

Optimal predictive inference

Susanne Still

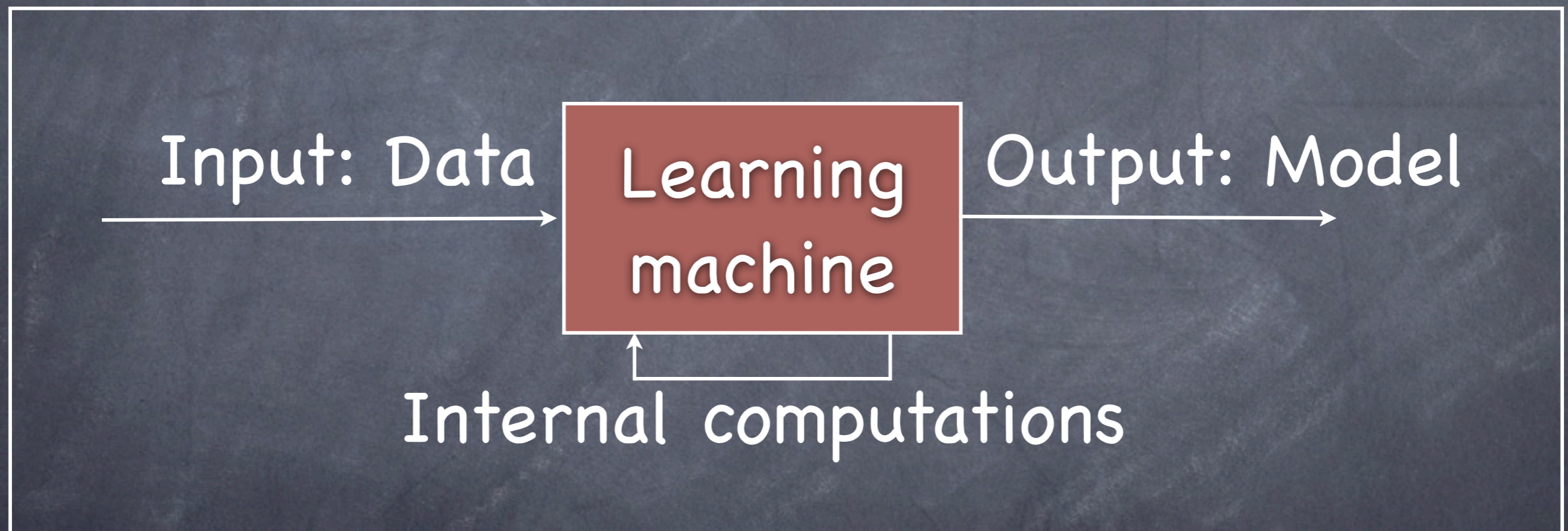
University of Hawaii at Manoa
Information and Computer Sciences

S. Still, J. P. Crutchfield. Structure or Noise? <http://lanl.arxiv.org/abs/0708.0654>

S. Still, J. P. Crutchfield, C. J. Ellison. Optimal Causal Inference.

<http://lanl.arxiv.org/abs/0708.1580>

The modeling (and machine learning) challenge:

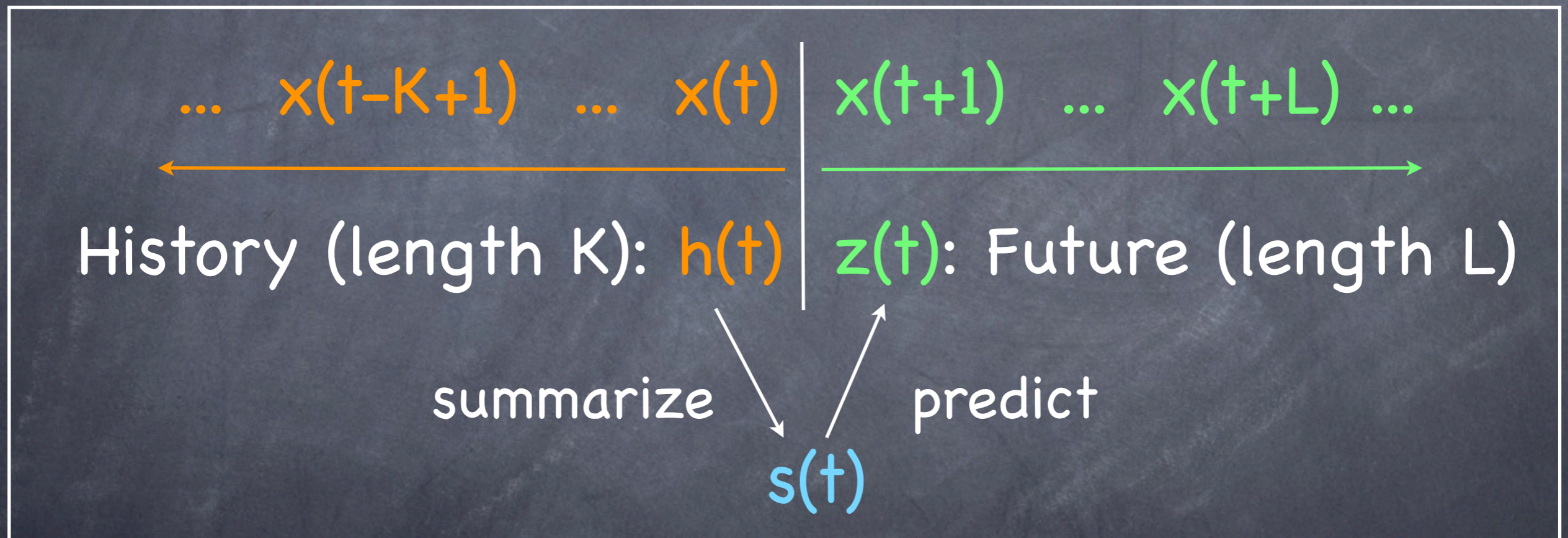


Fundamental question:
What is a good model?

Intuition behind the approach we take

1. A good model predicts well
 - > find maximally predictive model.
Keep predictive information!
2. A good model is compact
 - > do not keep irrelevant information.

- **System** → Produces observable $x(t)$
- **Observer**: Has access to a history $h(t)$ of length K .
Makes a representation of the system by creating internal states $s(t)$.



