A Crash Course in Information Theory

David P. Feldman

College of the Atlantic

18 June 2018

Table of Contents

Information Theory is...

- Both a subfield of electrical and computer engineering and

Info theory is commonly used across complex systems.

Goal for today: Give a solid introduction to the basic elements of information theory with just the right amount of math.

Information Theory is...

- Both a subfield of electrical and computer engineering and
- machinery to make statements about probability distributions and relations among them,

Info theory is commonly used across complex systems.

Goal for today: Give a solid introduction to the basic elements of information theory with just the right amount of math.

Information Theory is...

- Both a subfield of electrical and computer engineering and
- machinery to make statements about probability distributions and relations among them,
- including memory and non-linear correlations and relationships,

Info theory is commonly used across complex systems.

Goal for today: Give a solid introduction to the basic elements of information theory with just the right amount of math.

Information Theory is...

- Both a subfield of electrical and computer engineering and
- machinery to make statements about probability distributions and relations among them,
- including memory and non-linear correlations and relationships,
- that is complementary to the Theory of Computation.

Info theory is commonly used across complex systems.

Goal for today: Give a solid introduction to the basic elements of information theory with just the right amount of math.

Some Recommended Info Theory References

1. T.M. Cover and J.A. Thomas, *Elements of Information Theory*. Wiley, 1991. By far the best information theory text around.
2. Raymond Yeung, *A First Course in Information Theory*. Springer, 2006.
3. 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.
4. J.P. Crutchfield and D.P. Feldman, "Regularities Unseen, Randomness Observed: Levels of Entropy Convergence." *Chaos* **15**:25–53. 2003.
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

- 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?

- 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?

Different amounts of uncertainty?

- 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.

Different amounts of uncertainty?

- 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

Different amounts of uncertainty?

- 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 slide **uniquely** determine $H[X]$, up to a multiplicative constant.

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

$$H[X] \equiv -\sum_{x \in \mathcal{X}} \Pr(x) \log_2(\Pr(x)) \ . \tag{1}$$

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

The requirements on the previous slide **uniquely** determine $H[X]$, up to a multiplicative constant.
The Shannon entropy of a random variable X is given by:

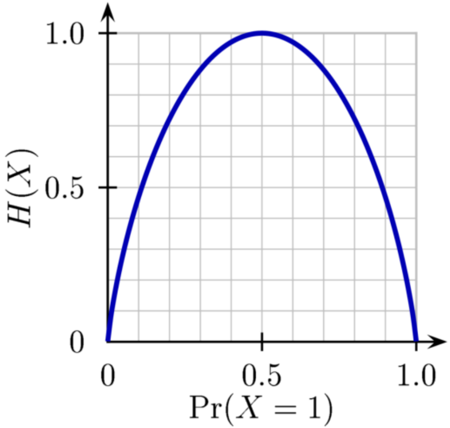
$$H[X] \equiv -\sum_{x \in \mathcal{X}} \Pr(x) \log_2(\Pr(x)) . \tag{1}$$

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

Examples

- **Fair Coin:** $\Pr(h) = \frac{1}{2}, \Pr(t) = \frac{1}{2}$. $H = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = \mathbf{1}$ **bit**.
- **Biased Coin:** $\Pr(h) = 0.6, \Pr(t) = 0.4$.
  $H = -0.6\log_2 0.6 - 0.4\log_2 0.4 = \mathbf{0.971}$ **bits**.
- **More Biased Coin:** $\Pr(h) = 0.9, \Pr(t) = 0.1$.
  $H = -0.9\log_2 0.9 - 0.1\log_2 0.1 = \mathbf{0.469}$ **bits**.
- **Totally Biased Coin:** $\Pr(h) = 1.0, \Pr(t) = 0.0$.
  $H = -1.0\log_2 1.0 - 0.0\log_2 0.0 = \mathbf{0.0}$ **bits**.

Binary Entropy



Entropy of a binary variable as a function of its bias.

Figure Source: original work by Brona, published at `https://commons.wikimedia.org/wiki/File:Binary_entropy_plot.svg`.

