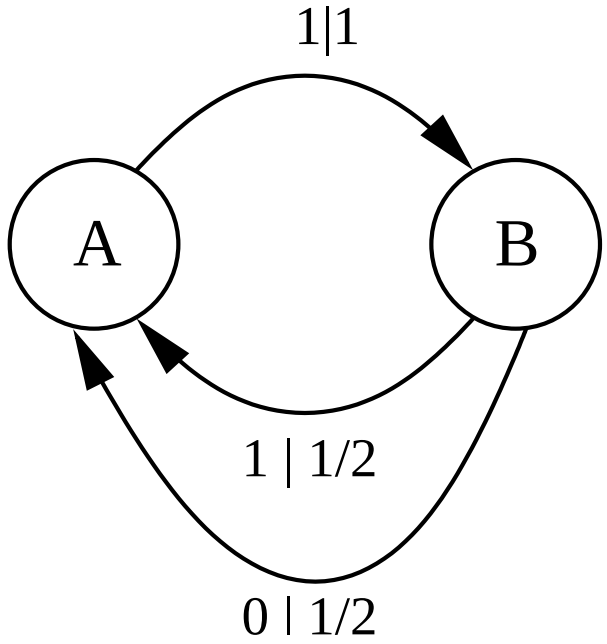
Discovery!

$$\ldots 10111110101110111010111 \ldots$$

- 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:



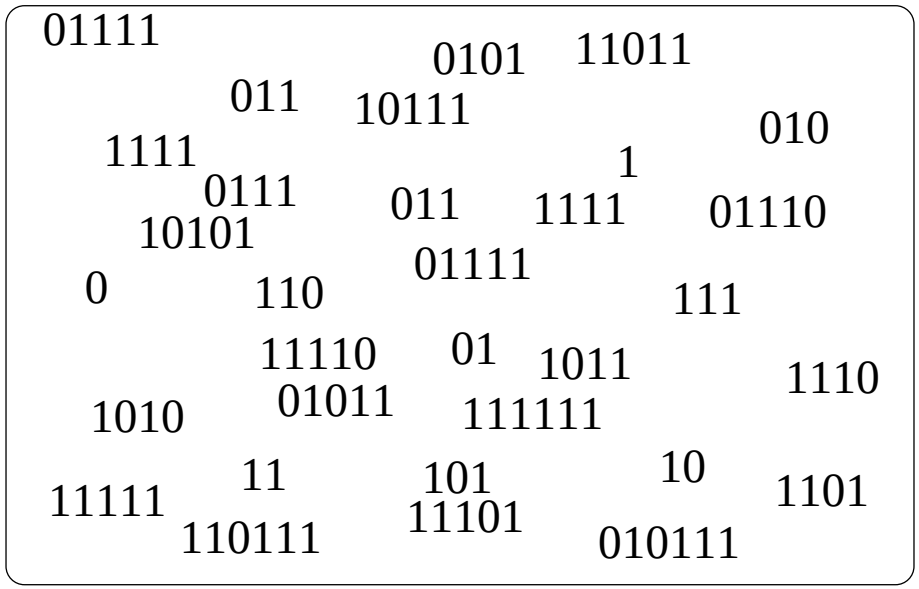
- From state $\mathbf{A}$, one sees a 1 with probability 1.
- From sate $\mathbf{B}$, one sees a 1 with probability $1/2$, and a 0 with probability $1/2$.
- This is a stochastic generalization of a finite state machine.
- Note that it is still *deterministic* in the sense that the output symbol (0 or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

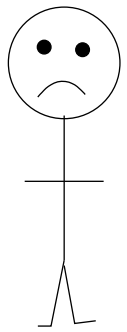
What do you Need to Remember in Order to Predict?

Space of all possible pasts.



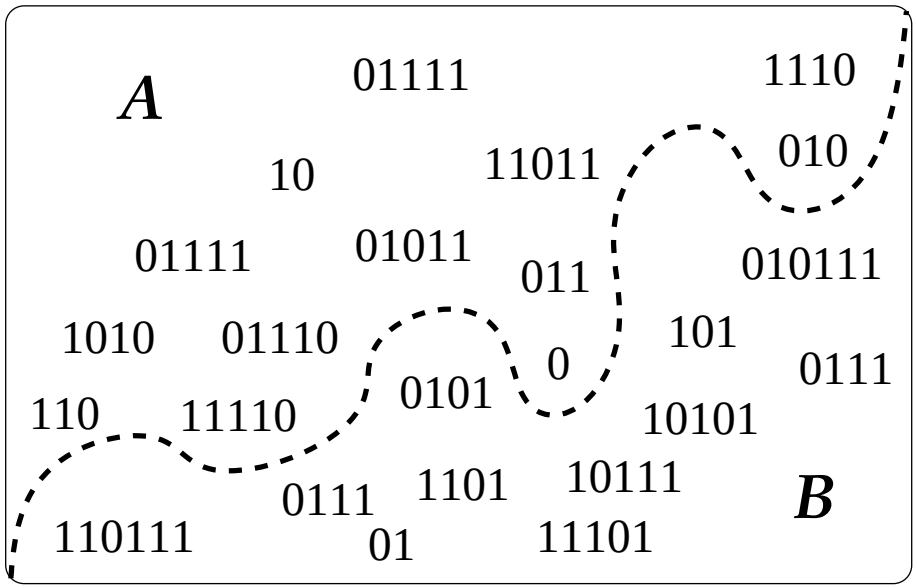
Do I really have to remember all this??

My memory isn't good enough.



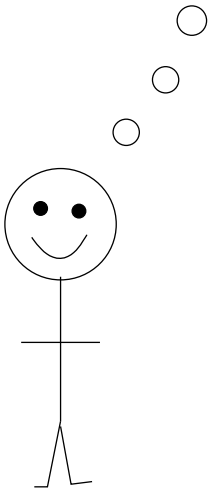
One Only Needs to Remember the Causal States.

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, $\Pr(\overrightarrow{S} \mid 0) = \Pr(\overrightarrow{S} \mid 1011)$. 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

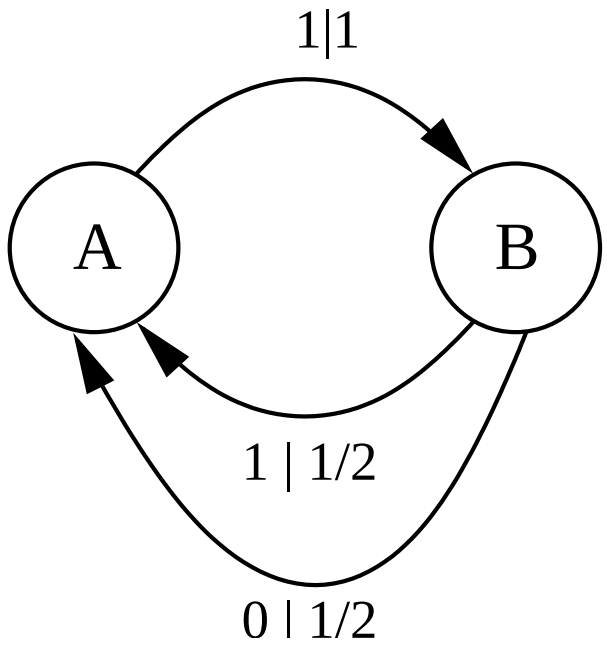
- The causal states together with the probability of transitions between causal states are an **ϵ-machine**, a minimal model capable of statistically reproducing the original configuration.
- The ϵ-machine tells us **how the system computes**.
- The "ϵ" reminds us that the measurement symbols upon which 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:

$$T^{(s)}_{\alpha\beta} \;=\; \Pr(\sigma_\beta, s|\sigma_\alpha) \;, \tag{3}$$

or

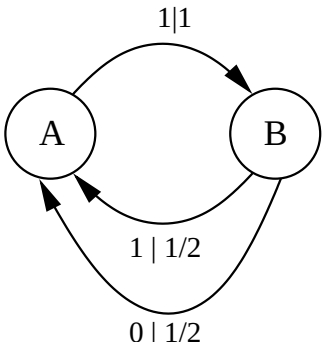
$$T^{(s)}_{\alpha\beta} \;=\; \alpha \stackrel{s}{\longrightarrow} \beta \;. \tag{4}$$

Initial Example Again



$$T^{(s=0)} = \begin{pmatrix} 0 & 0 \\ \frac{1}{2} & 0 \end{pmatrix} , \quad T^{(s=1)} = \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & 0 \end{pmatrix} . \tag{5}$$

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:

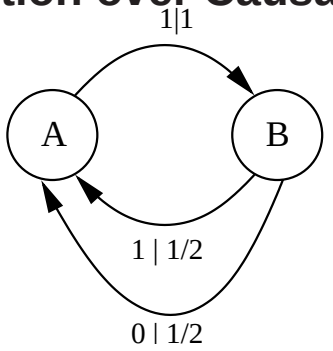
$$T_{\alpha\beta} = \sum_s T^{(s)}_{\alpha\beta} \ . \tag{6}$$

- For our example:

$$T = T^{(s=0)} + T^{(s=1)} \ , \tag{7}$$

$$T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \frac{1}{2} & 0 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & 0 \end{pmatrix} \ . \tag{8}$$

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:

$$pT = p \ . \tag{9}$$

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

$$p = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} \ . \tag{10}$$

- 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:

$$C_\mu \equiv -\sum_\alpha p_\alpha \log_2 p_\alpha \ . \tag{11}$$

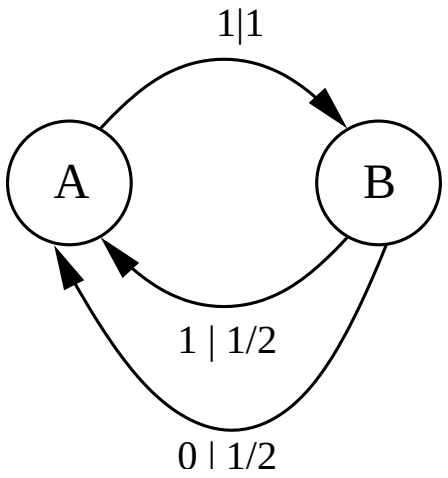
- To perform optimal prediction of the system one needs only to remember the causal states.
- The statistical complexity thus measures the minimum amount of memory needed to perform optimal prediction.
- The statistical complexity is a measure of the pattern or structure or regularity present in the 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:

$$h_\mu = -\sum_\alpha p_\alpha \sum_s \Pr(s|\sigma_\alpha) \log_2 \Pr(s|\sigma_\alpha) \ . \qquad (12)$$

- 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*:

$$I[\overrightarrow{S}; \overleftarrow{S}] = I[\overrightarrow{S}; \sigma] \ . \tag{13}$$

  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.