- What information do we need to keep for prediction?
- Simplifying assumptions: discrete time; finite states.

Making a model

- Natural processes execute computations and produce information.
- The “predictive information”, or “excess entropy” measures the information that the past carries about the future.

$$I_{\text{pred}} = \left\langle \log \frac{p(\text{future}; \text{past})}{p(\text{future})p(\text{past})} \right\rangle$$

Quantifying the intuition

Objectives

Good generalization

\Leftrightarrow Good prediction

Minimal coding cost
(complexity)

Quantified by

Predicted Information

$$I[s; z]$$

s: internal state at time t

z: future starting at t

Coding Rate

$$I[s; h]$$

h: history up to time t

Predicted Information: information model carries about future

$$I[s;z] = H[z] - H[z|s]$$

- Measures the reduction in uncertainty about the future, when the model state is known.
- If the state tells us nothing about future, then $H[z|s] = H[z]$ and $I[s;z] = 0$
- If knowing the state reduces the uncertainty about the future to $H[z|s] = 0$, then I is maximal: $I[s;z] = H[z]$.

Coding rate = historical information that is retained by the model:

$$I[s;h] = H[s] - H[s|h]$$

- $H[s]$: measures the "statistical complexity" of the model (Crutchfield and Young, '89). Computational mechanics.
- $H[s|h]$: measures how uncertain we are about whether the history h should be represented by the state s .
- Maximum entropy solution (Rose, 1990): $\max H[s|h]$
- $\min I[s;h] \rightarrow$ prefers model with minimal statistical complexity **and** maximal entropy.
- ($H[s|h] = 0$ for deterministic maps – the case in computational mechanics. Then $I[s;h] = H[s]$)

Making a model

- Finite state machine with internal states s
- Probabilistic map from input space to internal states: $p(s|h)$
- Together with prediction from model state s
 $p(z|s) = \langle p(z|h)p(h|s) \rangle$
- **Objective:** Construct s such that it is a maximally **predictive** and **efficient summary** of historical information.
- Find the optimal probabilistic map $p(s|h)$.

Optimization principle:

$$\max_{p(s|h)} (I[s; z] - \lambda I[s; h])$$

This is rate-distortion theory!!!

- Maps directly onto the “Information Bottleneck Method” (N. Tishby, F. Pereira and W. Bialek (1999) <http://lanl.arxiv.org/abs/physics/0004057>)
- past = data to compress
- future = relevant quantity

- Note: Objective function is equivalent to constructing the states such that they implement "causal shielding", a property of the causal states in computational mechanics.

$$\begin{aligned}
 I[z; h|s] &= \left\langle \log \left[\frac{p(z; h|s)}{p(z|s)p(h|s)} \right] \right\rangle_{p(z,h,s)} \stackrel{\text{use Markov condition, } p(z|h;s) = p(z|h)}{=} \left\langle \log \left[\frac{p(z|h)p(h|s)}{p(z|s)p(h|s)} \right] \right\rangle_{p(z,h,s)} \\
 &= \left\langle \log \left[\frac{p(z|h)}{p(z|s)} \right] \right\rangle_{p(z,h,s)} = \left\langle \left(\log \left[\frac{p(z|h)}{p(z)} \right] - \log \left[\frac{p(z|s)}{p(z)} \right] \right) \right\rangle_{p(z,h,s)} \\
 &= \left\langle \log \left[\frac{p(z|h)}{p(z)} \right] \right\rangle_{p(z,h)} - \left\langle \log \left[\frac{p(z|s)}{p(z)} \right] \right\rangle_{p(z,s)} \\
 &= I[z; h] - I[z; s]
 \end{aligned}$$

- Max $I[s; z]$ same as min $I[z; h|s]$, because $I[z; h]$ is const. The optimization principle now reads:

$$\min_{p(s|h)} (I[s; h] + \beta I[z; h|s])$$

- Writing the O.F. this way makes it obvious that this is rate distortion theory

$$\min_{p(s|h)} \left(I[s; h] + \beta \langle d(h, s) \rangle_{p(h, s)} \right)$$

with the distortion function

$$d(h, s) = D_{\text{KL}}[p(z|h) || p(z|s)]$$

- because

$$\begin{aligned} I[z; h|s] &= \left\langle \left\langle \log \left[\frac{p(z; h|s)}{p(z|s)p(h|s)} \right] \right\rangle_{p(z|h, s)} \right\rangle_{p(h, s)} \\ &= \left\langle \left\langle \log \left[\frac{p(z|h)}{p(z|s)} \right] \right\rangle_{p(z|h)} \right\rangle_{p(h, s)} \\ &= \langle D_{\text{KL}}[p(z|h) || p(z|s)] \rangle_{p(h, s)} \end{aligned}$$

Optimization principle:

$$\max_{p(s|h)} (I[s; z] - \lambda I[s; h])$$

Family of solutions:

$$p(s|h) = \frac{p(s)}{Z(h, \lambda)} \exp \left(-\frac{1}{\lambda} D_{KL}[p(z|h) || p(z|s)] \right)$$

Iterative Algorithm

$$p^{(j)}(s|h) = \frac{p^{(j)}(s)}{Z(h, \lambda)} \exp \left(-\frac{1}{\lambda} D_{KL}[p(z|h) || p^{(j)}(z|s)] \right)$$

$$p^{(j+1)}(s) = \sum_h p^{(j+1)}(s|h) p(h)$$

$$p^{(j+1)}(z|s) = \frac{1}{p^{(j+1)}(s)} \sum_h p(z|h) p^{(j)}(s|h) p(h)$$

Blahut-Arimoto type of algorithm (same as
"Information Bottleneck" algorithm).

Converges to local optimum.

Theorem: In the low temperature regime ($\lambda \rightarrow 0$)
the causal state partition is found.

(S. Still, J. P. Crutchfield, C. Ellison. Optimal Causal Inference. arXiv:0708.1580)

- The causal states reflect the underlying states of the system \rightarrow physically **meaningful solution**.

(J. P. Crutchfield and K. Young (1989) PRL 63:105–108)

- Causal states are **unique and minimal sufficient statistics**.