Average Surprise

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

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

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

Guessing Games 1

Consider a random variable X with four equally likely outcomes:
$\Pr(a) = \Pr(b) = \Pr(c) = \Pr(d) = \frac{1}{4}$.

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

Guessing Games 1

Consider a random variable X with four equally likely outcomes:
$\mathrm{Pr}(a) = \mathrm{Pr}(b) = \mathrm{Pr}(c) = \mathrm{Pr}(d) = \frac{1}{4}$.

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

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 2

What's the best strategy for guessing Y?

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

Guessing Games 2

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, $H[Y] = 1.75$!!

Entropy Measures Average Number of Guesses

Strategy: try to divide the probability in half with each guess.

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 more common occurrences of X.
- This is identical to the strategy for guessing outcomes.

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 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

Shannon's noiseless source coding theorem:

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

Also known as 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

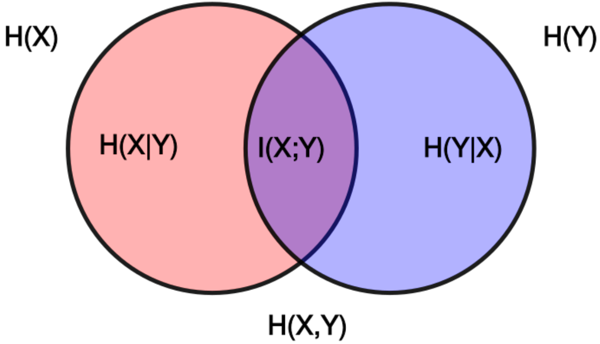
Definition

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

Relationships

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

Information Diagram



The information diagram shows the relationship among joint and conditional entropies and the mutual information.

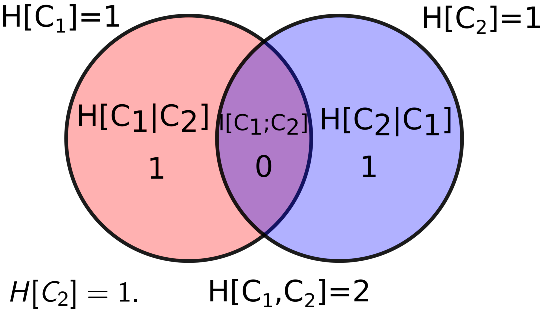
Figure Source: Konrad Voelkel, released to the public domain.

`https://commons.wikimedia.org/wiki/File:Entropy-mutual-information-relative-entropy-relation-diagram.svg`

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	$\frac{1}{4}$	$\frac{1}{4}$

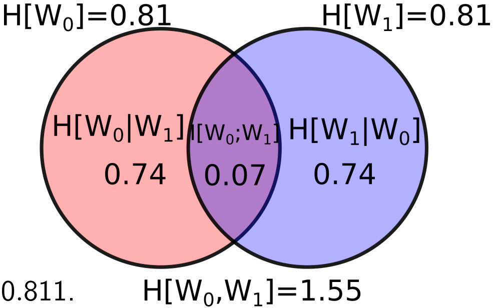


- $H[C_1] = 1$ and $H[C_2] = 1$.
- $H[C_1, C_2] = 2$.
- $H[C_1|C_2] = 1$. Even if you know what C_2 is, you're still uncertain about C_1.
- $I[C_1; C_2] = 0$. Knowing C_1 does not reduce your uncertainty of C_2 at all.
- C_1 carries no information about C_2.

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	$\frac{1}{8}$	$\frac{1}{8}$



- $H[W_0] = 0.811 = H[W_1] = 0.811.$
- $H[W_0, W_1] = 1.549.$
- 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.

Estimating Entropies

By the way...

- Probabilities $\Pr(x)$, etc., can be estimated empirically.
- Just observe the occurrences c_i of different outcomes and estimate the frequencies:

$$\Pr(x_i) = \frac{c_i}{\sum_j c_j} \, .$$

No big deal.

Estimating Entropies

By the way...