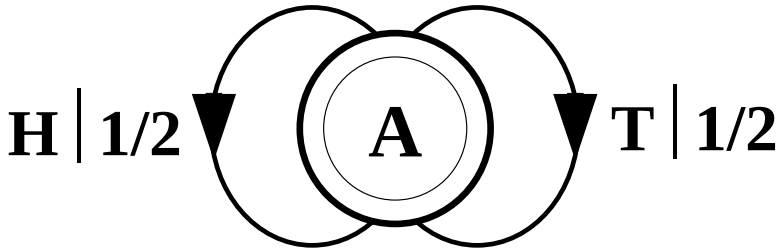
$$C_\mu \geq \mathbf{E} \; . \tag{14}$$

$$\text{Memory needed for model} \geq \text{Memory of the process itself} \; . \tag{15}$$

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

Example I

Fair Coin:

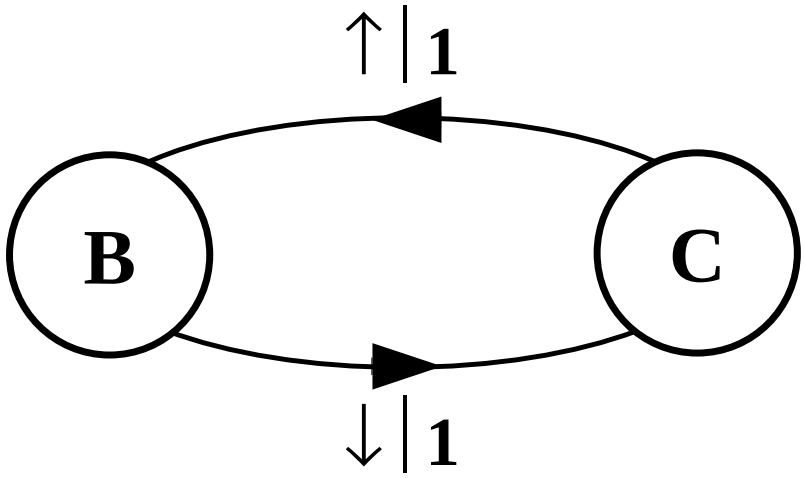


$$\cdots \mathrm{HHTHTHTTTHTHTHTTHTHH} \cdots$$

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

Example II

Period 2 Pattern:

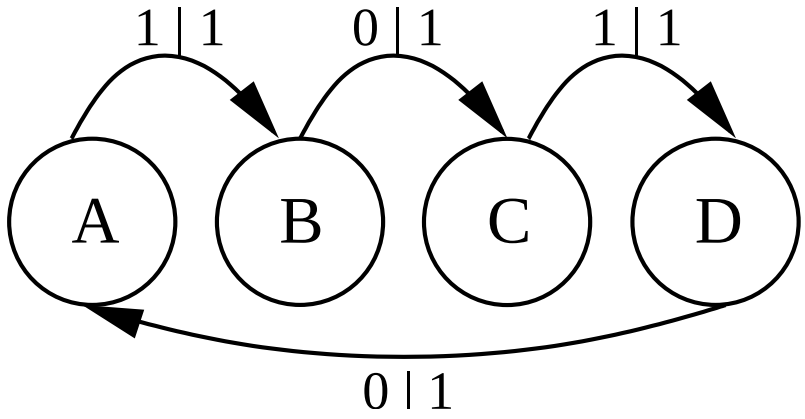


$$\cdots \uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow\uparrow\downarrow \cdots$$

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?

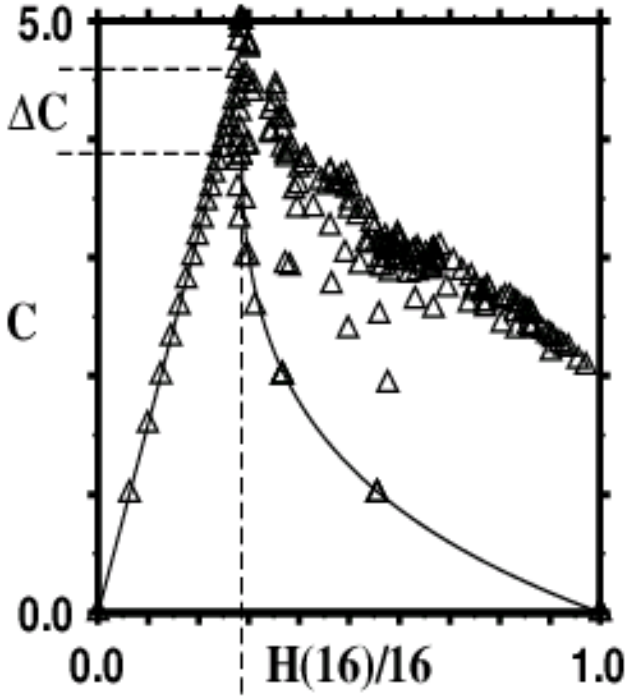
Summary:

- 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.

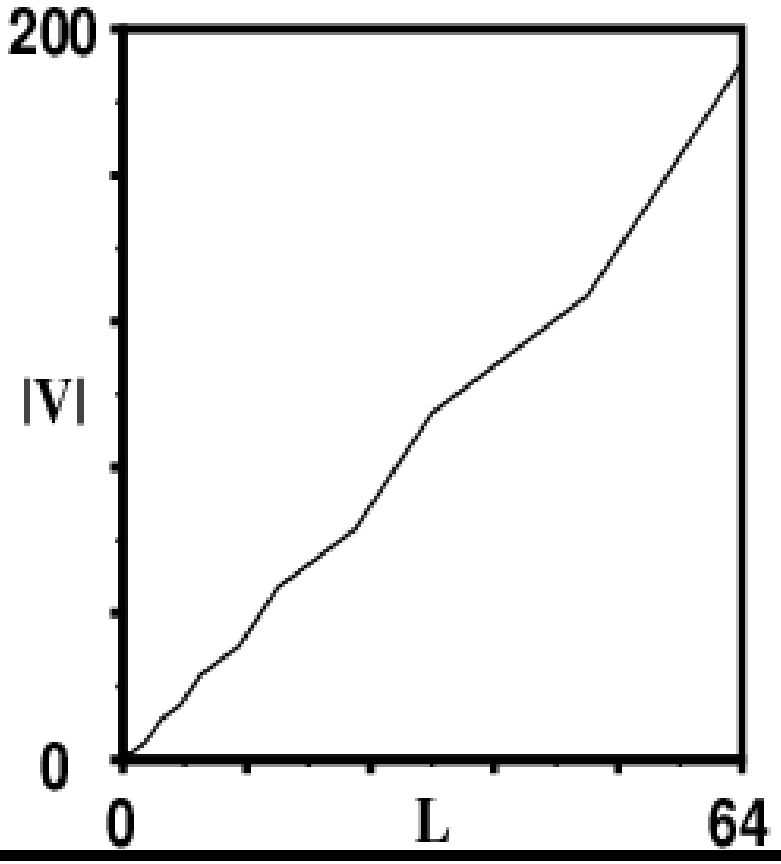


- The linear region on the left corresponds to periodic behavior.
- To the right of the linear region, the behavior is chaotic.

© David P. Feldman

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.

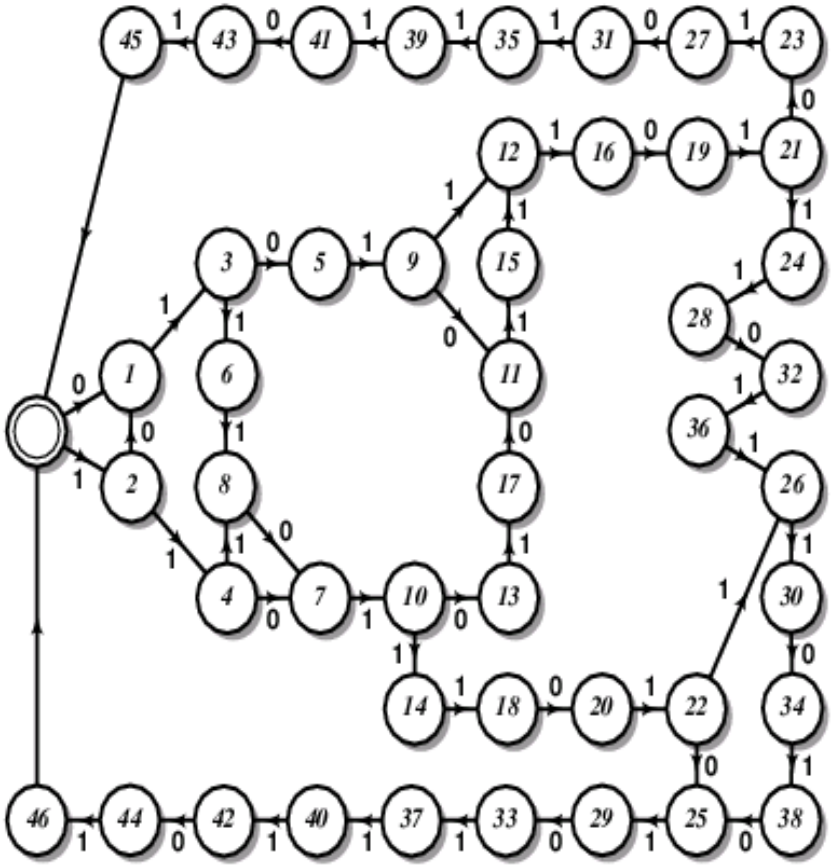


- This suggests that there is no longer a finite representation at the lowest level of the Chomsky Hierarchy.
- The nature of the divergence leads one to a higher-level computational model for the system.