(J. P. Crutchfield and C. R. Shalizi (1999) Phys.Rev.E 59(1): 275–283)

Proof Sketch:

$$\text{Recall: } p(s|h) = \frac{p(s)}{Z(h, \lambda)} \exp \left(-\frac{1}{\lambda} D_{KL}[p(z|h) || p(z|s)] \right)$$

1. In low temp. regime \rightarrow deterministic partition:

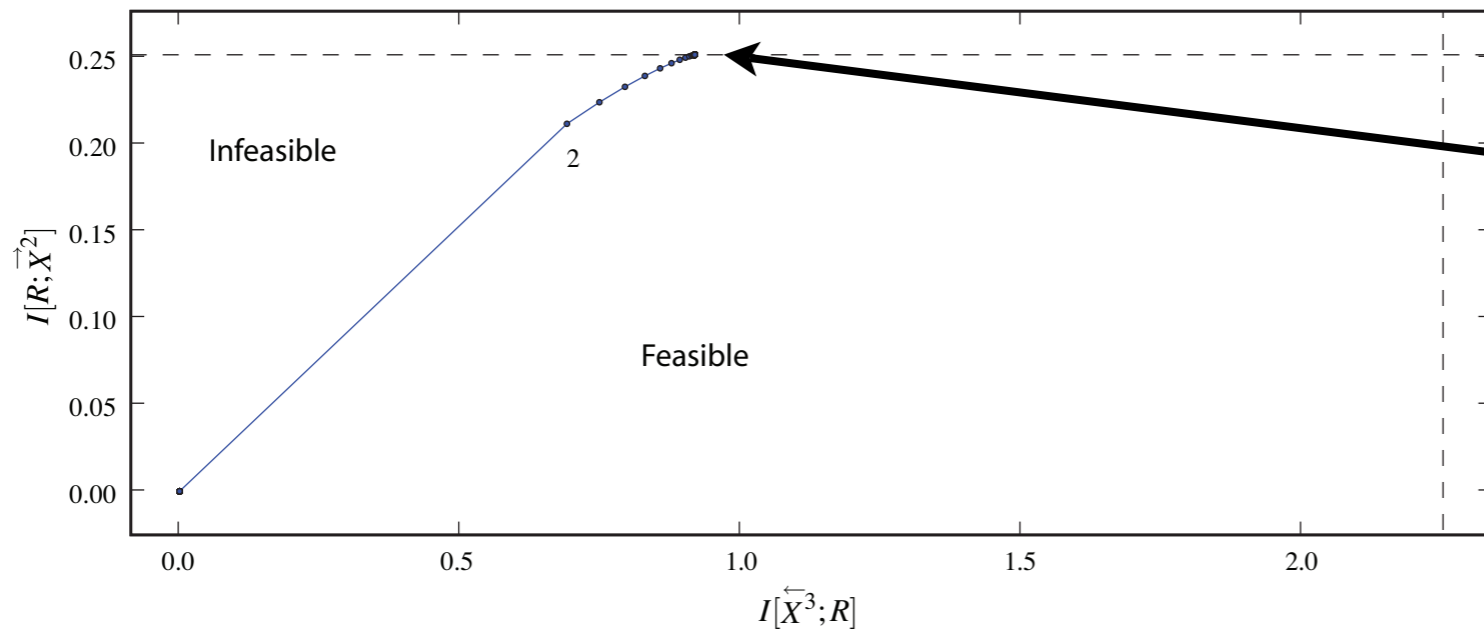
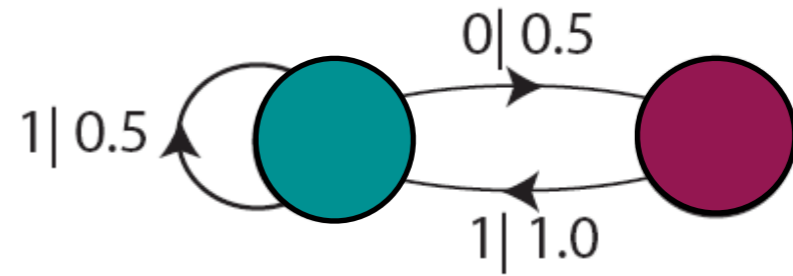
$$p(s|h) = \delta_{ss^*}$$

$$\text{with: } s^* = \arg \min_s D_{KL}[p(z|h) || p(z|s)]$$

2. Histories h with same conditional future distributions, $p(z|h) = p(z|s)$, are assigned to the same category s . This defines an equivalence relation, or probabilistic "bisimulation" (Milner, '84), which is precisely the causal state partition of computational mechanics.