- Probabilities $\Pr(x)$, etc., can be estimated empirically.
- Just observe the occurrences c_i of different outcomes and estimate the frequencies:

$$\Pr(x_i) = \frac{c_i}{\sum_j c_j} .$$

No big deal.

However, this will lead to a biased under-estimate for $H[X]$. For more accurate ways of estimate $H[X]$, see, e.g.,

- Schürmann and Grassberger. Chaos 6:414-427. 1996.
- Kraskov, Stögbauer, and Grassberger. *Phys Rev E* 69.6: 066138. 2004.

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.

Relative Entropy

The **Relative Entropy** or the **Kullback-Leibler** distance between two distributions $p(x)$ and $q(x)$ is:

$$D(p||q) \equiv \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} .$$

$D(p||q)$ is how much more random X appears if one assumes it is distributed according to q when it is actually distributed according to p.

$D(p||q)$ is measure of "entropic distance" between p and q.

Relative Entropy: Example

$X \in \{a, b, c, d\}$
$p : p(a) = 1/2, p(b) = 1/4, p(c) = 1/8, p(d) = 1/8$
$q : q(a) = 1/4, q(b) = 1/4, q(c) = 1/4, q(d) = 1/4$

$$D(p||q) \equiv \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} \, ,$$

$$D(p||q) \equiv \sum_{x \in \mathcal{X}} -p(x) \log_2 q(x) - H(p) \, .$$

The first term on the right is the expected code length if we used the code for q for a variable that was actually distributed according to p.

Relative Entropy: Example, continued

$X \in \{a, b, c, d\}$
$p : p(a) = 1/2, p(b) = 1/4, p(c) = 1/8, p(d) = 1/8$
$q : q(a) = 1/4, q(b) = 1/4, q(c) = 1/4, q(d) = 1/4$
Optimal code for X distributed according to q:

$$a \longrightarrow 01\ ,\quad b \longrightarrow 00\ ,\quad c \longrightarrow 10\ ,\quad d \longrightarrow 11$$

$$D(p||q) \equiv \sum_{x \in \mathcal{X}} -p(x) \log_2 q(x) - H(p)\ .$$

Ave length of code for X using q coding if X is distributed according to p:

$$\frac{1}{2}(2) + \frac{1}{4}(2) + \frac{1}{8}(2) + \frac{1}{8}(2) = 2$$

Relative Entropy: Example, continued further

$X \in \{a, b, c, d\}$
$p : p(a) = 1/2, p(b) = 1/4, p(c) = 1/8, p(d) = 1/8$
$q : q(a) = 1/4, q(b) = 1/4, q(c) = 1/4, q(d) = 1/4$

Recall that $H(p) = 1.75$. Then

$$D(p||q) \equiv \sum_{x \in \mathcal{X}} -p(x) \log_2 q(x) - H(p) \; .$$

$$D(p||q) = 2 - 1.75 = 0.25 \; .$$

So using the code for q when X is distributed according to p adds 0.25 to the average code length.

Exercise: Show that $D(q||p) = 0.25$.

Relative Entropy Summary

- $D(p||q)$ is not a proper distance. It is not symmetric and does not obey the triangle inequality.
- Arises in many different learning/adapting and statistics contexts.
- Measures the "coding mismatch" or "entropic distance" between p and q.

Summary and Reflections

- Information theory provides a natural language for working with probabilities.
- Information theory is *not* a theory of semantics or meaning.
- Information theory is used throughout complex systems.
- Often shows common mathematical structures across different domains and contexts.

Information Theory: Part II Applications to Stochastic Processes

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

$$\cdots 0011001001010110100110011101 0110 \cdots$$

- 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

Stochastic Process Notation

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

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

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

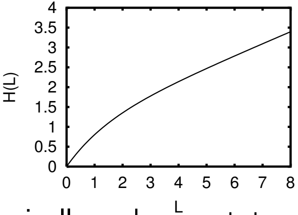
$$\cdots 11010100101101010101001001010010 \cdots$$

Entropy Growth

- Entropy of L-block:

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

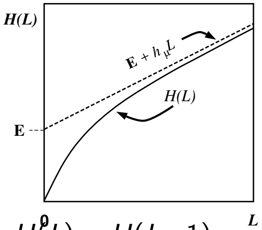
- $H(L)$ = average uncertainty about the outcome of L consecutive variables.