© David P. Feldman

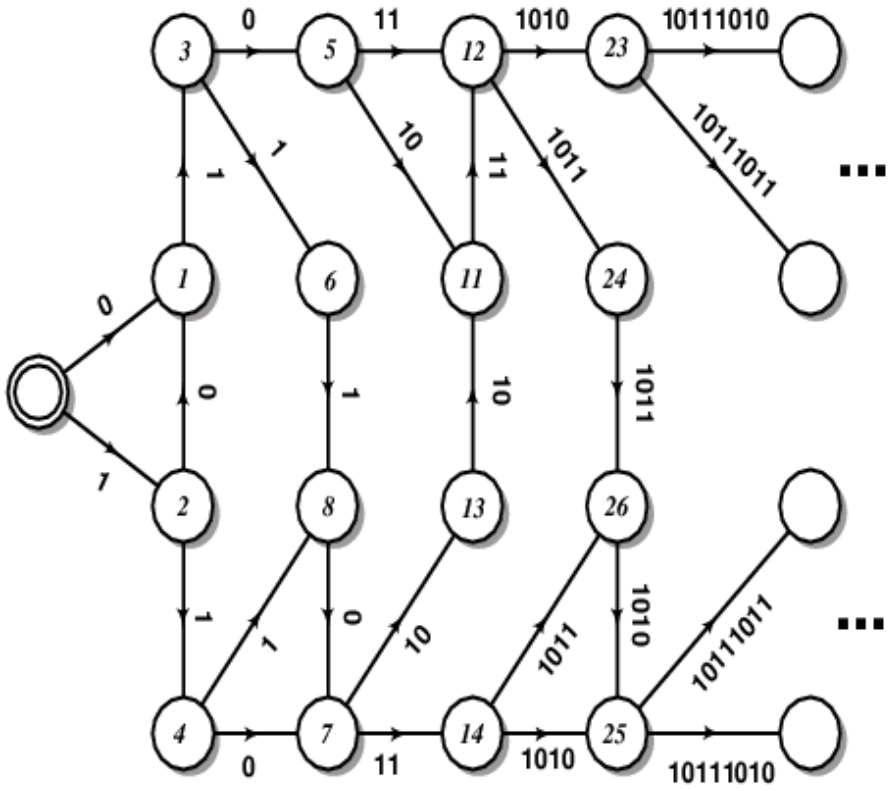
Critical Machine, continued

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



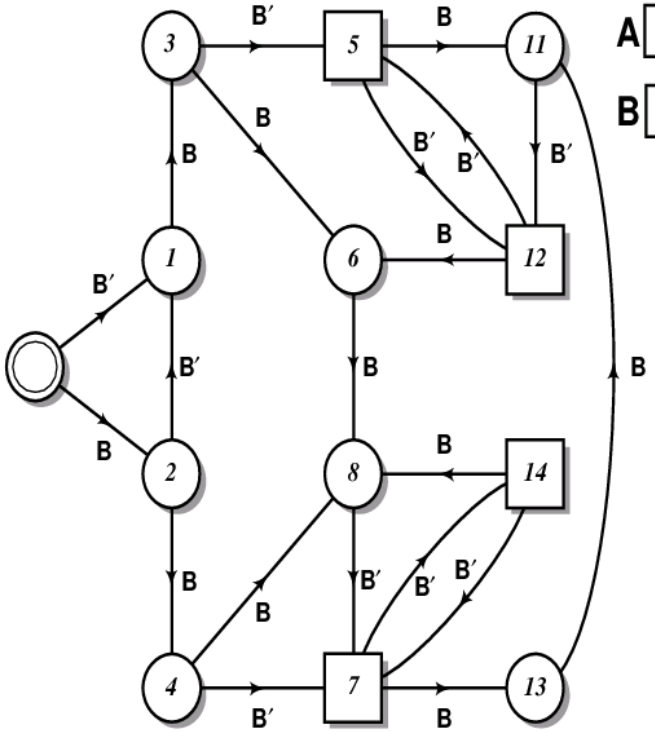
Critical Machine, continued

- 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 $B = 1010$, $B' = 1011$.)

Emergence!

- **Main Point:** Diverging model size at lower level of Chomsky hierarchy necessitated a new model at a higher level (context sensitive languages).
- Something genuinely new—qualitatively different—emerges at the transition point.

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.

© David P. Feldman

Part VIII

Complexity vs. Entropy and Edges of Chaos

© David P. Feldman

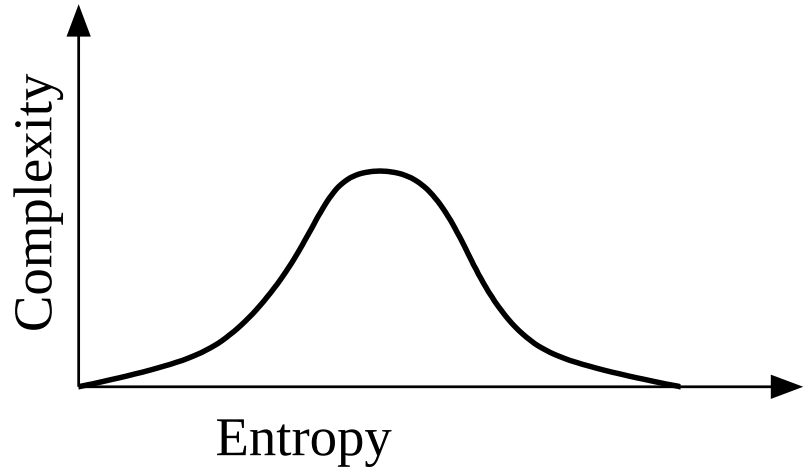
Complexity vs. Entropy

- 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?

© David P. Feldman

One approach: Prescribing Complexity vs. Entropy Behavior

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

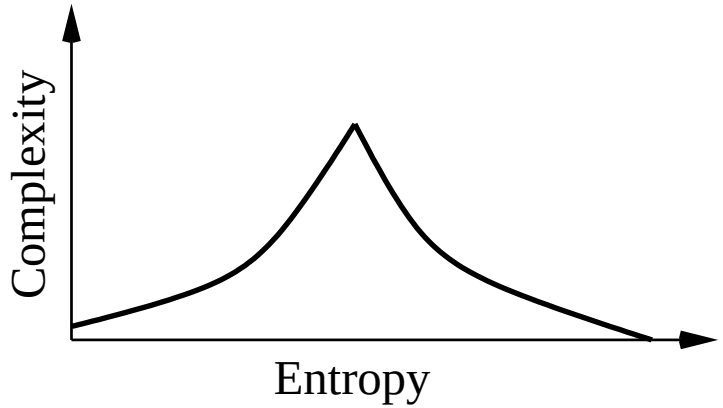


- This plot is often used as the central criteria for defining complexity.

Complexity-Entropy Phase Transition?

Edge of Chaos?

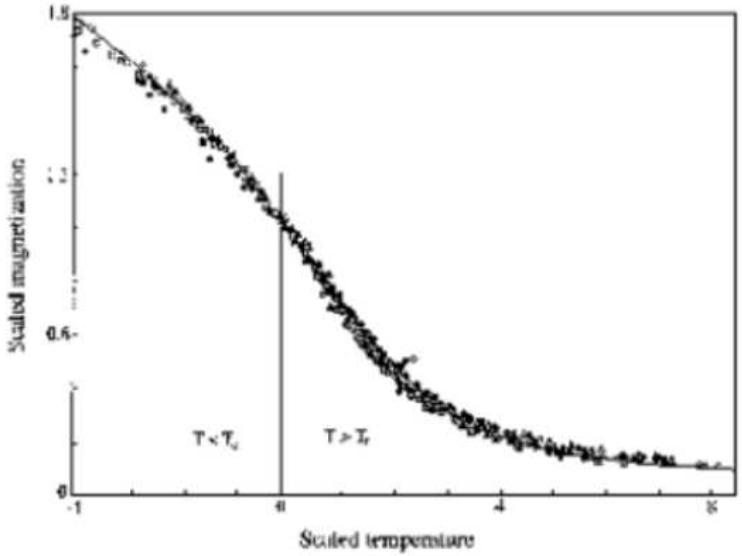
- Additionally, it has been conjectured that there is a sharp transition in complexity as a function of entropy:



- 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_2Fe.