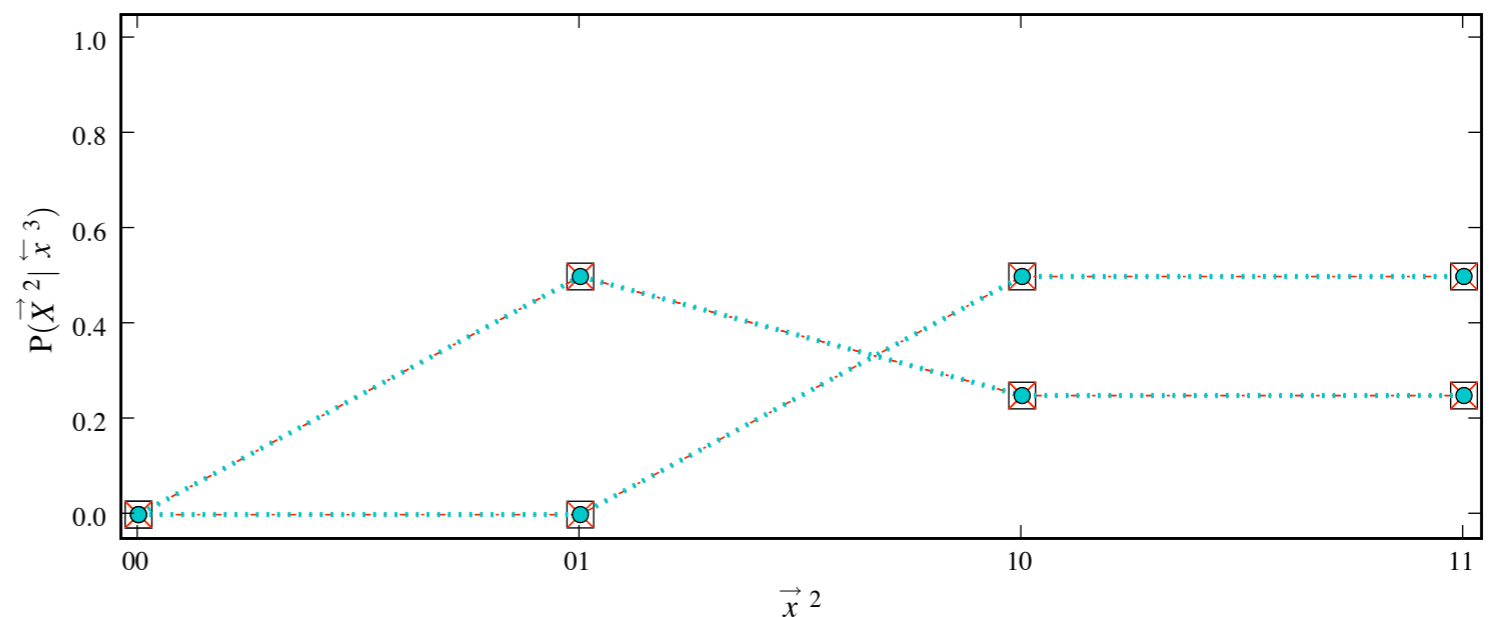
Example 1: Golden Mean

Produces all binary sequences
except those with 00



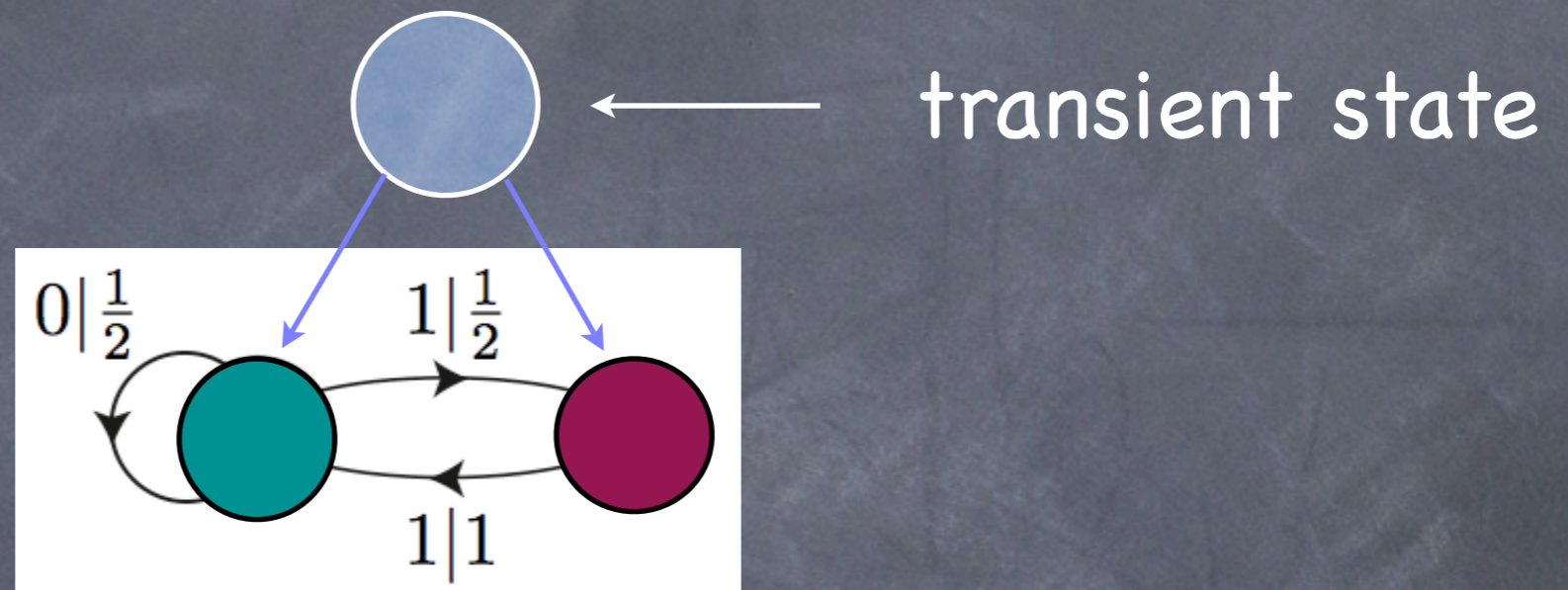
Algorithm finds
the 2 states that
describe the
process fully.

Conditional future
distributions
associated with 2
state model



Example 2: Even Process

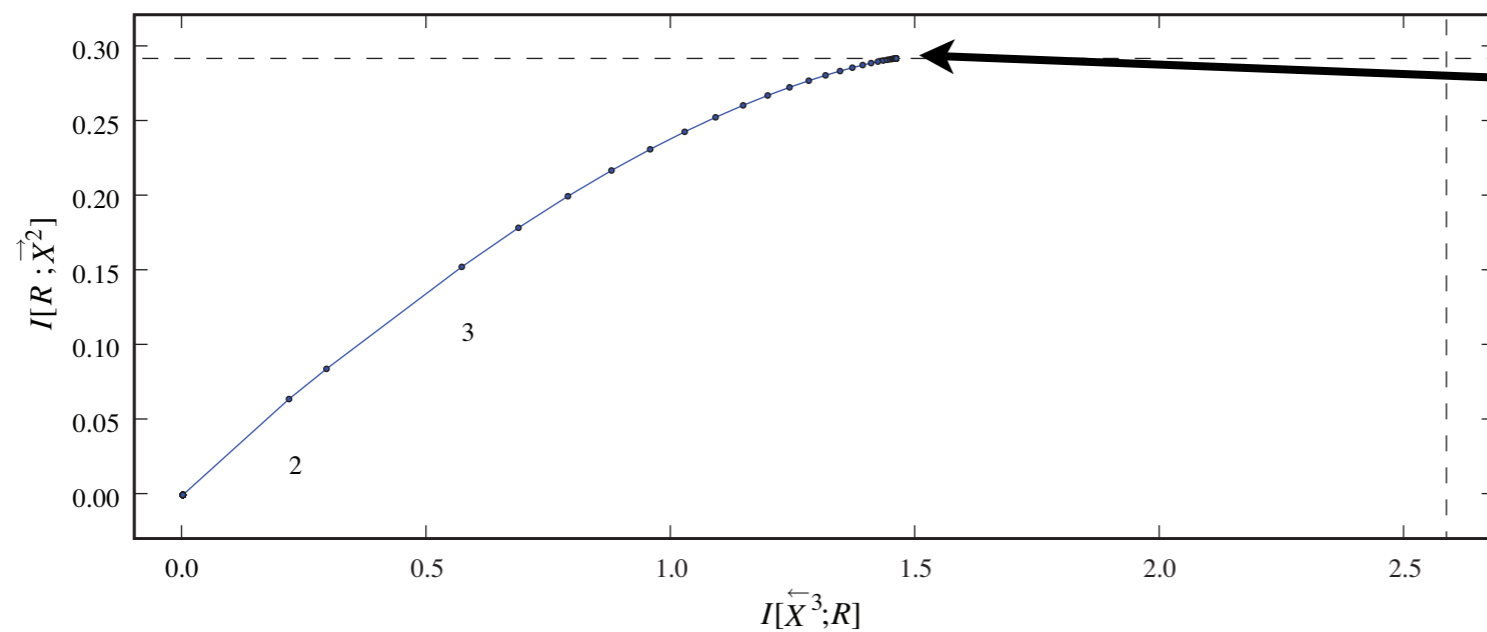
Even blocks of 1s.



