Information in Processes II

Reading for this lecture:

EIT, Chapter 4 and Secs. 5.1-5.6 and 7.1-7.7.
MET in CMech Readings.

Entropy Growth for Stationary Stochastic Processes: $\Pr(\overleftrightarrow{S})$

Block Entropy:

$$H(L) = H\left(\Pr(s^L)\right) = - \sum_{s^L \in \mathcal{A}^L} \Pr(s^L) \log_2 \Pr(s^L)$$

Monotone increasing: $H(L) \geq H(L-1)$

Adding a random variable cannot decrease entropy:

$$H(S_1, S_2, \ldots, S_L) \leq H(S_1, S_2, \ldots, S_L, S_{L+1})$$

No measurements, no information: $H(0) = 0$

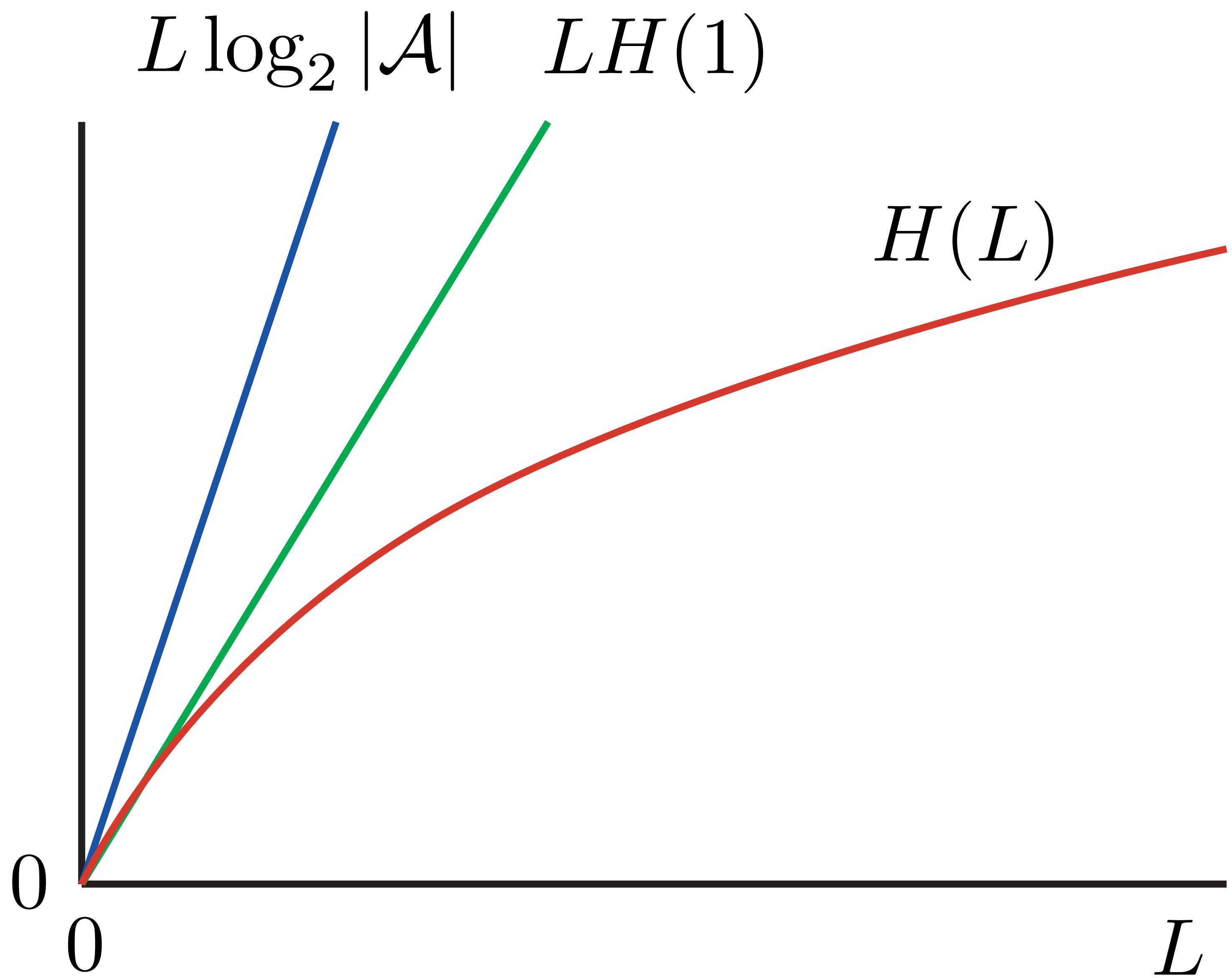
Bounds:

(1) Crude: $H(L) \leq L \log_2 |\mathcal{A}|$

(2) 1-block Markov: $H(L) \leq L H(1)$

Entropy Growth for Stationary Stochastic Processes ...

Block Entropy Curves:



Information in Processes ...

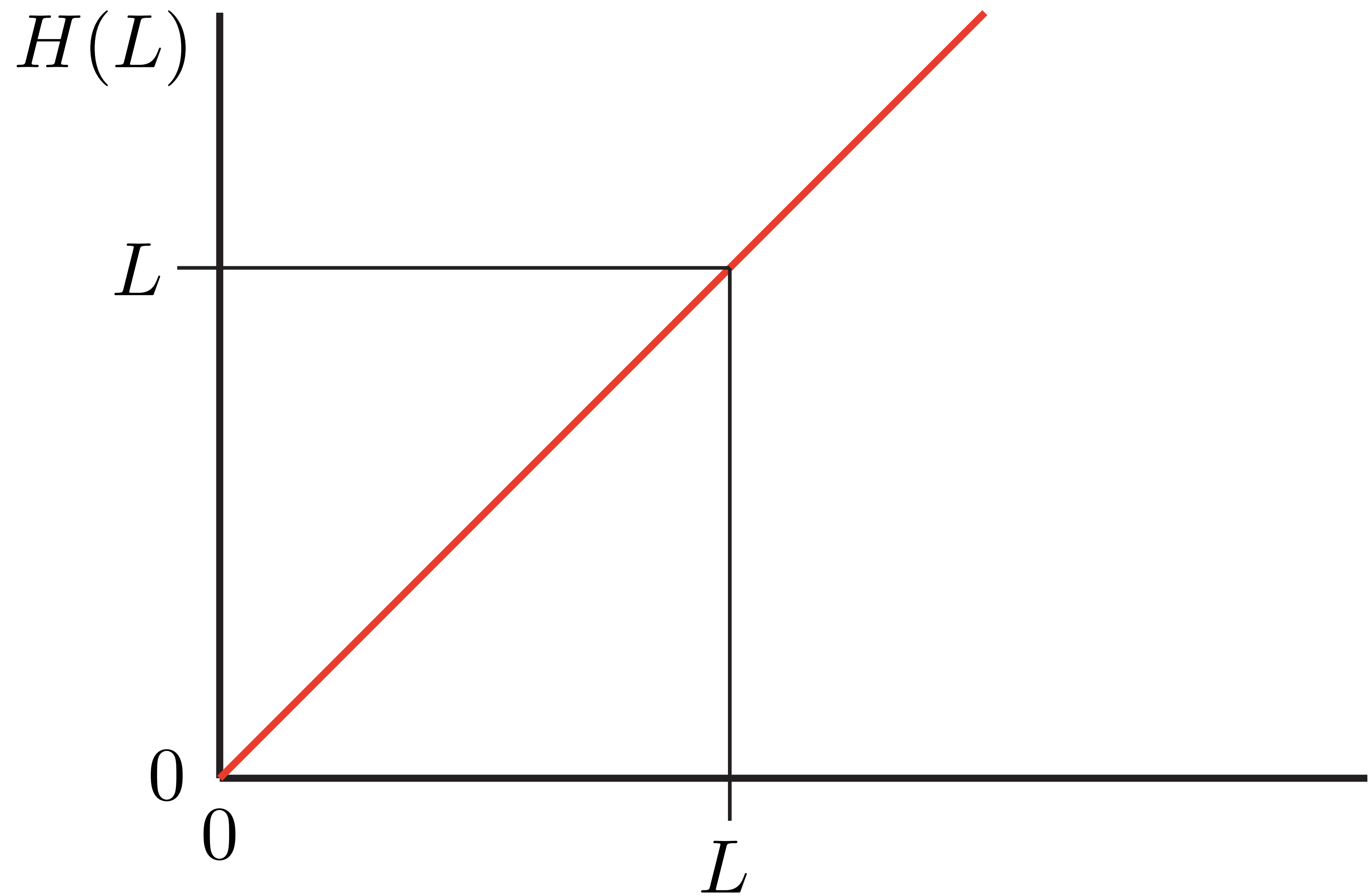
Entropy Growth for Stationary Stochastic Processes ...

Block Entropy ...

Example: Fair Coin

$$\Pr(s^L) = \frac{1}{2^L}$$

$$H(L) = L$$



Information in Processes ...