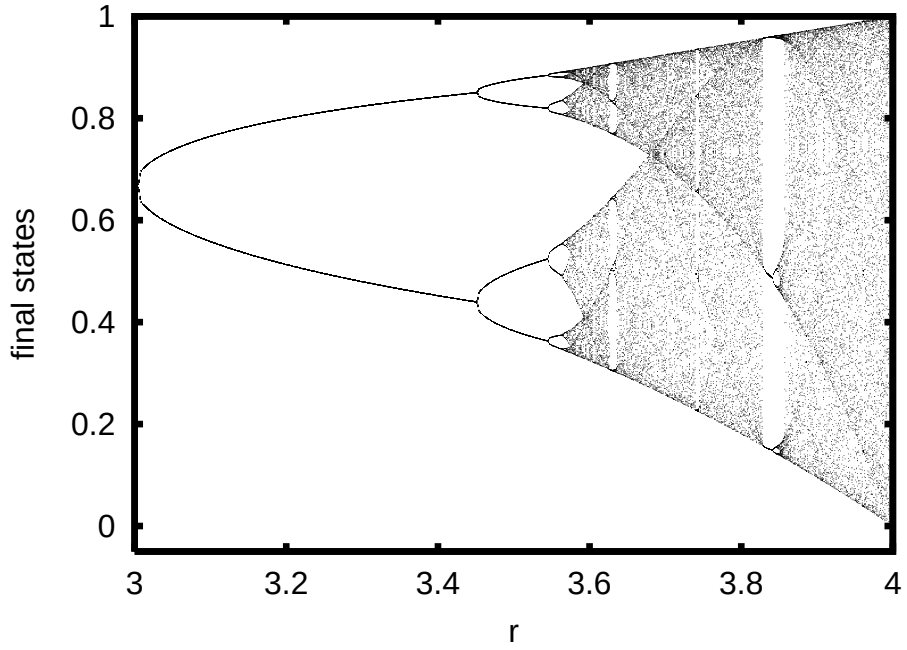
- These materials are very different, but clearly possess some deep similarities.
- Perhaps there is a similar data collapse for some appropriate definitions of complexity and entropy.
- Note: One could trivially obtain this by simply defining complexity to be a single-valued function of the entropy.

Figure source: H.E. Stanley, *Rev. Mod. Phys*. **71**:S358. 1999.

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?

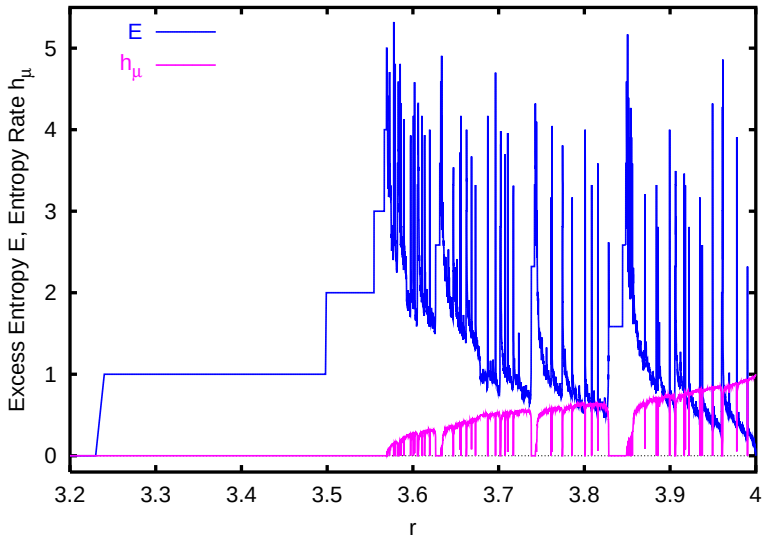
Logistic Equation: Bifurcation Diagram



- 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.



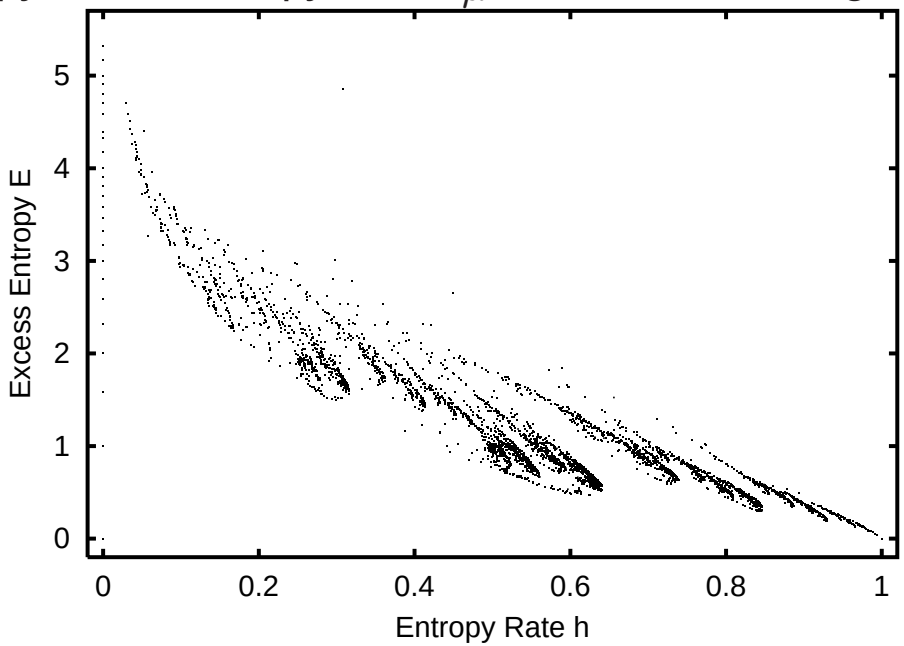
- Note that $\mathbf{E}$ and h_μ depend on a complicated way on r.
- Hard to see how complexity and entropy are related.
- Numerical results. For each r, 1×10^7 symbols were generated. The largest L was 30 for low entropy sequences. r was varied by increments of 0.0001.

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:

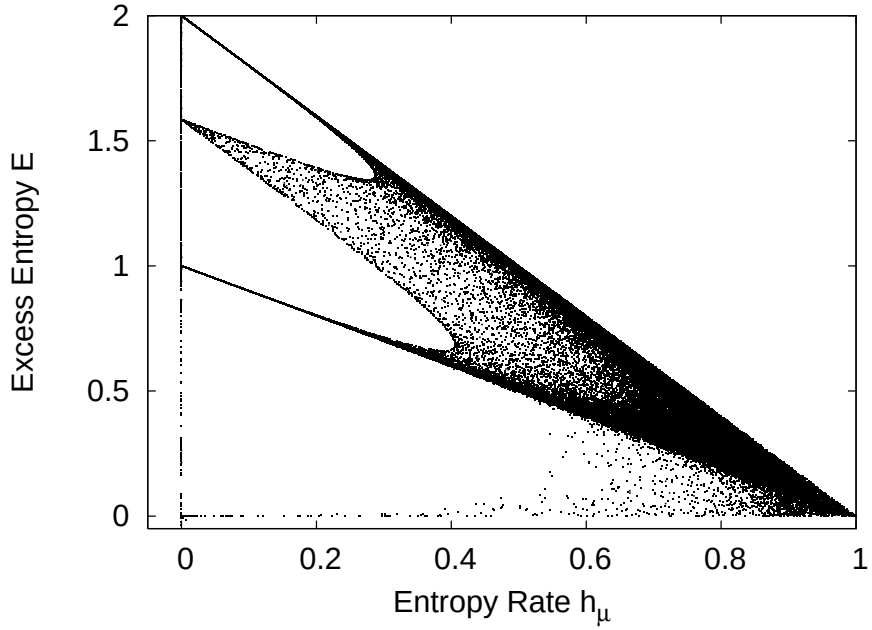
$$\mathcal{H} \equiv -J_1 \sum_i s_i s_{i+1} - J_2 \sum_i s_i s_{i+2} - B \sum s_i \ .$$

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

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

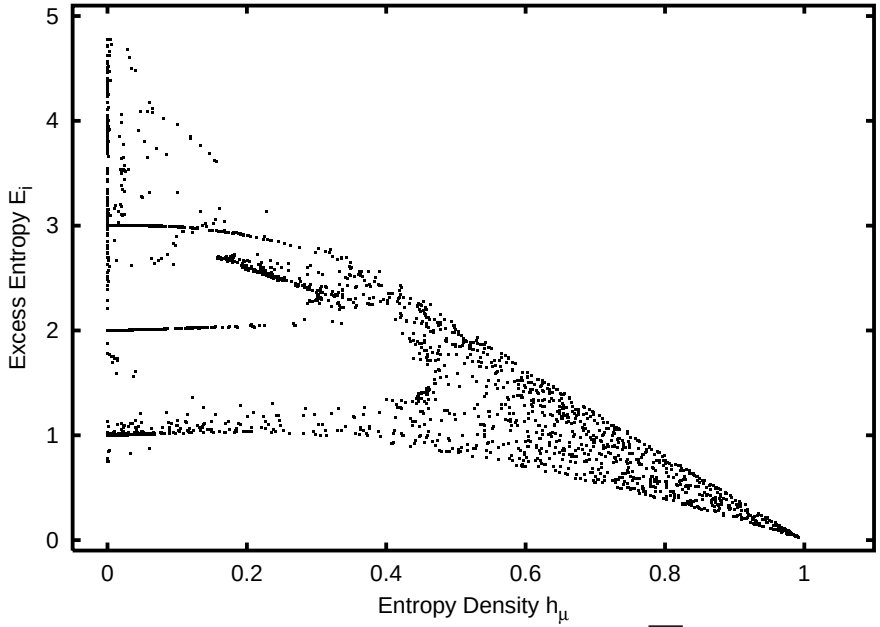
- Ising models are very generic models of spatially extended, discrete degrees of freedom that have some interaction that makes them want to either do the same or the opposite thing.