- Irreducible forbidden words: $F = \{010, 01110, 011110, \dots\}$
- No finite order markov process can model the even process.

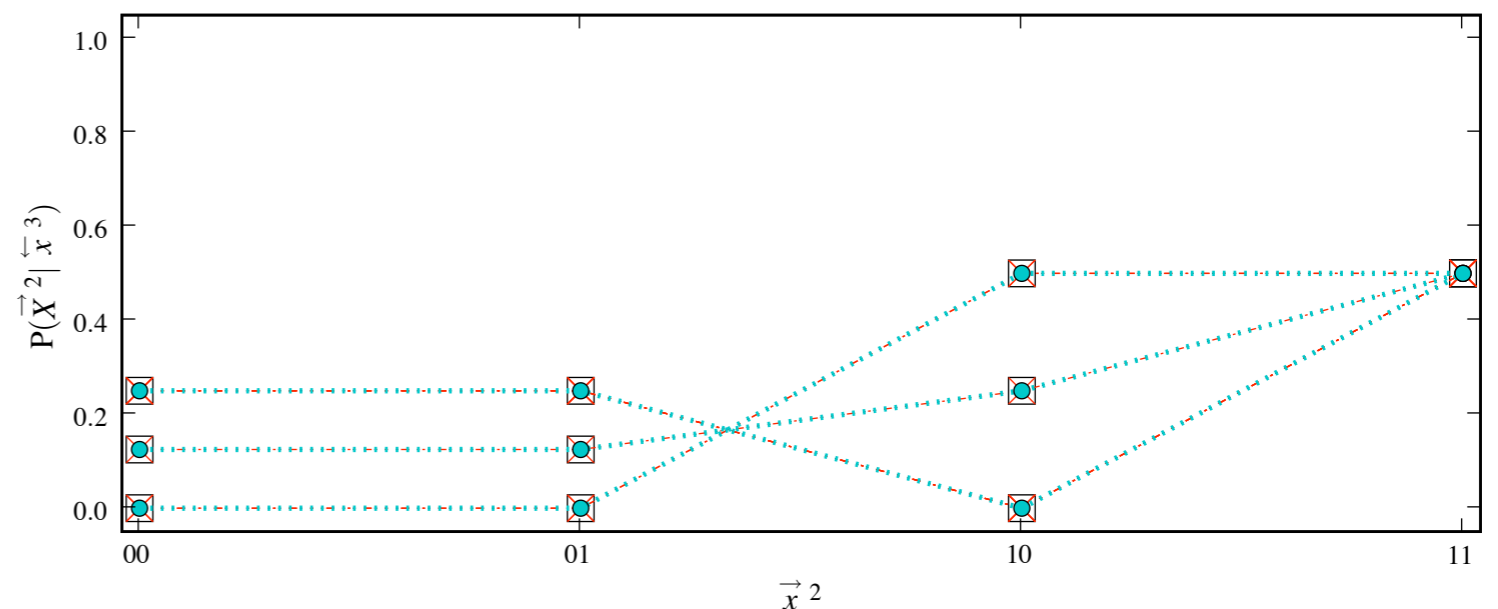
Example 2: Even Process

Generates all sequences with blocks 1s of even length



Algorithm finds the 3 states that describe the process fully.