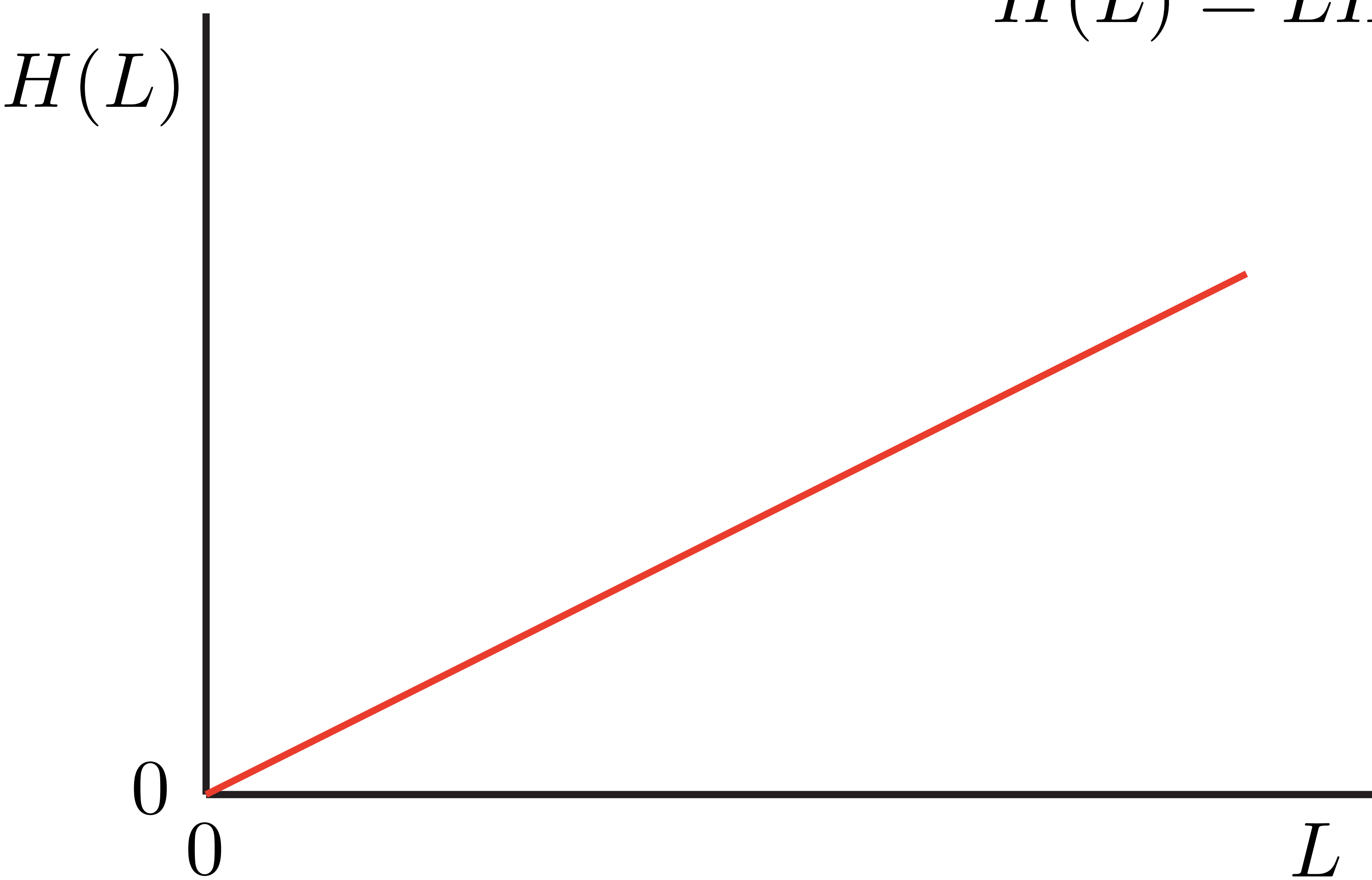
Entropy Growth for Stationary Stochastic Processes ...

Block Entropy ...

Example: Biased Coin

$$\Pr(s^L) = p^n (1-p)^{L-n}$$

$$H(L) = LH(p)$$



For any IID process:

$$H(L) = LH(S_1)$$

Information in Processes ...

Entropy Growth for Stationary Stochastic Processes ...

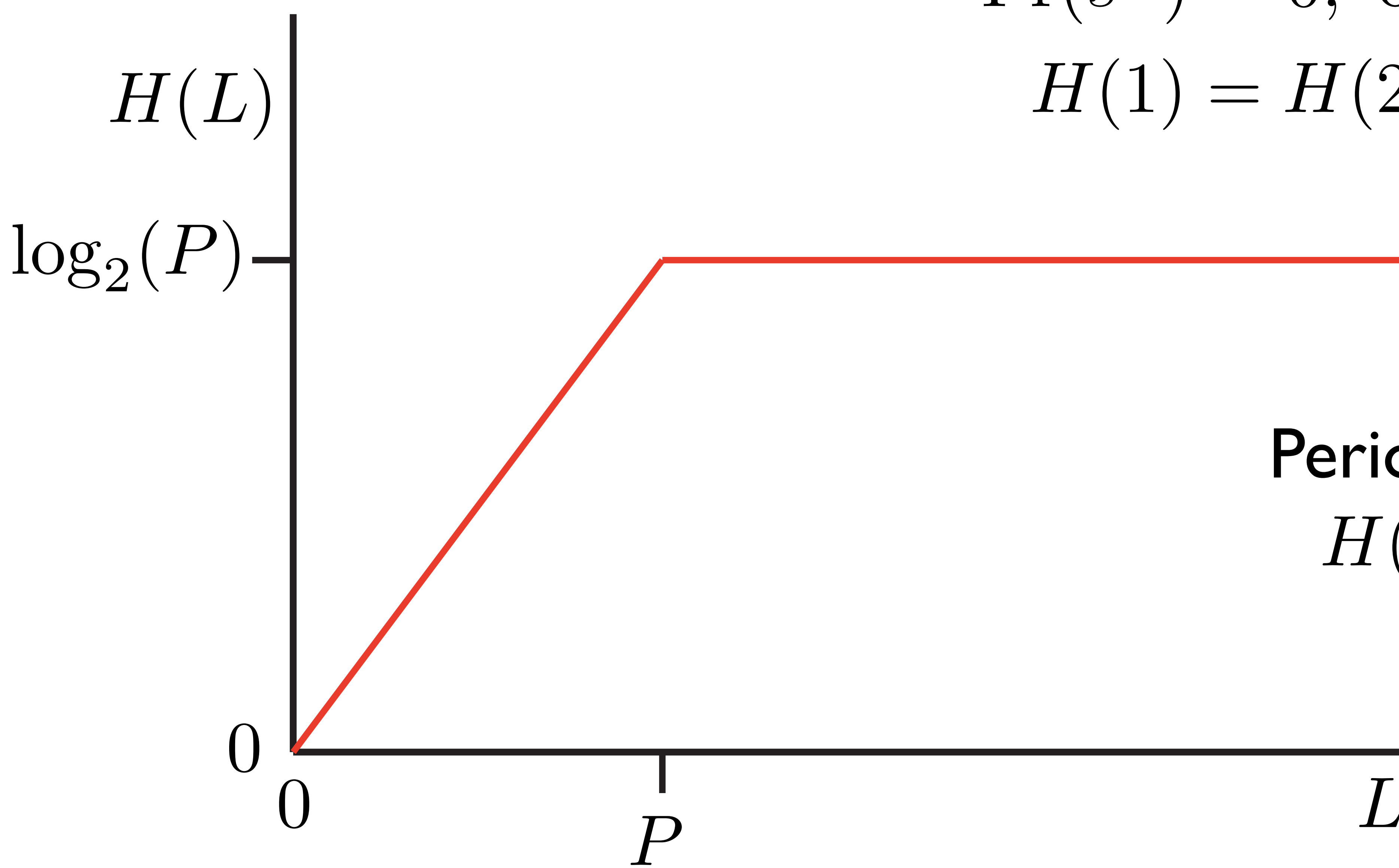
Block Entropy ...

Example: Period-2 Process

$$\ldots 0101010101 \ldots$$

$$\Pr(0) = \Pr(1) = \tfrac{1}{2}$$
$$\Pr(01) = \Pr(10) = \tfrac{1}{2}$$
$$\Pr(101) = \Pr(010) = \tfrac{1}{2}$$
$$\vdots$$
$$\Pr(s^L) = 0, \text{ otherwise}$$

$$H(1) = H(2) = H(L \geq 1) = 1$$



Period-P Process:

$$H(L \geq P) = \log_2(P)$$

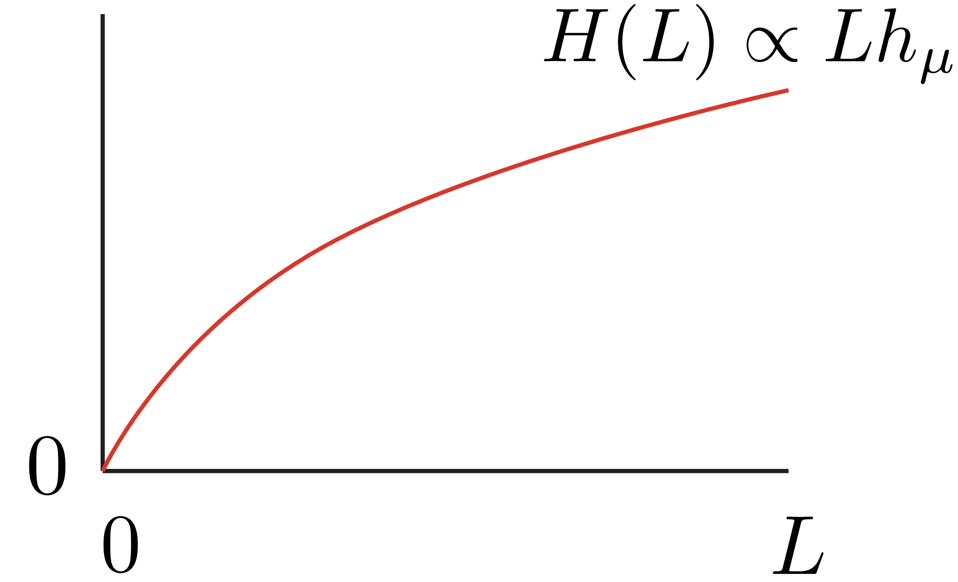
Information in Processes ...