Complexity-Entropy Diagram for 1D Ising Models



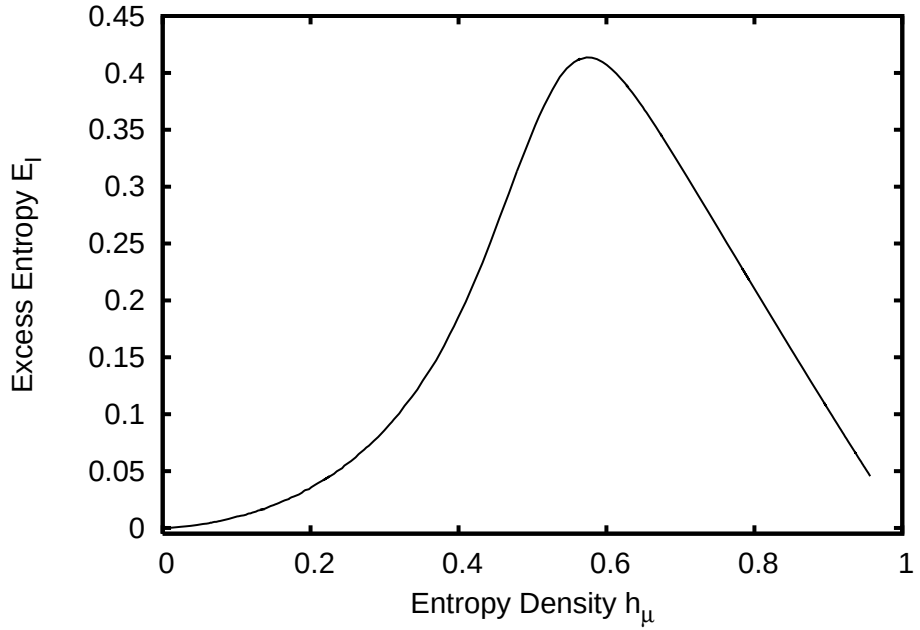
- Excess entropy $\mathbf{E}$ vs. entropy rate h_μ for the one-dimensional Ising model with anti-ferromagnetic couplings.
- Model parameters are chosen uniformly from the following ranges: $J_1 \in [-8, 0]$, $J_2 \in [-8, 0]$, $T \in [0.05, 6.05]$, and $B \in [0, 3]$.
- Note how different this is from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

Complexity-Entropy Diagram for 2D Ising Models



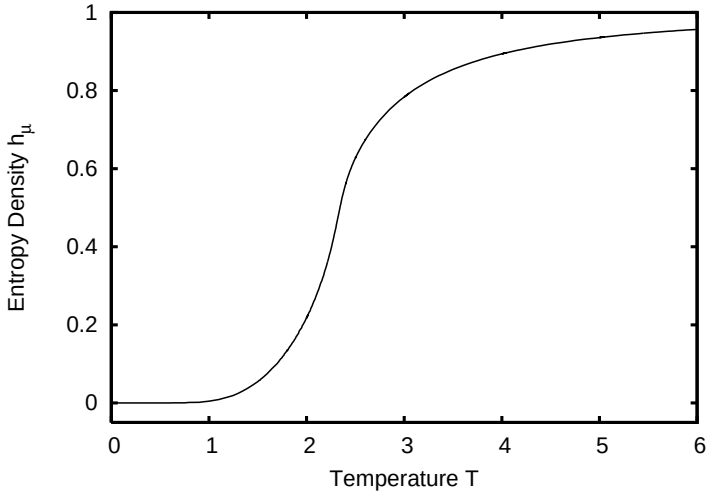
- Mutual information form of the excess entropy $\mathbf{E}_{\mathrm{i}}$ vs. entropy density h_μ for the two-dimensional Ising model with AFM couplings
- Model parameters are chosen uniformly from the following ranges: $J_1 \in [-3, 0]$, $J_2 \in [-3, 0]$, $T \in [0.05, 4.05]$, and $B = 0$.
- Surprisingly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

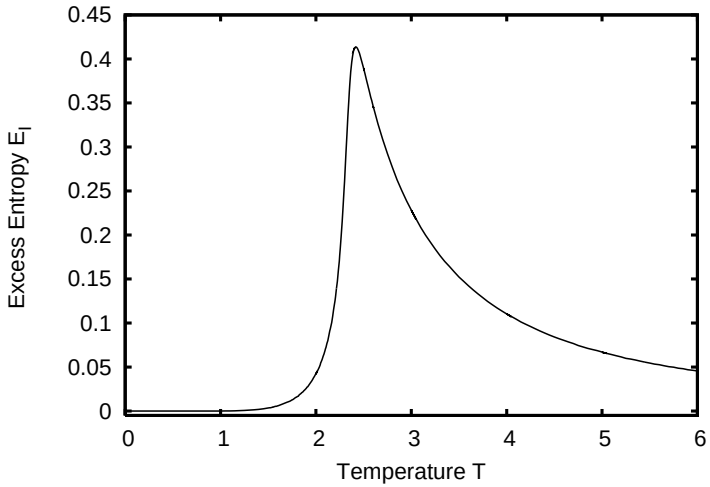
Complexity-Entropy Diagram for 2D Ising Model Phase Transition



- Convergence form of the excess entropy $\mathbf{E}_c$ 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.

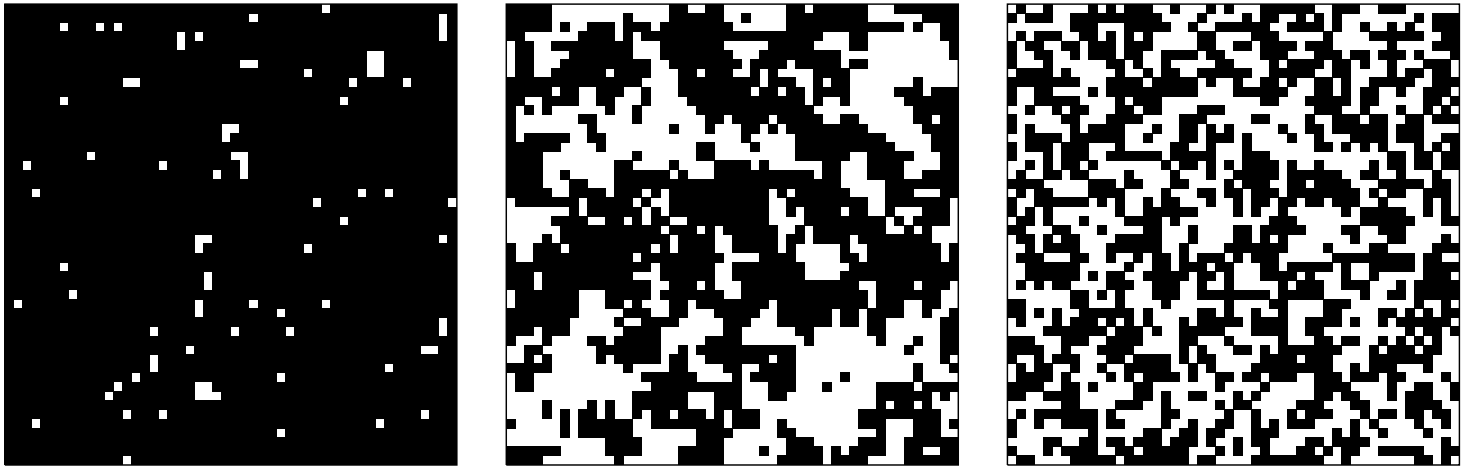
Complexity-Entropy Diagram for 2D Ising Model Phase Transition, continued





- Convergence form of the excess entropy $\mathbf{E}_\mathrm{c}$ 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



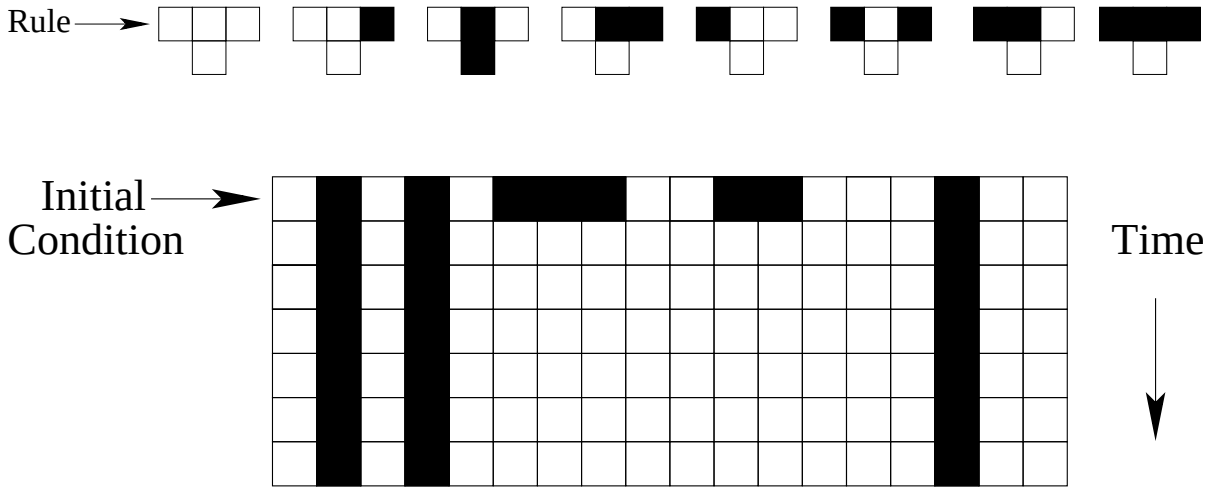
- Typical configurations for the 2D Ising model below, at, and above the critical temperature.