Conditional future distributions associated with 3 state model



Hierarchy of optimal models

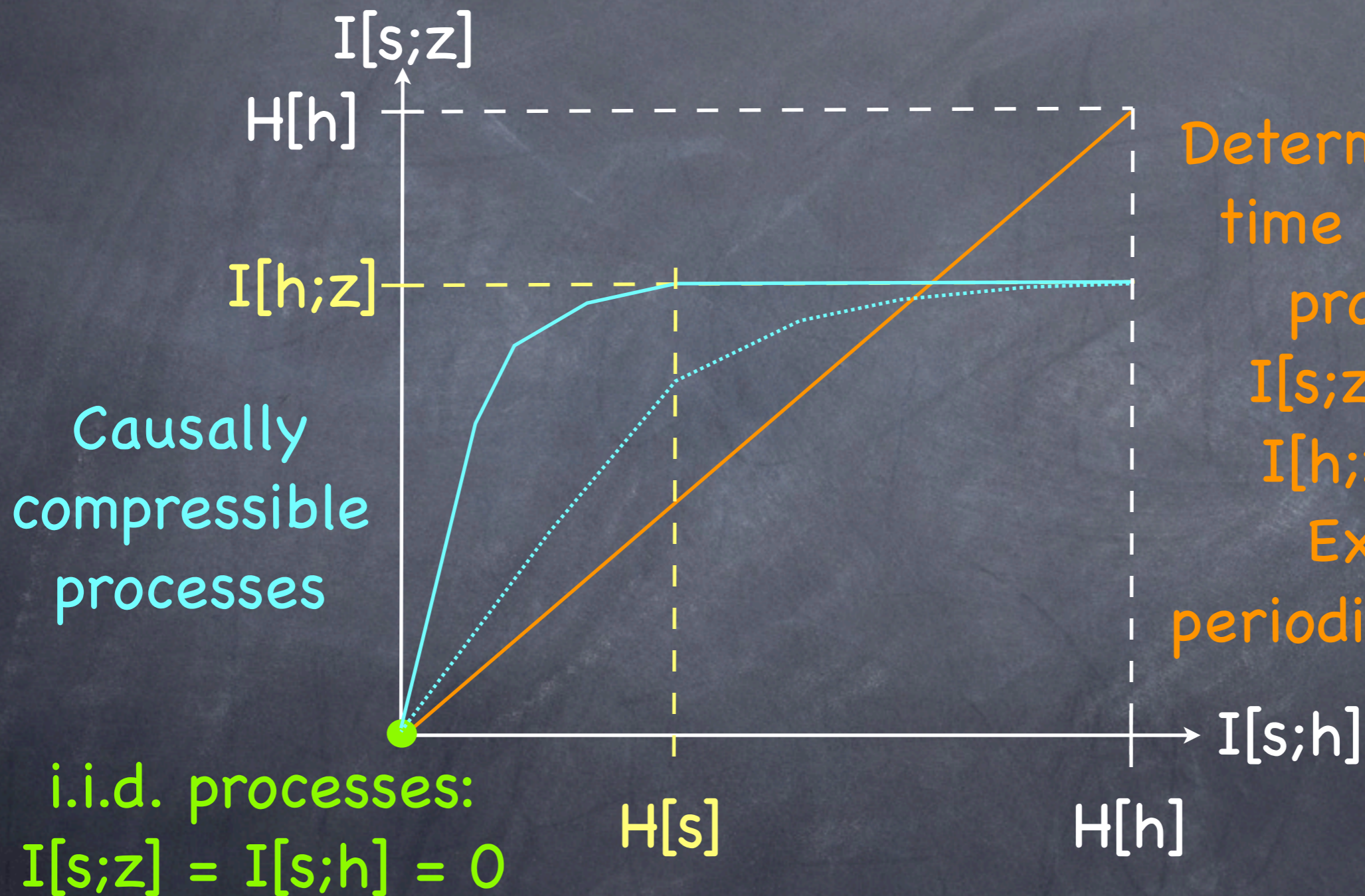
- As temperature $\rightarrow 0$, we find causal states. Those capture the **full** predictive information!
- But, in general we may not want to keep all detail!
- We can find more compact models. Those have larger prediction error.
- Compared to computational mechanics, here we have an extension to non-deterministic models.

Causal compressibility

- Study the full range of optimal models to learn about the “causal compressibility” of a stochastic process at hand.
- Encoded in the rate–distortion curve.

- **Fully** causally compressible: full predictive information can be kept with a model that has a complexity smaller than $H[h]$.
- **Not** causally compressible:
 - Deterministic and time reversible processes (RD curve on the diagonal)
 - i.i.d. processes (RD curve degenerated to a point at the origin)
- **Weakly** causally compressible: RD curve is close to a straight line (small curvature).
- **Strongly** causally compressible: large curvature.

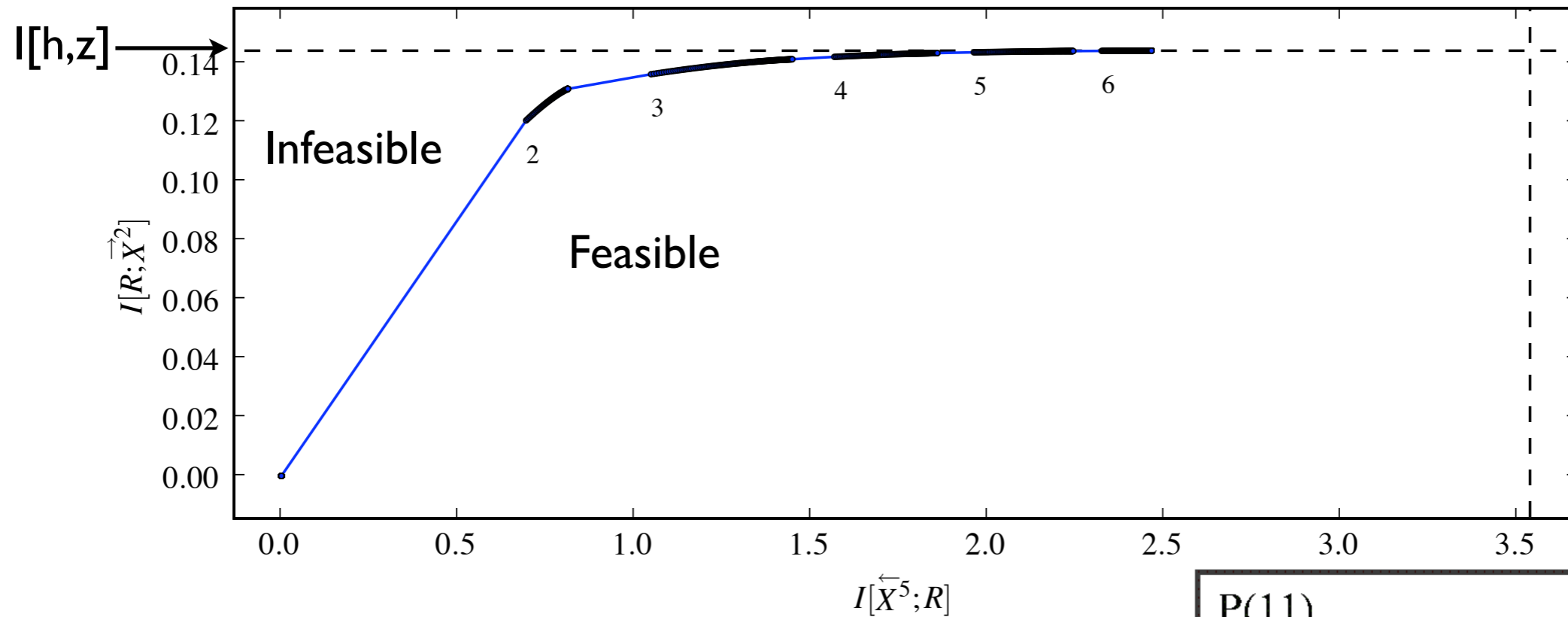
Causal compressibility



Deterministic and
time reversible
processes:
 $I[s;z] = I[s;h]$
 $I[h;z] = H[h]$
Example:
periodic processes