Entropy Rates for Stationary Stochastic Processes:

Entropy per symbol is given by the Source Entropy Rate:

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

(When limits exists.)



Interpretations:

- Asymptotic growth rate of entropy
- Irreducible randomness of process
- Average description length (per symbol) of process

Use: Typical sequences have probability: $\Pr(s^L) \approx 2^{-L \cdot h_\mu}$

(Shannon-MacMillian-Breiman Theorem)

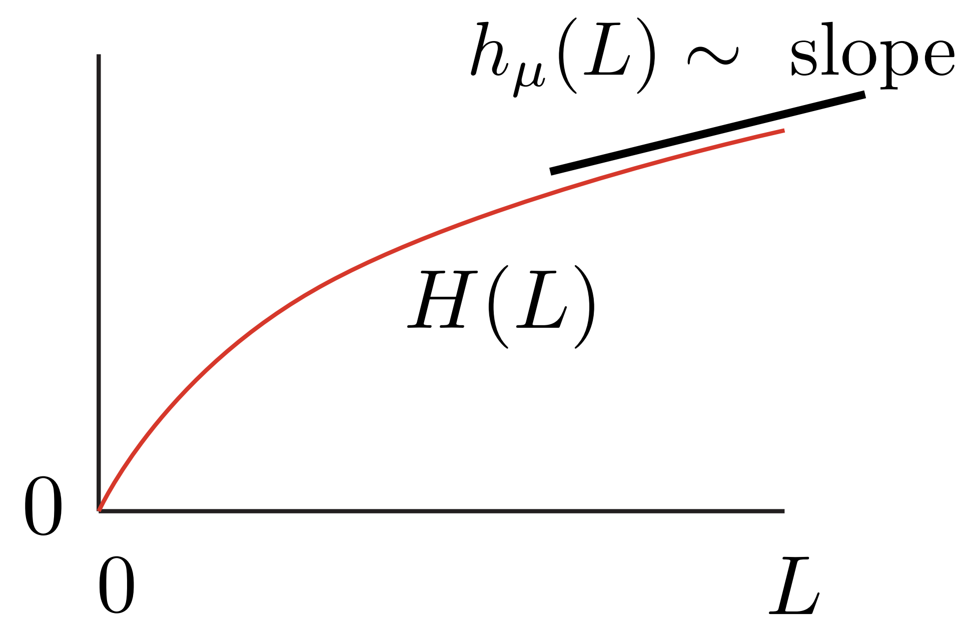
Information in Processes ...

Entropy Rates for Stationary Stochastic Processes ...

Length-L Estimate of Entropy Rate:

$$\widehat{h}_\mu(L) = H(L) - H(L-1)$$

$$\widehat{h}_\mu(L) = H(s_L | s_1 \cdots s_{L-1})$$



Boundary condition:

$$\widehat{h}_\mu(0) = \log_2 |\mathcal{A}|$$: no measurements, all events possible

$$\widehat{h}_\mu(1) = H(1)$$

Monotonic decreasing: $\widehat{h}_\mu(L) \leq \widehat{h}_\mu(L-1)$

Conditioning cannot increase entropy:

$$H(s_L | s_1 \cdots s_{L-1}) \leq H(s_L | s_2 \cdots s_{L-1}) = H(s_{L-1} | s_1 \cdots s_{L-2})$$

Information in Processes ...

Entropy Rates for Stationary Stochastic Processes:

Entropy rate ...

$$\widehat{h}_\mu = \lim_{L \to \infty} \widehat{h}_\mu(L) = \lim_{L \to \infty} H(s_0 | \overleftarrow{s}^L) = H(s_0 | \overleftarrow{s})$$

Interpretations:

- Uncertainty in next measurement, given past
- A measure of unpredictability
- Asymptotic slope of block entropy

Alternate entropy rate definitions agree:

$$\widehat{h}_\mu = h_\mu$$

Information in Processes ...

Entropy Rate for a Markov chain: $\{V, T\}$

$$\begin{aligned} h_\mu &= \lim_{L \to \infty} h_\mu(L) \\ &= \lim_{L \to \infty} H(v_L | v_1 \cdots v_{L-1}) \\ &= \lim_{L \to \infty} H(v_L | v_{L-1}) \end{aligned}$$

Assuming asymptotic state distribution:

Process in statistical equilibrium

Process running for a long time

Forgotten it's initial distribution

Closed-form:

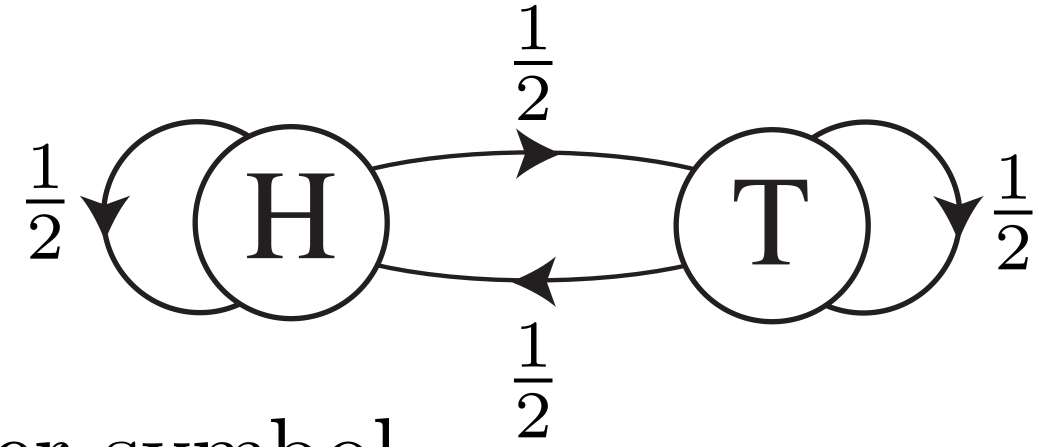
$$h_\mu = -\sum_{v \in V} p_v(\infty) \sum_{v' \in V} T_{vv'} \log_2 T_{vv'}$$

Information in Processes ...

Entropy Rate for Markov chains ...

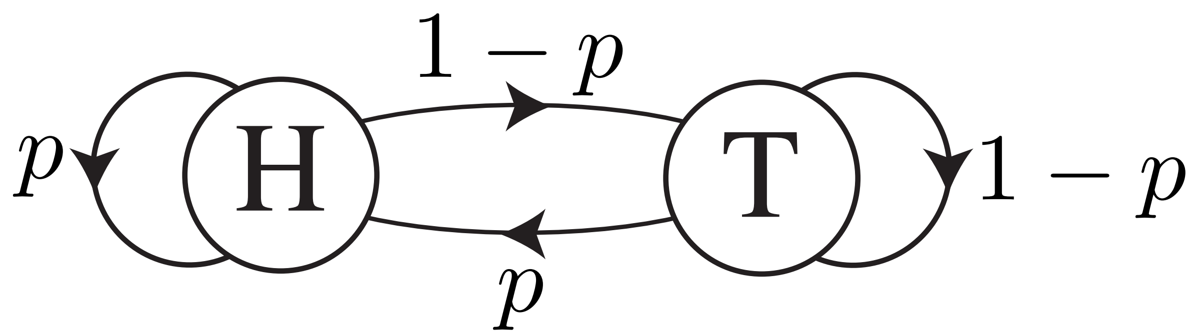
Examples:

(1) Fair Coin:



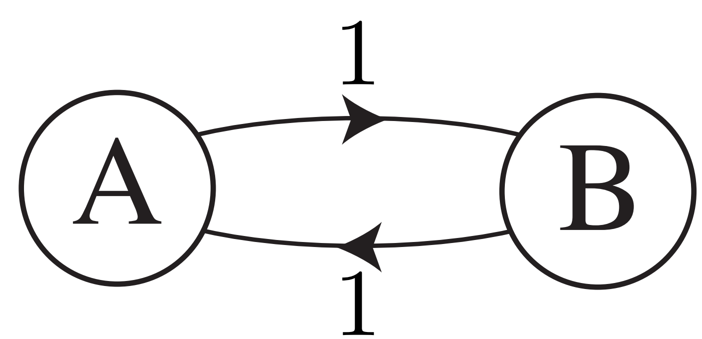
$$h_\mu = 1 \text{ bit per symbol}$$

(2) Biased Coin:



$$h_\mu = H(p) \text{ bits per symbol}$$

(3) Period-2 Process:



$$h_\mu = 0 \text{ bits per symbol}$$

Entropy Rate for Unifilar Hidden Markov Chain:

Internal: $\{V, T\}$
Observed: $\{T^{(s)} : s \in \mathcal{A}\}$

Closed-form for entropy rate:

$$h_\mu = - \sum_{v \in V} p_v(\infty) \sum_{s \in \mathcal{A}} \sum_{v' \in V} T^{(s)}_{vv'} \log_2 T^{(s)}_{vv'}$$

Due to unifilarity:
Observed sequences are (effectively) unique paths in UHMC

Entropy Rate for Unifilar Hidden Markov Chain ...