© David P. Feldman

Cellular Automata

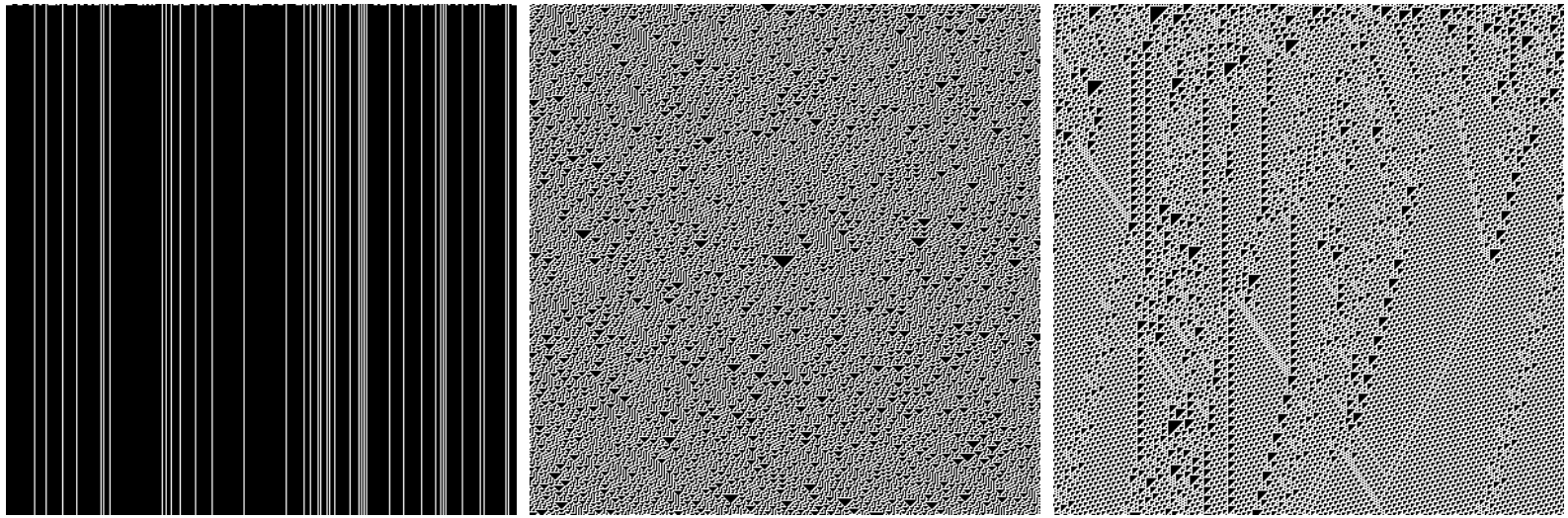
- The next row in the grid is determined by the row directly above it according to a given rule
- Start with a random initial condition

Example:



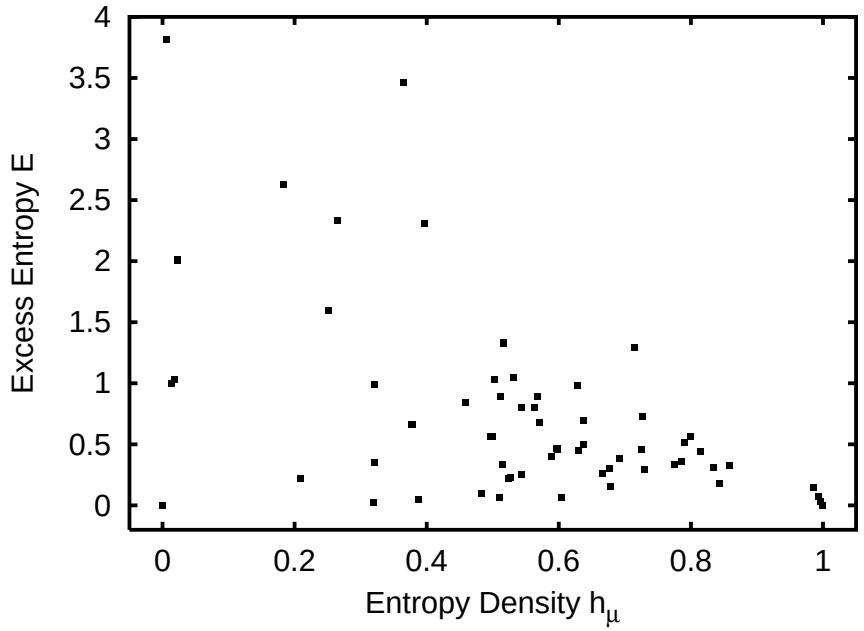
- The number of cells away from the center cell that the rule considers is known as the radius of the CA.

Different Rules Yield Different Patterns



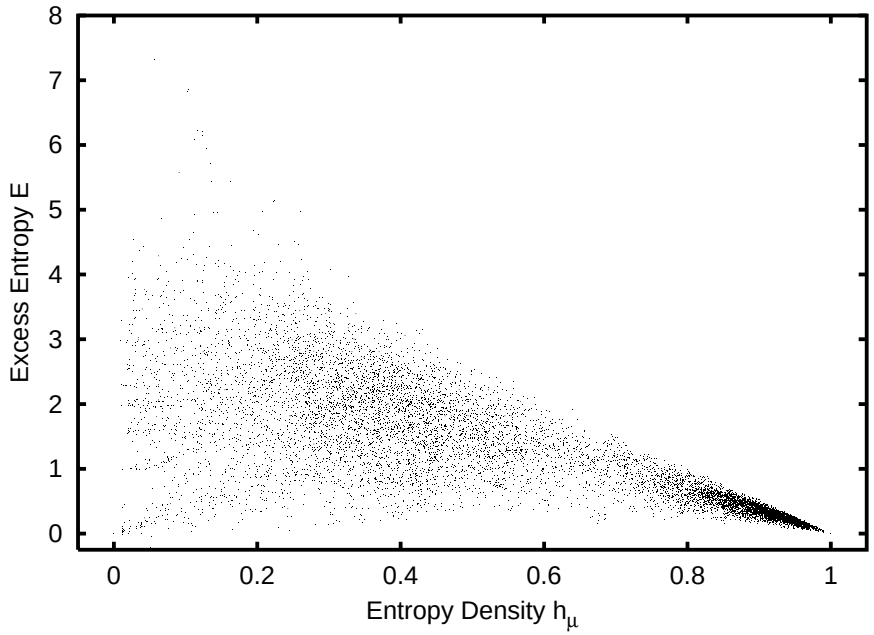
- Each pattern is for a different rule.

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.

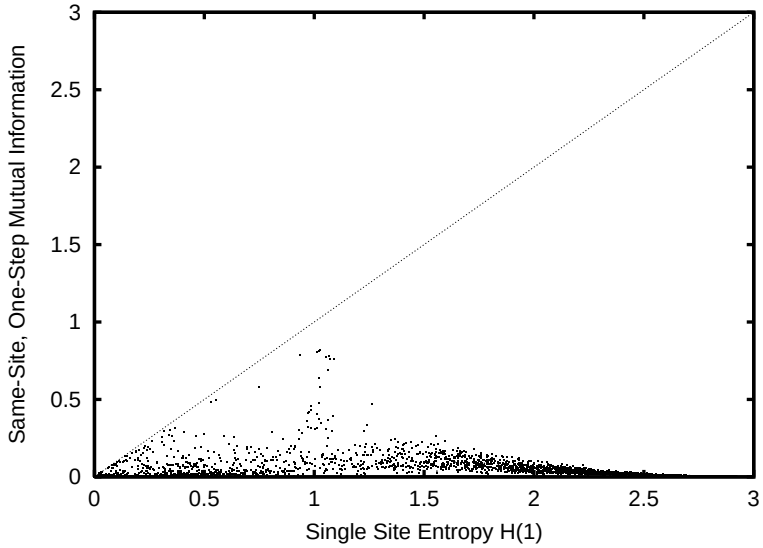
Langton's Lambda

- 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:

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

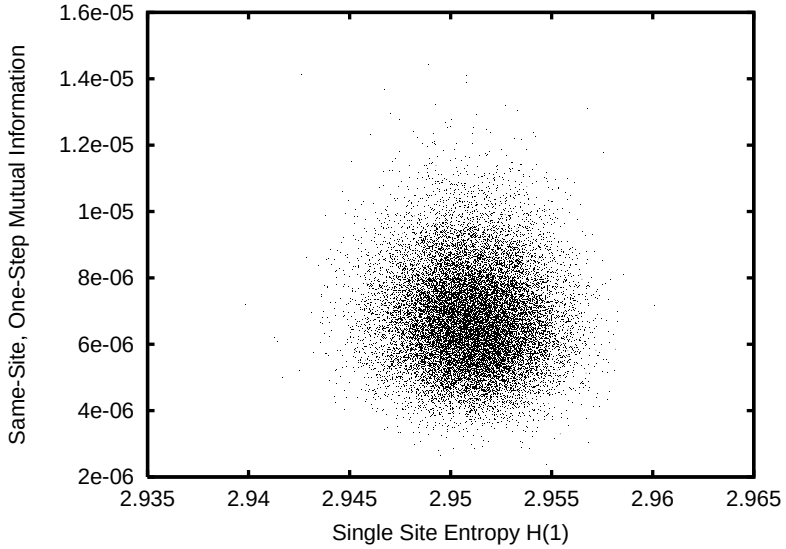
- Langton considered two-dimensional, 8-states, radius-1 cellular automata.
- There are around $10^{30,000}$ such CAs!
- Sweep λ from 0 to 1 in increments of 0.01. Randomly generate 50 different CA rules for each λ.
- For each CA, calculate single-site entropy and one-step mutual information.