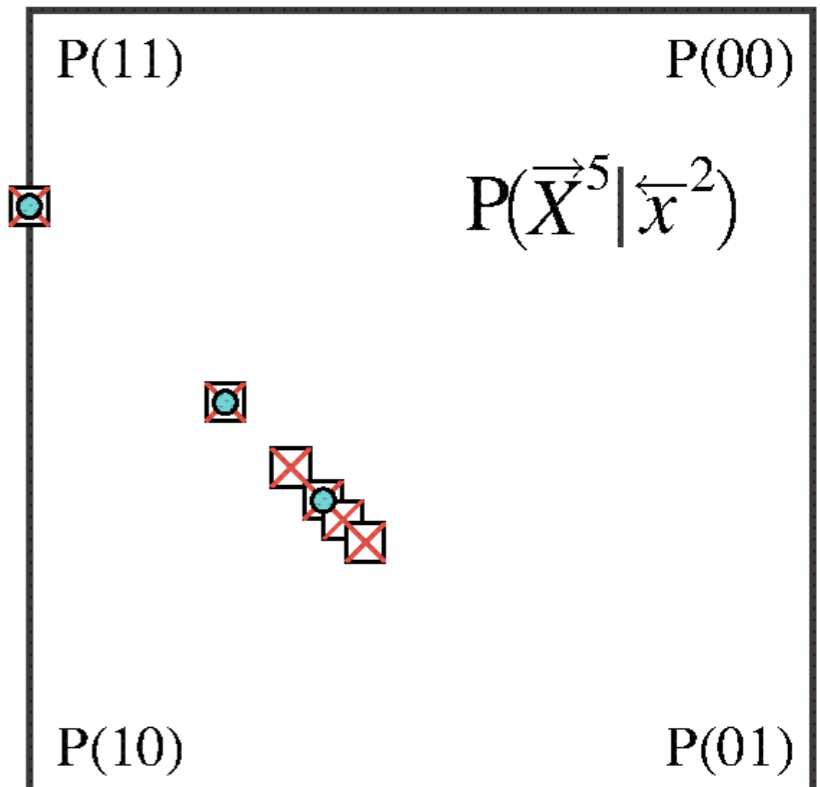
i.i.d. processes:
 $I[s;z] = I[s;h] = 0$

Example (SNS)



Future distributions:

- ✗ full historical information
- best 6-state model
- best 3-state model



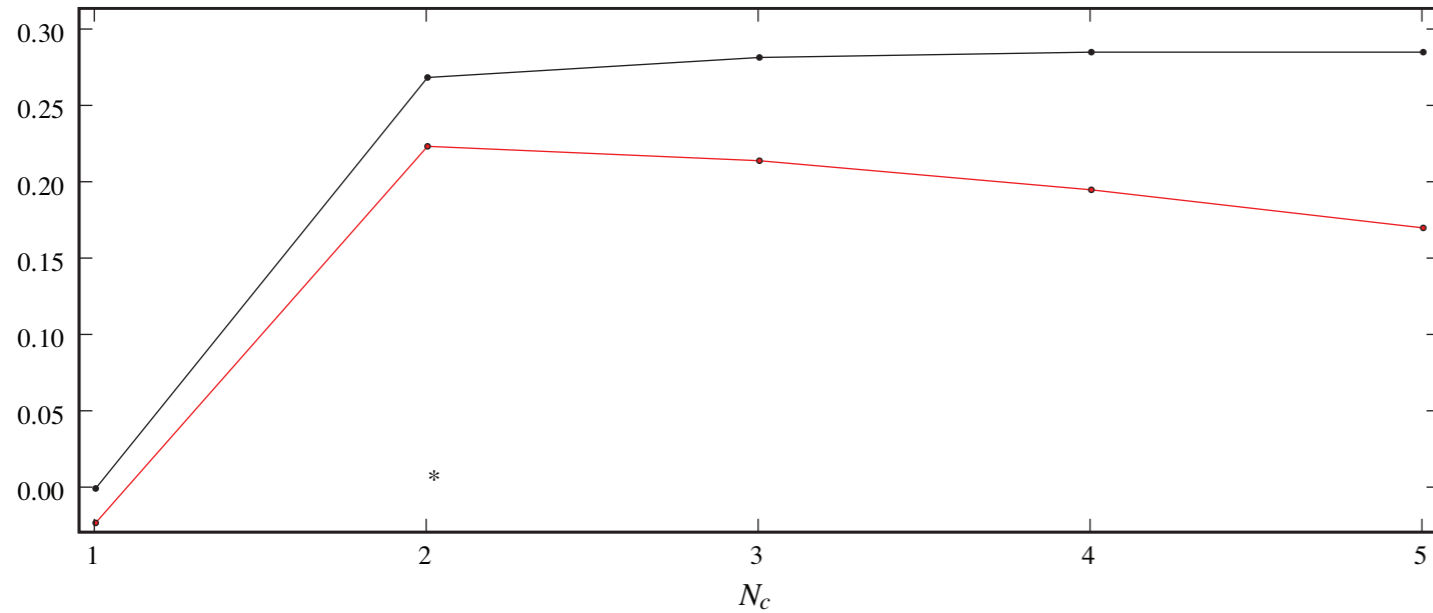
Learning from finite data

- So far, we assumed knowledge of $P(z|h)$
- In practice we have to estimate this distribution from finite data.
- Sampling errors result in overestimate of predictive information \Rightarrow could result in overfitting!
- It looks as if there are more causal states than there really are.

Finite Data

- Find the maximum number of states we can use safely without overfitting.
- Idea: compensate for sampling error in the objective function! Taylor expansion gives estimate of error.
- Result: Corrected curve has a maximum => easy to detect optimal number of states.