Example: Why are modems noisy?

Recall previous prefix code example

Distribution:
$$
\begin{aligned}
p(a) &= \tfrac{1}{2} \\
p(b) &= \tfrac{1}{4} \\
p(c) &= \tfrac{1}{8} \\
p(d) &= \tfrac{1}{8}
\end{aligned}
$$

$$H(X) = 1.75 \text{ bits}$$

Codebook:
$$
\begin{aligned}
C(a) &= 0 \\
C(b) &= 10 \\
C(c) &= 110 \\
C(d) &= 111
\end{aligned}
$$

$$R(C) = 1.75 \text{ bits per message}$$

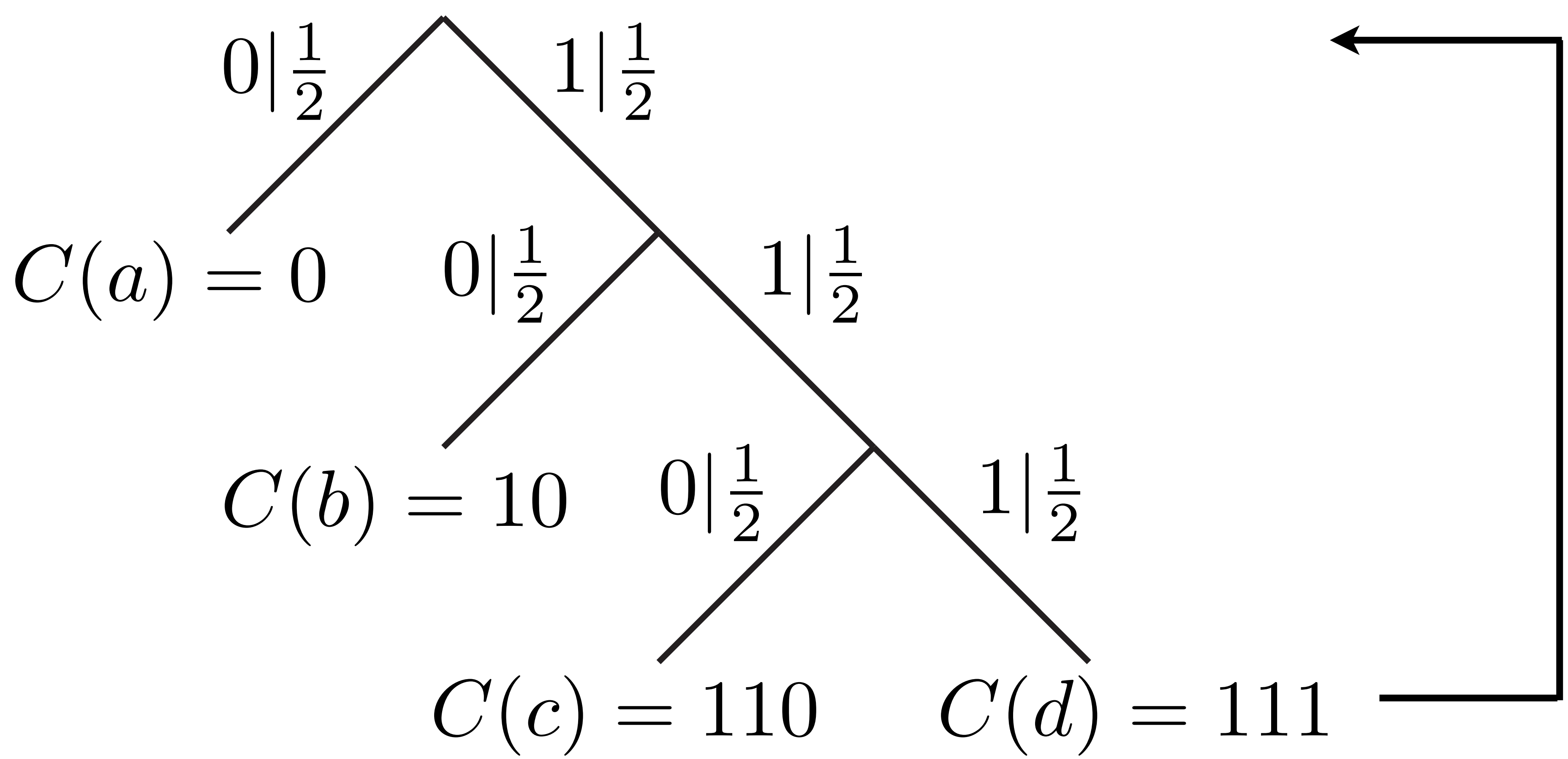
What is entropy rate (per output bit) of encoded stream?

Information in Processes ...

Entropy Rate for Deterministic Hidden Markov Chain ...

Example: Why are modems noisy?

How often are codewords generated?



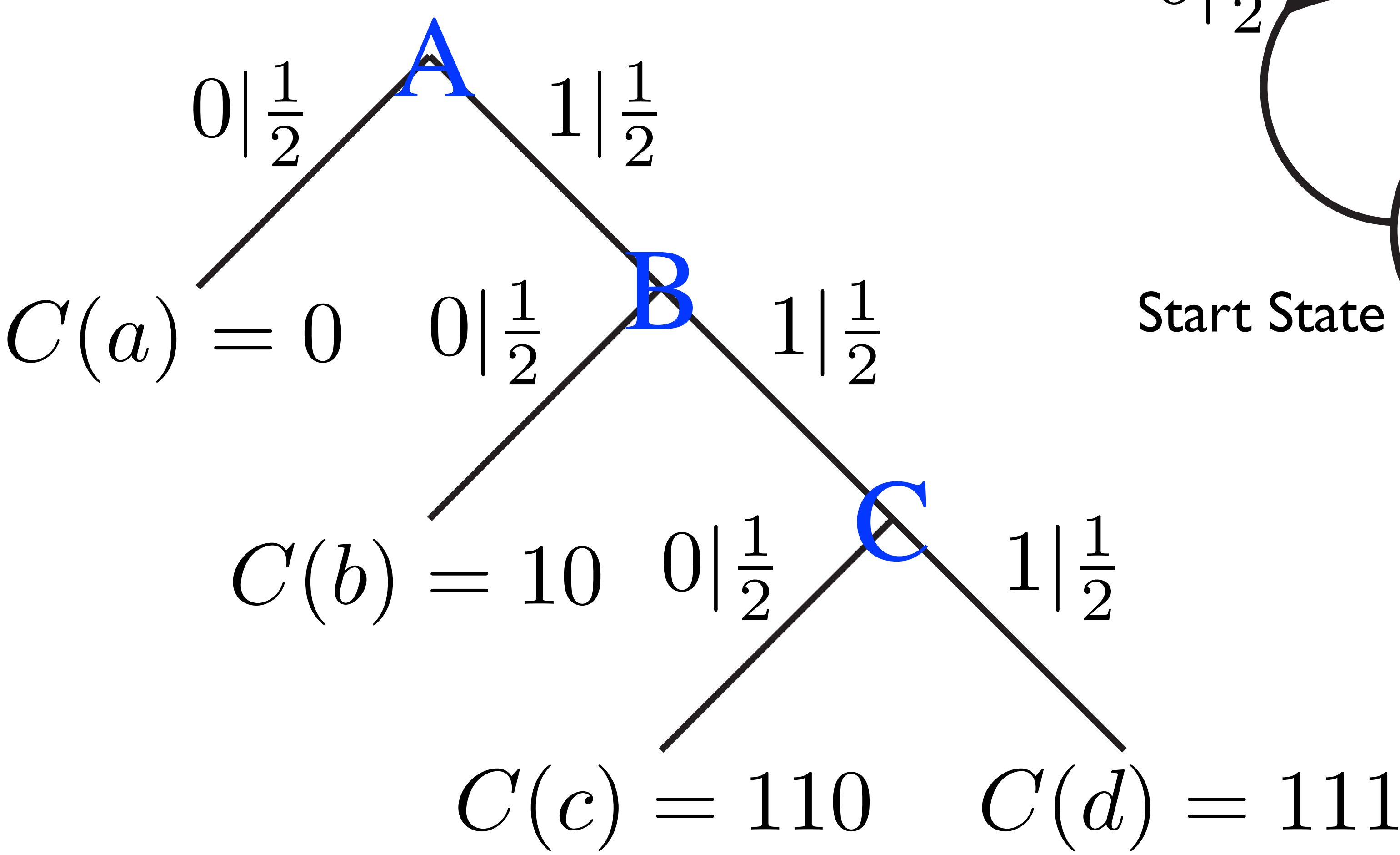
Encoding (output of channel) is a hidden Markov chain:

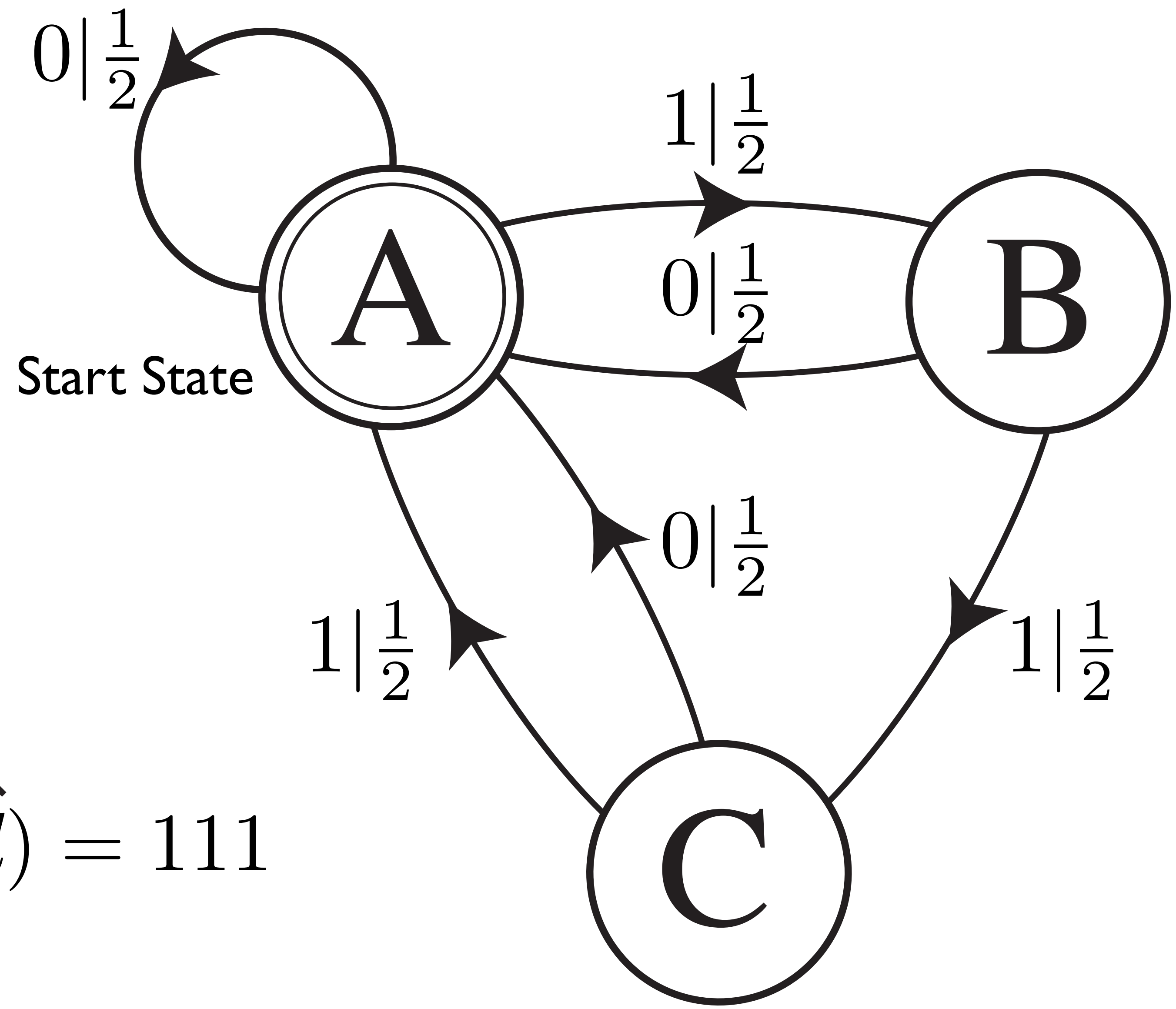
Leaves connect to top tree node

Entropy Rate for Deterministic Hidden Markov Chain ...

Example: Why are modems noisy?

Identify tree nodes with states of a hidden Markov chain





Entropy Rate for Deterministic Hidden Markov Chain ...

Example: Why are modems noisy?

Equivalent hidden Markov chain

$$T = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 \end{pmatrix}$$

$$p_V(\infty) = (p_A, p_B, p_C) = (\tfrac{4}{7}, \tfrac{2}{7}, \tfrac{1}{7})$$

It's unifilar:

$$T^{(0)} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{pmatrix} \qquad T^{(1)} = \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 \end{pmatrix}$$

Entropy Rate for Deterministic Hidden Markov Chain ...

Example: Why are modems noisy?

Calculate entropy rate directly:

$$
\begin{aligned}
h_\mu &= - \sum_{v \in V} p_v(\infty) \sum_{s \in \mathcal{A}} \sum_{v' \in V} T^{(s)}_{vv'} \log_2 T^{(s)}_{vv'} \\
&= \tfrac{4}{7} \cdot 1 + \tfrac{2}{7} \cdot 1 + \tfrac{1}{7} \cdot 1 \\
&= 1 \text{ bit}
\end{aligned}
$$

Encoding provides *full* utilization of binary channel.

Entropy Rate for Deterministic Hidden Markov Chain ...

Example: Why are modems noisy?

Compare:

4-symbol source is redundant:

$$\begin{aligned} \mathcal{R} &= \log_2 |\mathcal{A}| - H(X) \\ &= 2 - 1.75 = 0.25 \text{ bits} \end{aligned}$$

Does not use all of 4-symbol channel.

Prefix code mapped 4-symbol, suboptimal source
into a new source that uses all available capacity.

Modems do the same: Maximize use of capacity by sending a
code stream that is as close to maximum entropy as possible.

Entropy Rate for Nonunifilar Hidden Markov Chain:

Internal: $\{V, T\}$
Observed: $\{T^{(s)} : s \in \mathcal{A}\}$

Entropy rate: No closed-form!

$$h_\mu \neq - \sum_{v \in V} \pi_v \sum_{s \in \mathcal{A}} \sum_{v' \in V} T^{(s)}_{vv'} \log_2 T^{(s)}_{vv'}$$

Upper and Lower Bounds:

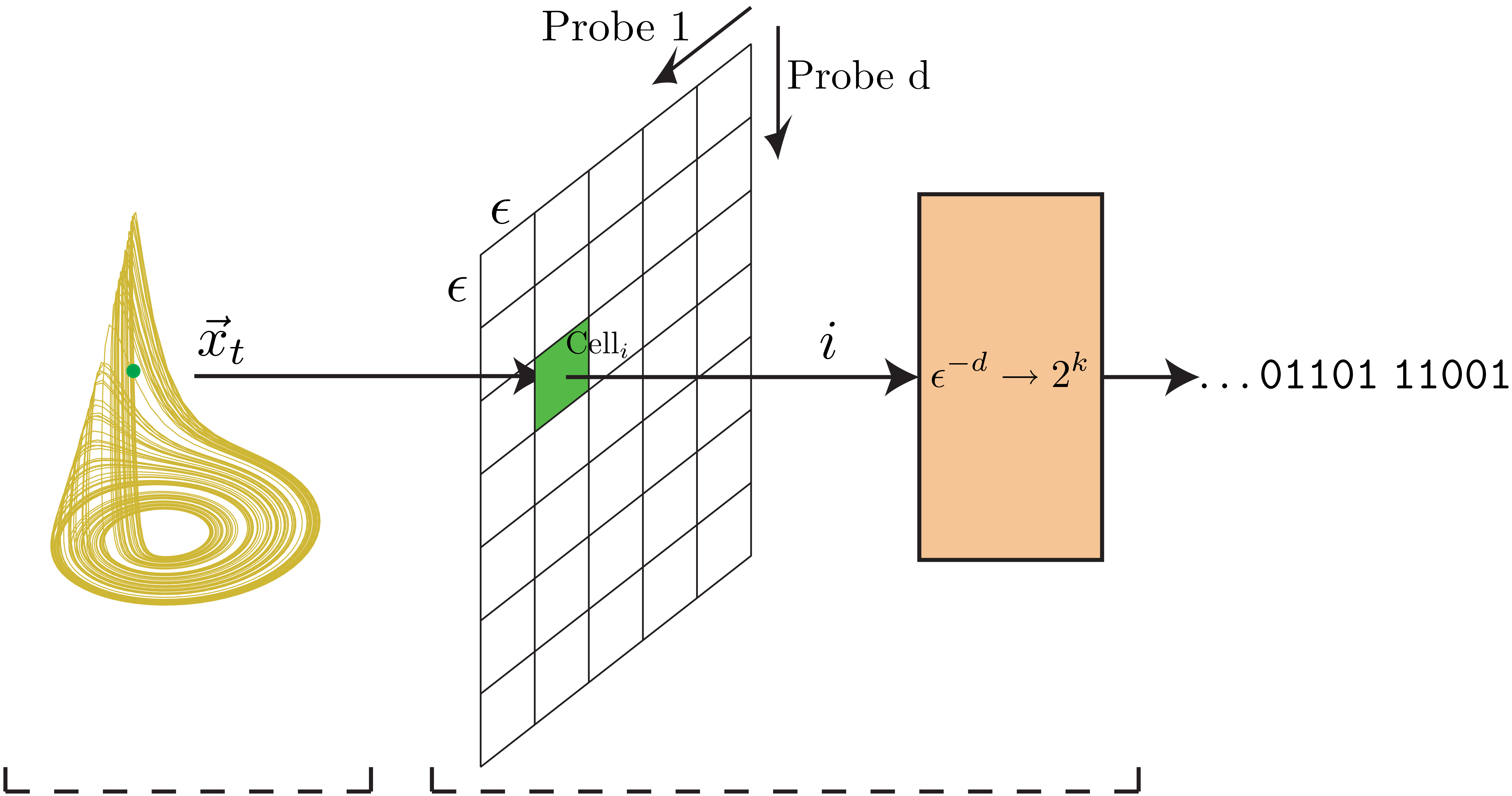
$$H(S_L | V_1 S_1 \cdots S_{L-1}) \leq h_\mu(L) \leq H(S_L | S_1 \cdots S_{L-1})$$

Unrealistic for inference: Must know about internal states.
Unrealistic for analysis: Simulate chain, do empirical estimate.