"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

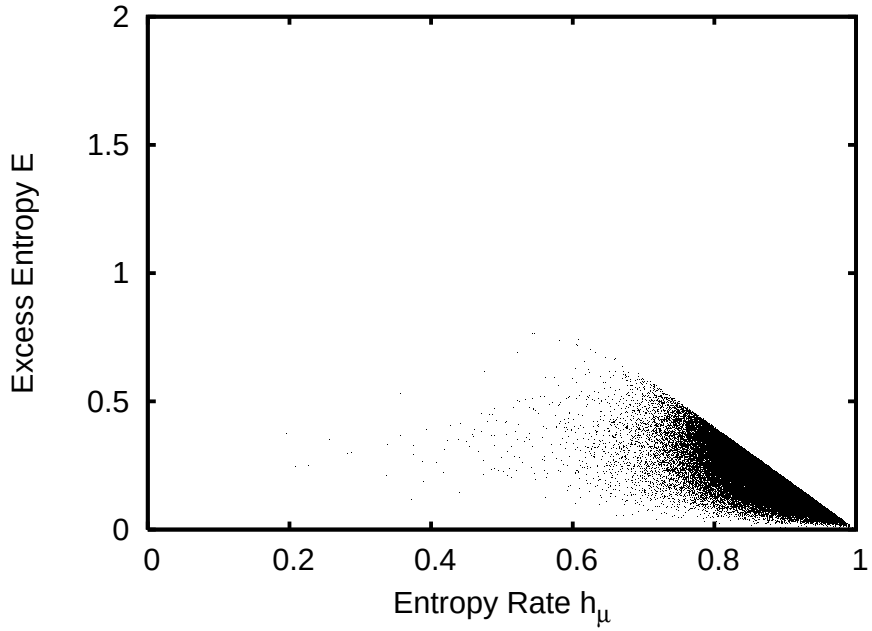


- Single-site entropy $H(1)$ and same-site, one-step mutual information I_2^t for $r = 1$, $K = 8$ two-dimensional cellular automata.
- $20,000$ rules were sampled randomly.
- Note the very small variation in $H(1)$ and I_2^t. All data in this figure would appear as a single dot in the lower right hand 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.

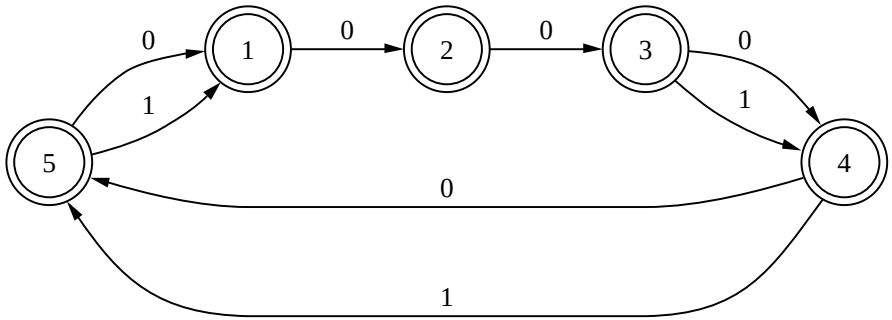
Complexity-Entropy Diagram for Markov Models

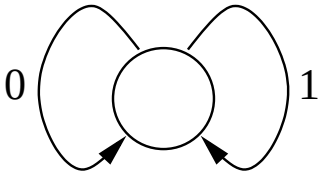


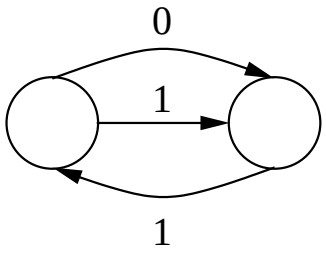
- Excess entropy $\mathbf{E}$ vs. entropy rate h_μ for $100,000$ random Markov models.
- The Markov models here have four states, corresponding to dependence on the previous two symbols, as in the 1D NNN Ising model.
- Transition probabilities chosen uniformly on $[0, 1]$ and then normalized.
- Note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

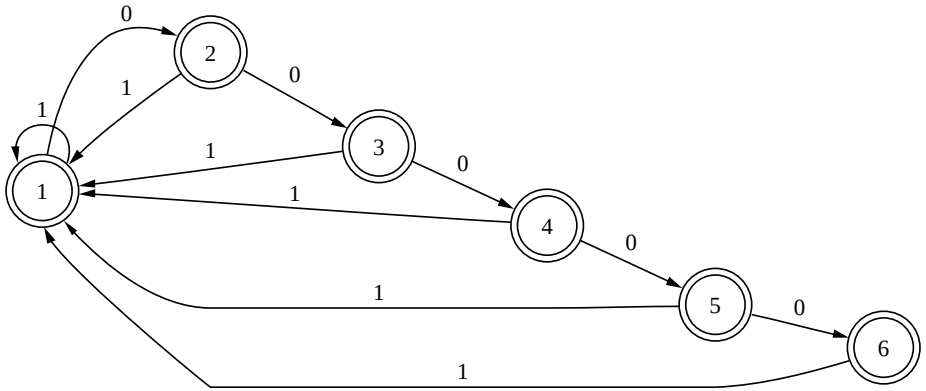
Topological Markov Chain Processes

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

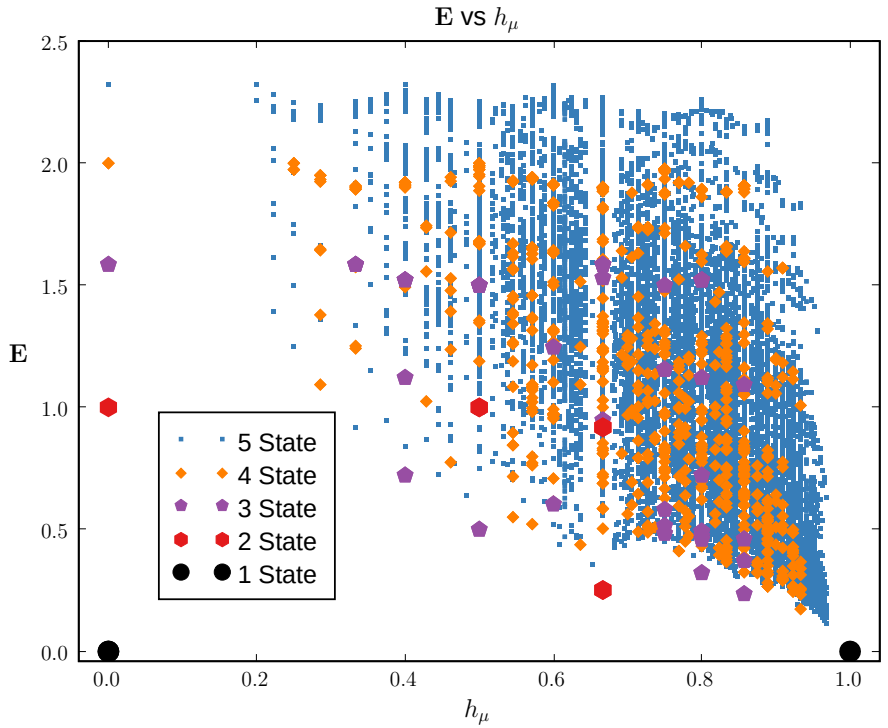








Complexity-Entropy Diagram for Topological Processes



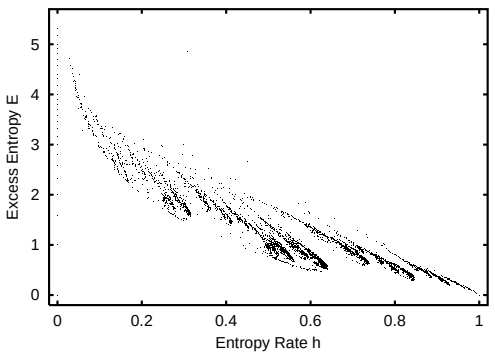
- $h_\mu, \mathbf{E}$ pairs for all $14,694$ distinct topological processes of $n = 1$ to $n = 6$ states.
- Enumeration algorithm by Carl McTague, **E** calculation by Chris Ellison.
- Note the prevalence of high-entropy, high-complexity processes.

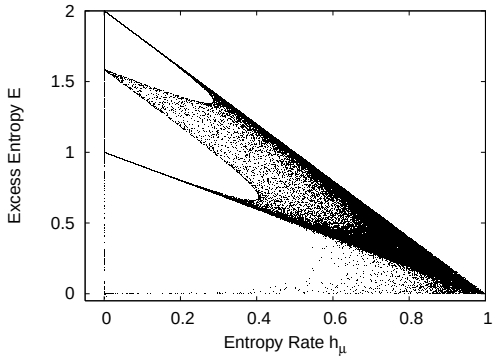
A Gallery of Complexity-Entropy Diagrams

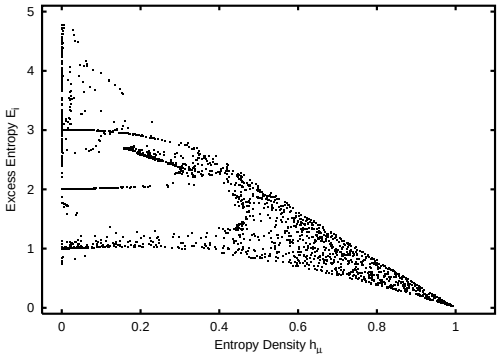
The next slide shows, left to right, top to bottom, complexity-entropy diagrams for:

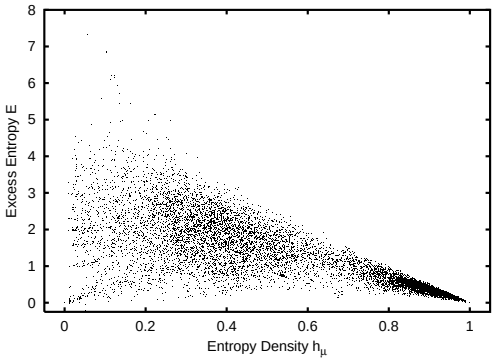
1. Logistic Equation
2. One-Dimensional Ising model with nearest- and next-nearest-neighbor interactions
3. Two-Dimensional Ising model with nearest- and next-nearest-neighbor interactions
4. One-Dimensional radius-2 cellular automata
5. Random Markov chains
6. All 6-state topological processes