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

Entropy Rate

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



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

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

Entropy Rate, continued

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

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

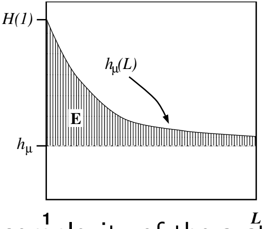
- $h_\mu(L) = H[S_L | S_1 S_1 \ldots S_{L-1}]$
- I.e., $h_\mu(L)$ is the average uncertainty of the next symbol, given that the previous L symbols have been observed.

Interpretations of Entropy Rate

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

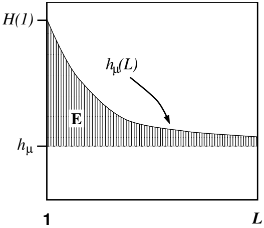
How does $h_\mu(L)$ approach h_μ?

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



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

The Excess Entropy



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

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

- **E** is thus the total amount of randomness that is "explained away" by considering larger blocks of variables.

Mutual information

- One can show that **E** is equal to the mutual information between the "past" and the "future":

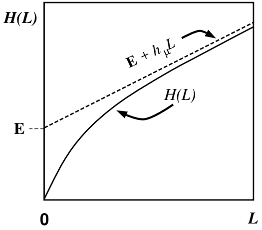
$$\mathbf{E} = I(\overleftarrow{S}; \overrightarrow{S}) \equiv \sum_{\{\overleftrightarrow{s}\}} \Pr(\overleftrightarrow{s}) \log_2 \left[\frac{\Pr(\overleftrightarrow{s})}{\Pr(\overleftarrow{s})\Pr(\overrightarrow{s})} \right] .$$

- **E** is thus the amount one half "remembers" about the other, the reduction in uncertainty about the future given knowledge of the past.
- Equivalently, **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

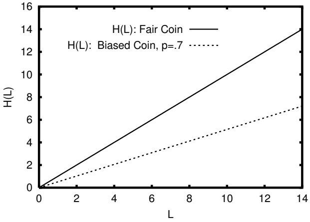
- **E** is the y-intercept of the straight line to which $H(L)$ asymptotes.
- $\mathbf{E} = \lim_{L\to\infty} [H(L) - h_\mu L]$.



Excess Entropy Summary

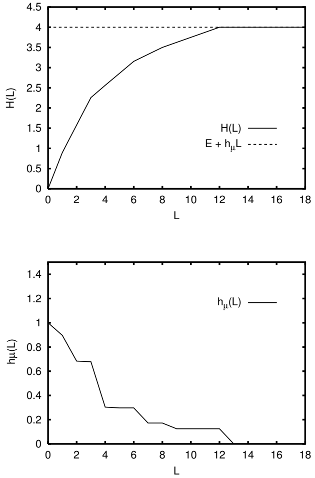
- Is a structural property of the system — measures a feature complementary to entropy.
- Measures memory or spatial structure.
- Lower bound for statistical complexity, minimum amount of information needed for minimal stochastic model of system

Example I: Fair Coin



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

Example II: Periodic Sequence



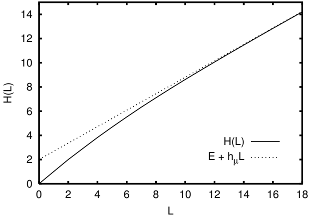
- Sequence: $\ldots 1010111011101110 \ldots$

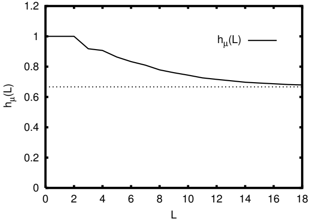
Example II, continued

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

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

Example III: Random, Random, XOR





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

Example III, continued

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

Excess Entropy: Notes on Terminology

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

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

Excess Entropy: Selected References and Applications

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

Excess Entropy: Selected References and Applications, continued

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