Entropy rate? But there exists a way ... stay tuned!

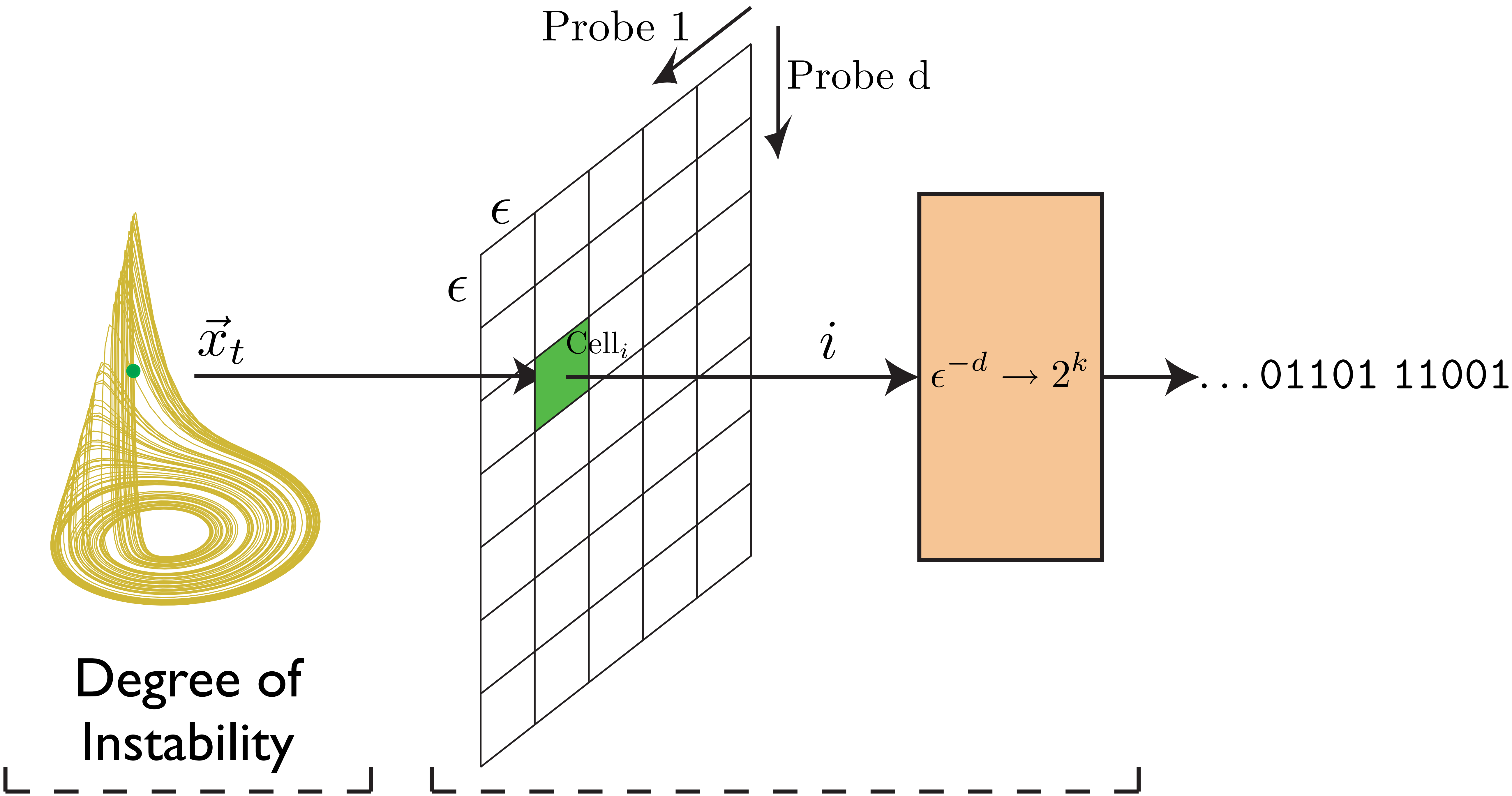
What is the connection between information in processes and chaotic dynamical systems?

Information in Processes ...

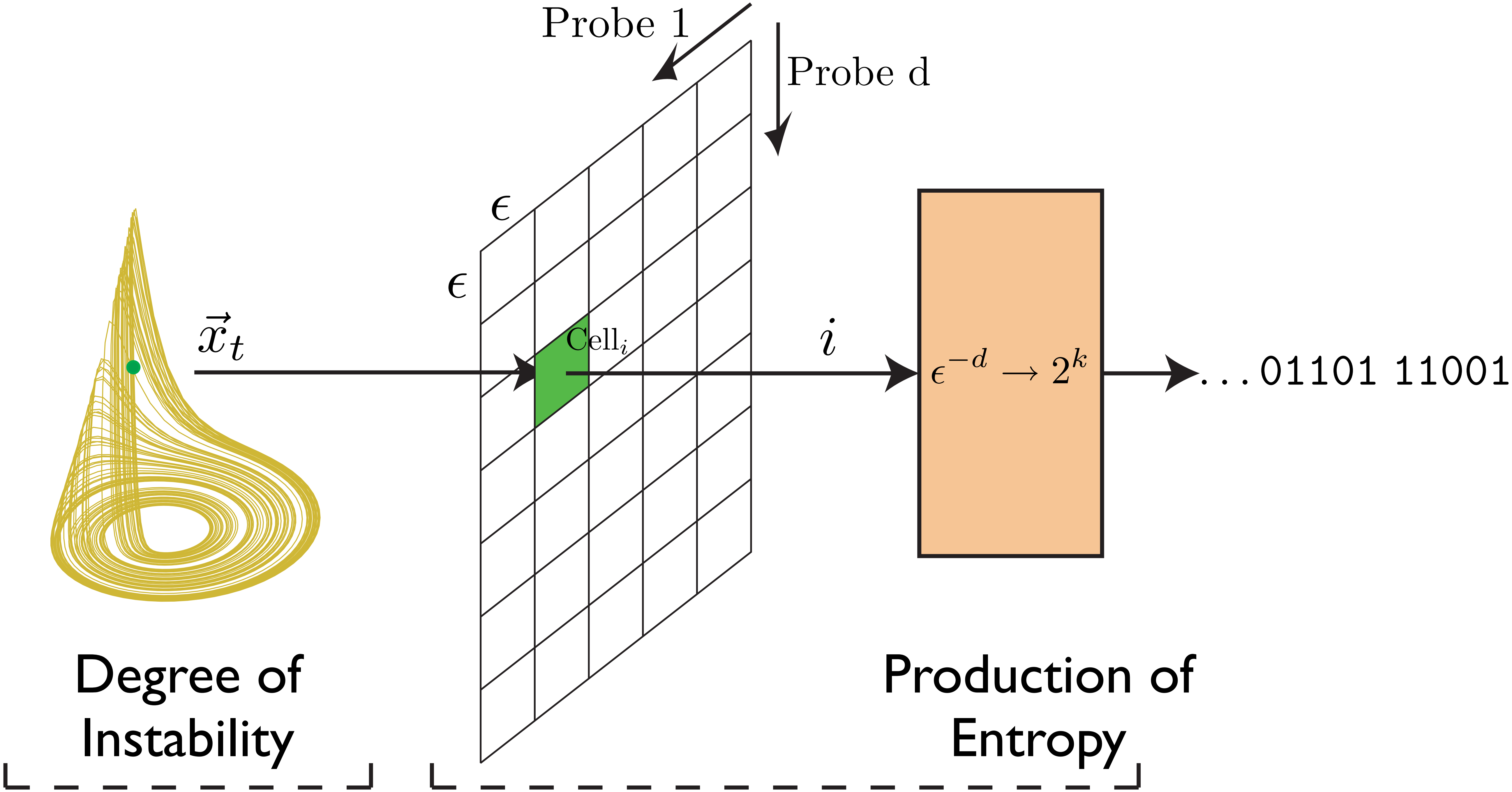


Lecture 16: Natural Computation & Self-Organization, Physics 256A (Winter 2014); Jim Crutchfield

Information in Processes ...