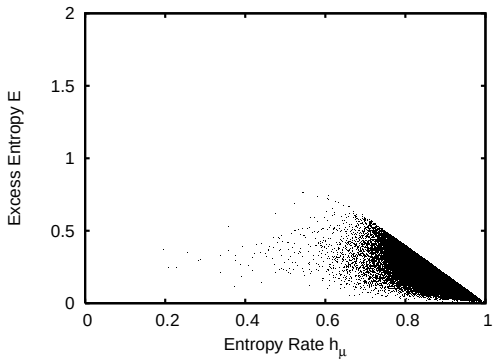
A Mosaic of Complexity-Entropy Diagrams

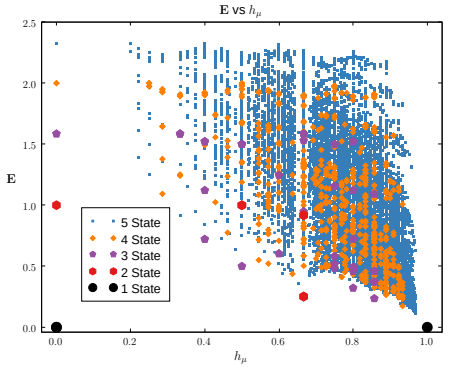






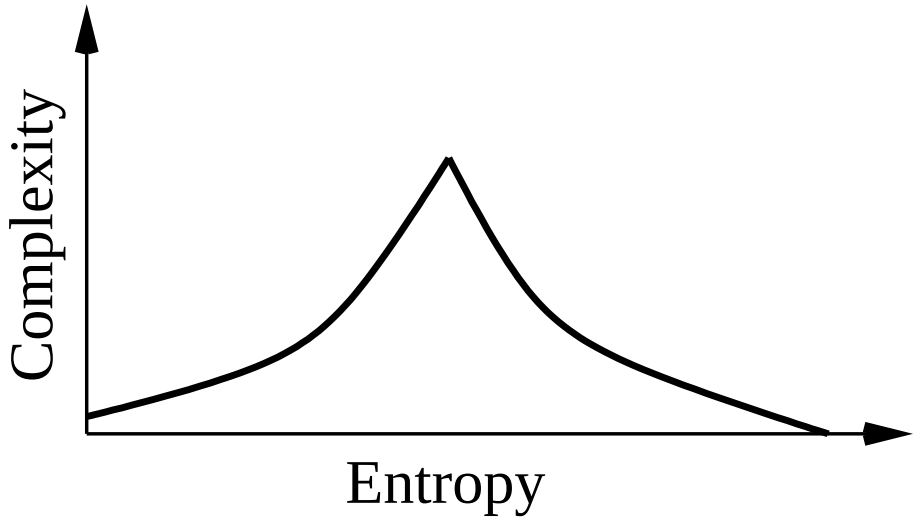






Complexity-Entropy Diagrams: Summary

- 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 IX

Thoughts on the Subjectivity of Complexity

© David P. Feldman

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

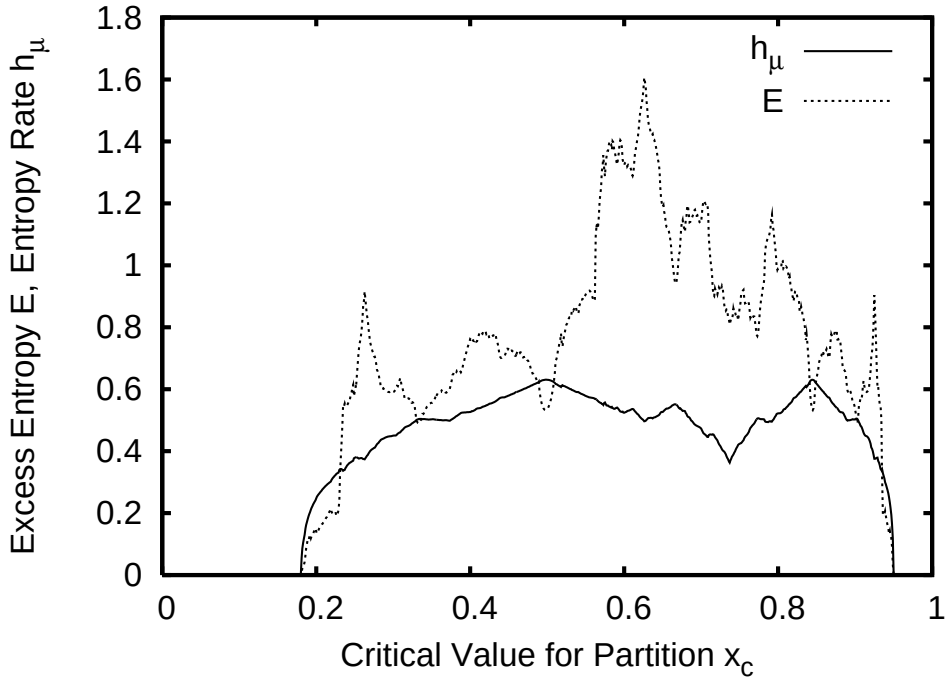
- Iterate the logistic equation: $x_{n+1} = f(x_n)$, where $f(x) = rx(1-x)$.
- Result is a sequence of numbers. E.g., $0.445, 0.894, 0.22, 0.344, \ldots$.
- Generate symbol sequence via:

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

- 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)

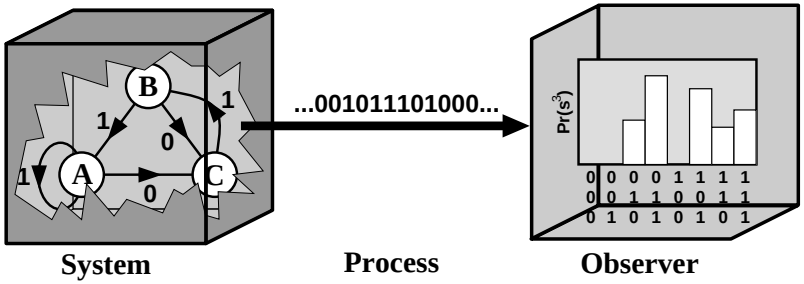
- Our estimates for h_μ and $\mathbf{E}$ depend strongly on x_c.
- Using an $x_c \neq 0.5$ leads to an h_μ is always lower than the true value.
- Using an $x_c \neq 0.5$ can lead to an over- or an under-estimate of $\mathbf{E}$.



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

Example III: A Randomness Puzzle

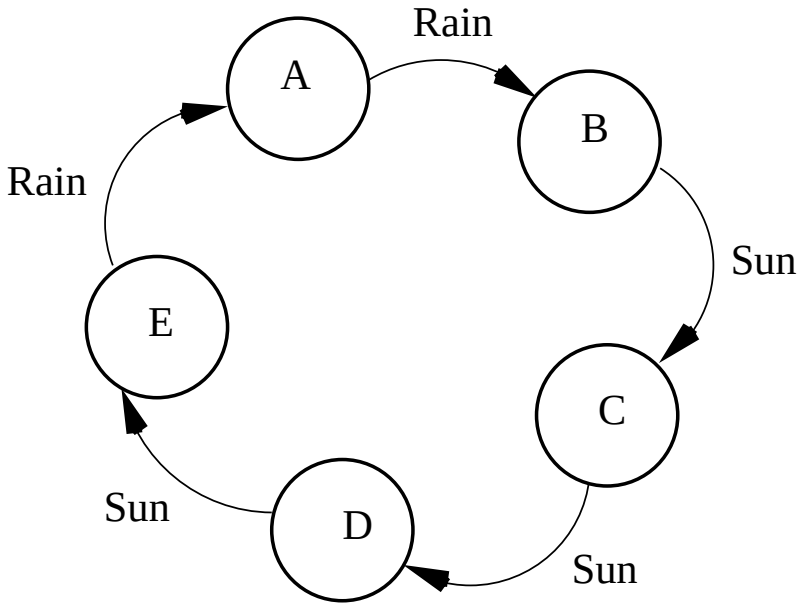
- Suppose we consider the binary expansion of π. Calculate its entropy rate h_μ and we'll find that it's 1.
- How can π be random? Isn't there a simple, deterministic algorithm to calculate digits of π?
- It is not random if one uses Kolmogorov complexity, since there is a short algorithm to produce 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** $\mathbf{T}$:

$$\mathbf{T} \equiv \sum_{L=0}^{\infty} \mathcal{H}(L) \ . \tag{16}$$

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:

$$\mathbf{T}_{\max} \sim \frac{P}{2} \log_2 P \, , \tag{17}$$

and

$$\mathbf{T}_{\min} \sim \frac{1}{2} \log_2^2 P \, , \tag{18}$$

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

$$\mathbf{T}_{\max} \approx 1024 \, , \text{ and } \mathbf{T}_{\min} \approx 32 \, . \tag{19}$$

- 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.

© David P. Feldman

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 X

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

© David P. Feldman

Conclusion

- Information Theory and Computation Theory, together with various dimensions and similar measures developed in the study of fractals and chaos, provide a set of tools that can be used to analyze and quantify randomness, memory, structure, pattern, complexity, and so on.
- The relations between order and disorder, complexity and randomness, are varied and subtle.
- These tools can help us, both practically and conceptually, when thinking about complexity and emergence.

© David P. Feldman

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.