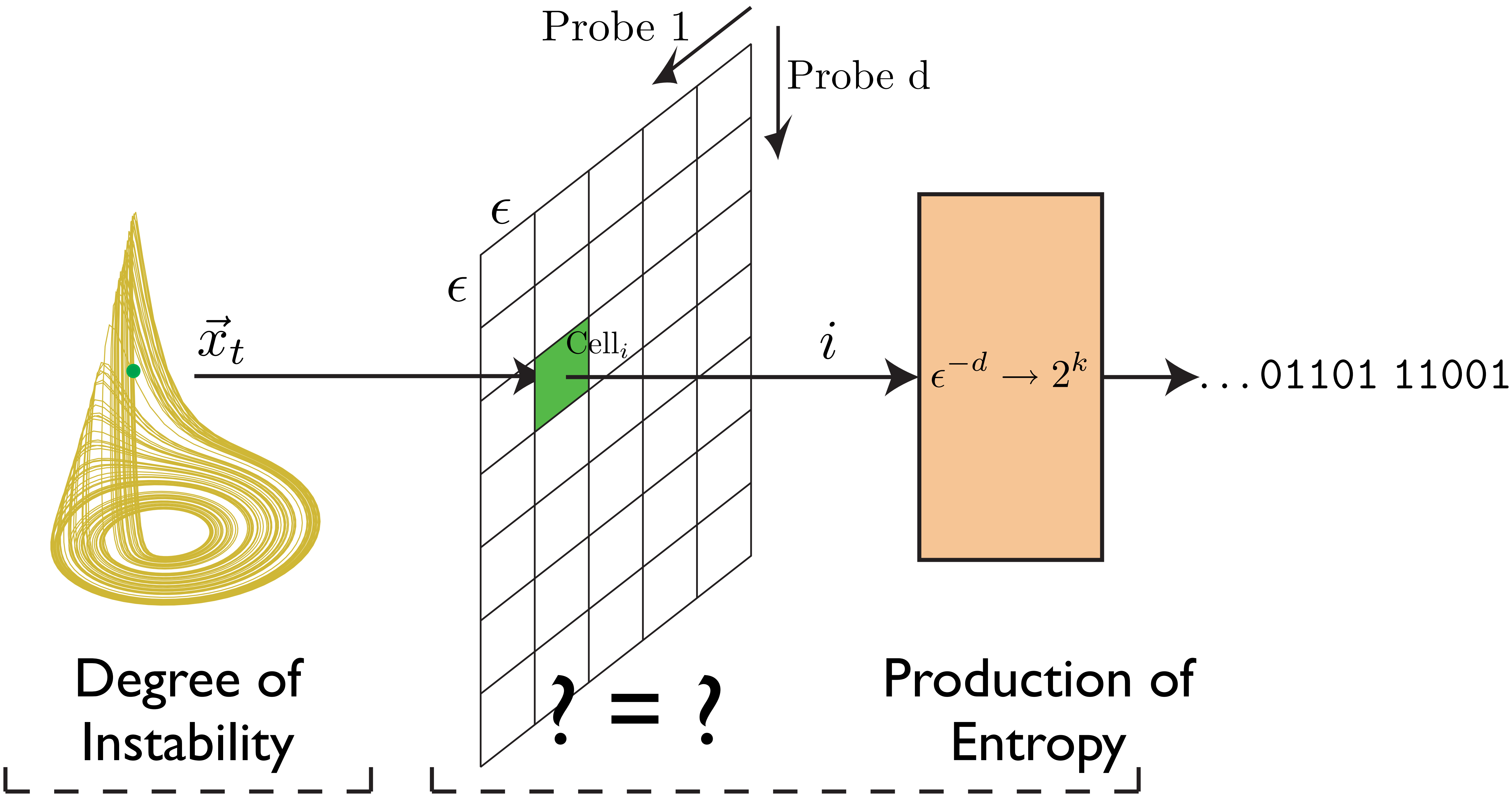


Information in Processes ...



Lecture 16: Natural Computation & Self-Organization, Physics 256A (Winter 2014); Jim Crutchfield

Information in Processes ...



Symbolic dynamics revisited ...

When are partitions good?

When symbol sequences encode orbits:

$$
\begin{array}{ccc}
M & \xrightarrow{\mathcal{T}} & M \\
\Delta \uparrow & & \Delta \uparrow \\
\mathcal{A}^{\mathbb{Z}} & \xrightarrow{\sigma} & \mathcal{A}^{\mathbb{Z}}
\end{array}
$$

Diagram commutes:

$$\mathcal{T}(x) = \Delta \circ \sigma \circ \Delta^{-1}(x)$$

Good kinds of instruments:

Markov partitions

Generating partitions

Metric entropies and mixing:

Entropy of partition:

$$H(\mathcal{P}) = -\sum_{i=1}^{k} \Pr(\mathcal{P}_i) \log_2 \Pr(\mathcal{P}_i)$$

Metric entropy given partition:

$$h_\mu(f, \mathcal{P}) = \lim_{N \to \infty} \frac{1}{N} H\left(\bigvee_{n=0}^{N} f^{-n}(\mathcal{P}) \right)$$

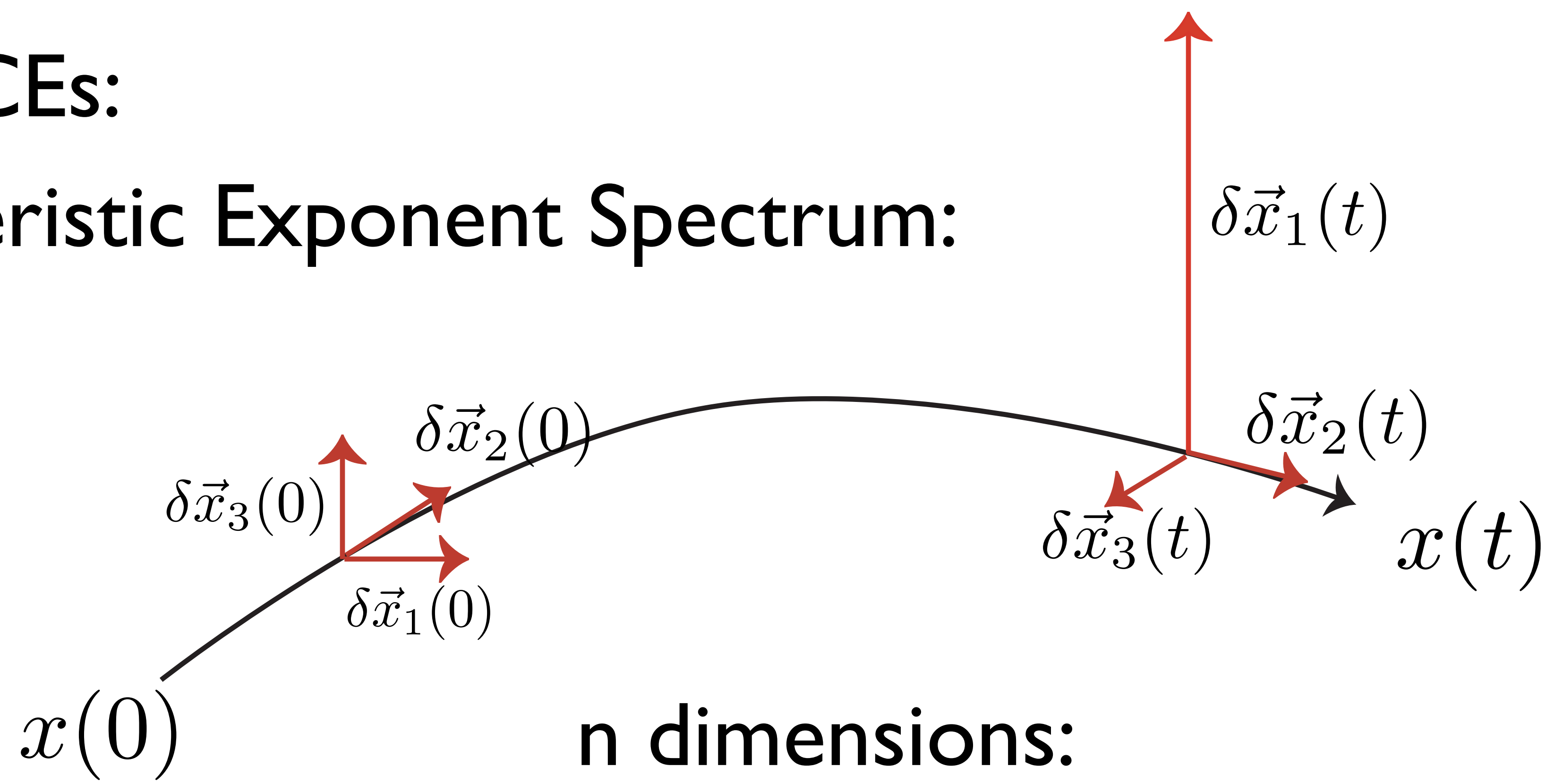
How well partition cells
are mixed together
=
Production of entropy

Theorem: $h_\mu(f) = \sup_{\mathcal{P}} h_\mu(f, \mathcal{P})$

Corollary:
Using generating partition, metric entropy of symbolic dynamics is that of the hidden dynamical system $h_\mu(f) = h_\mu(f, \mathcal{P})$.

Entropy rate and LCEs:

Lyapunov Characteristic Exponent Spectrum:



n dimensions:

$$\chi = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}, \;\; \lambda_i \geq \lambda_{i+1}$$

$$\lambda_i = \lim_{t \to \infty} \lim_{||\delta\vec{x}_i(0)|| \to 0} \frac{1}{t} \log_2 \frac{||\delta\vec{x}_i(t)||}{||\delta\vec{x}_i(0)||}$$

$$\{\delta\vec{x}_1, \delta\vec{x}_2, \ldots, \delta\vec{x}_n\}, \delta\vec{x}_i \cdot \delta\vec{x}_j = 0, i \neq j$$

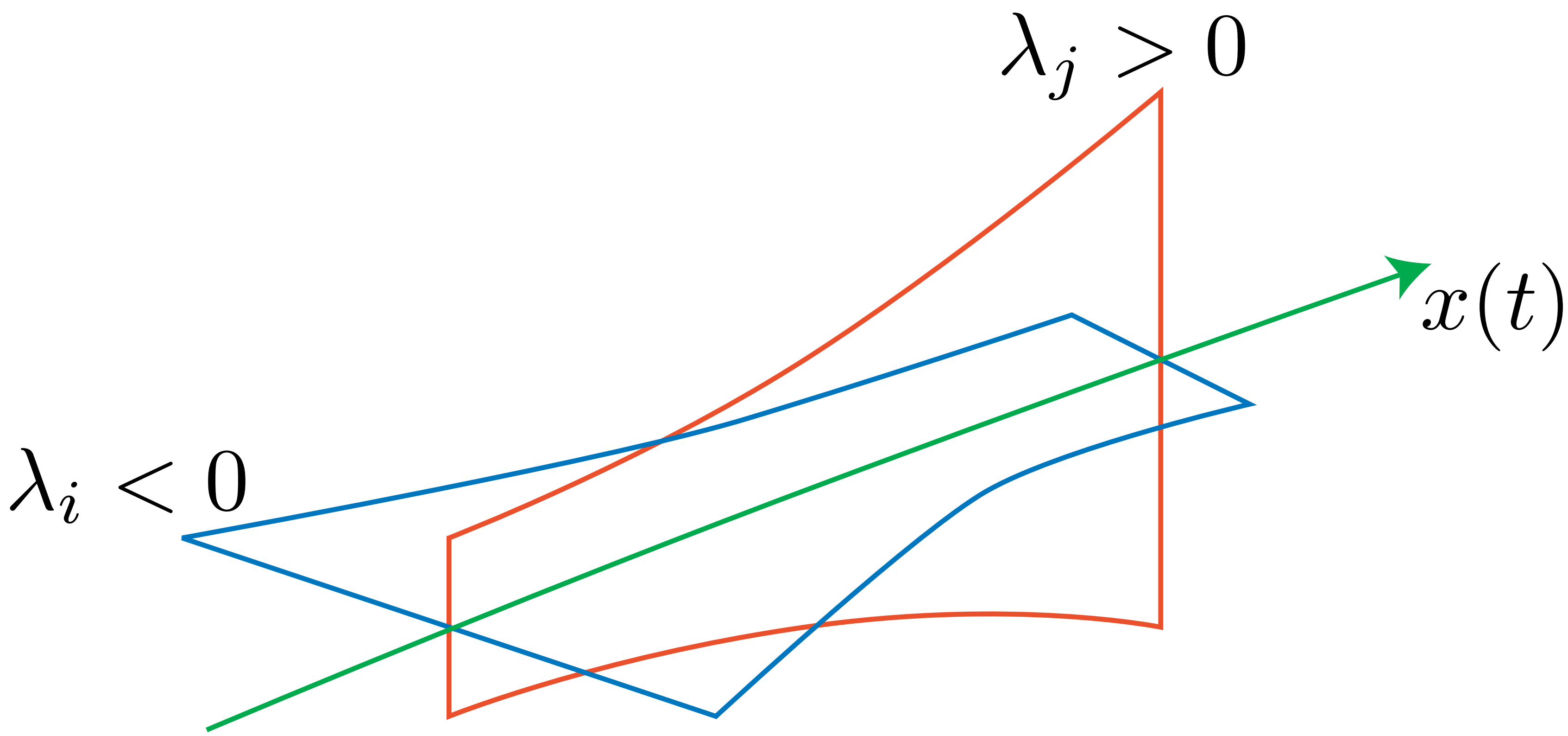
Information Dynamics ...

Entropy rate and LCEs:

LCE Spectrum gives "Geometry" of Submanifolds:

$$\lambda_i < 0 \iff \text{stable manifold}$$

$$\lambda_i > 0 \iff \text{unstable manifold}$$



Entropy rate and LCE Spectrum:

$$h_\mu = \sum_{\lambda_i > 0} \lambda_i$$

Rate of information production:
Relate a geometric property (LCE spectrum) to
how well subsets are mixed into each other (entropy rate).

Concrete statement of how a continuous-state dynamical system is an information source.

Dynamics and information theory are intimately related.

Information Dynamics ...

Ergodic Hierarchy:

Bernoulli system:
Most random

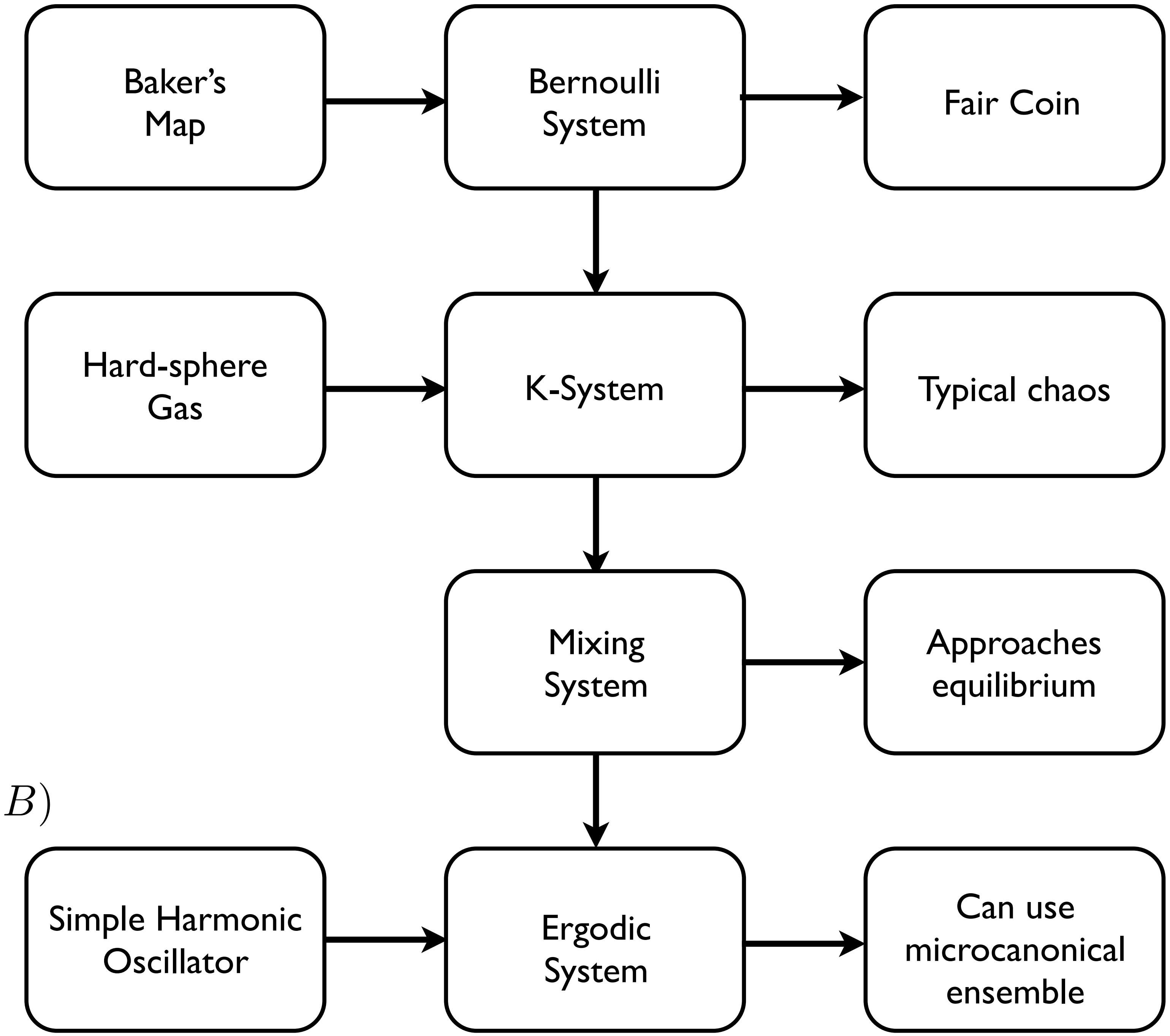
Kolmogorov system:
Present asymptotically independent of distant past

Mixing system:
Subsets mixed

$$\lim_{n \to \infty} \Pr\left(A \cap f^{-n}(B)\right) = \Pr(A) \Pr(B)$$

Ergodic system:
Time- & state- averages equal

$$\Pr(A) = |A|^{-1}$$



(See [MET] in CMech Readings)

Information in Processes ...

Reading for next lecture:

CMR article RURO.