Golden Mean Process



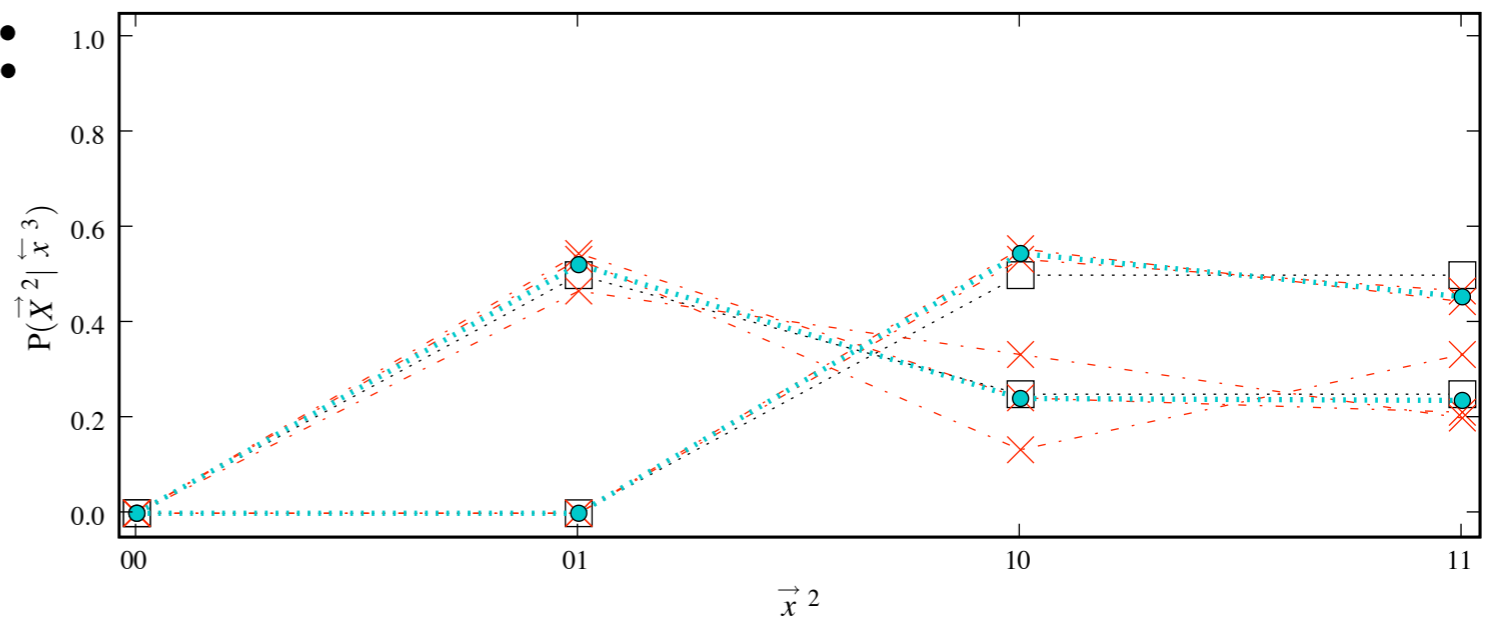
Correction finds
optimal number
of states (2)

Future distributions:

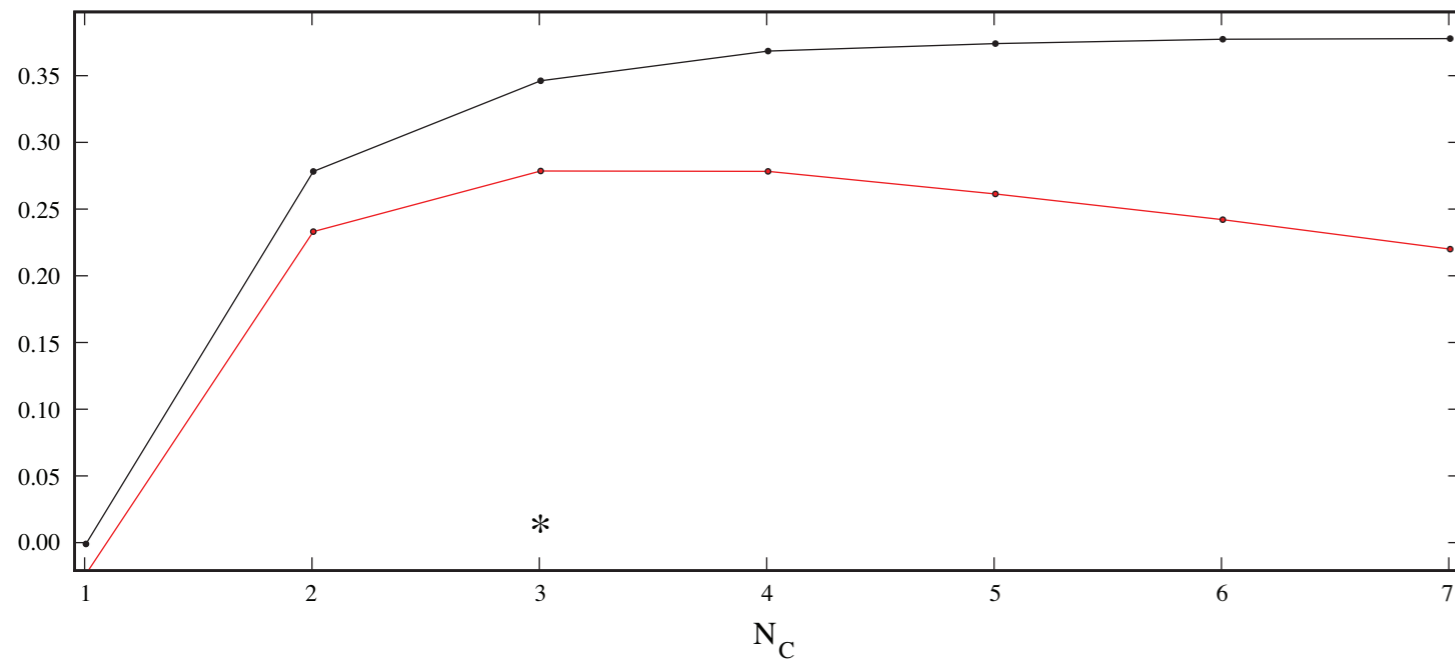
✗ Finite Data

□ ideal (truth)

● result (algorithm)



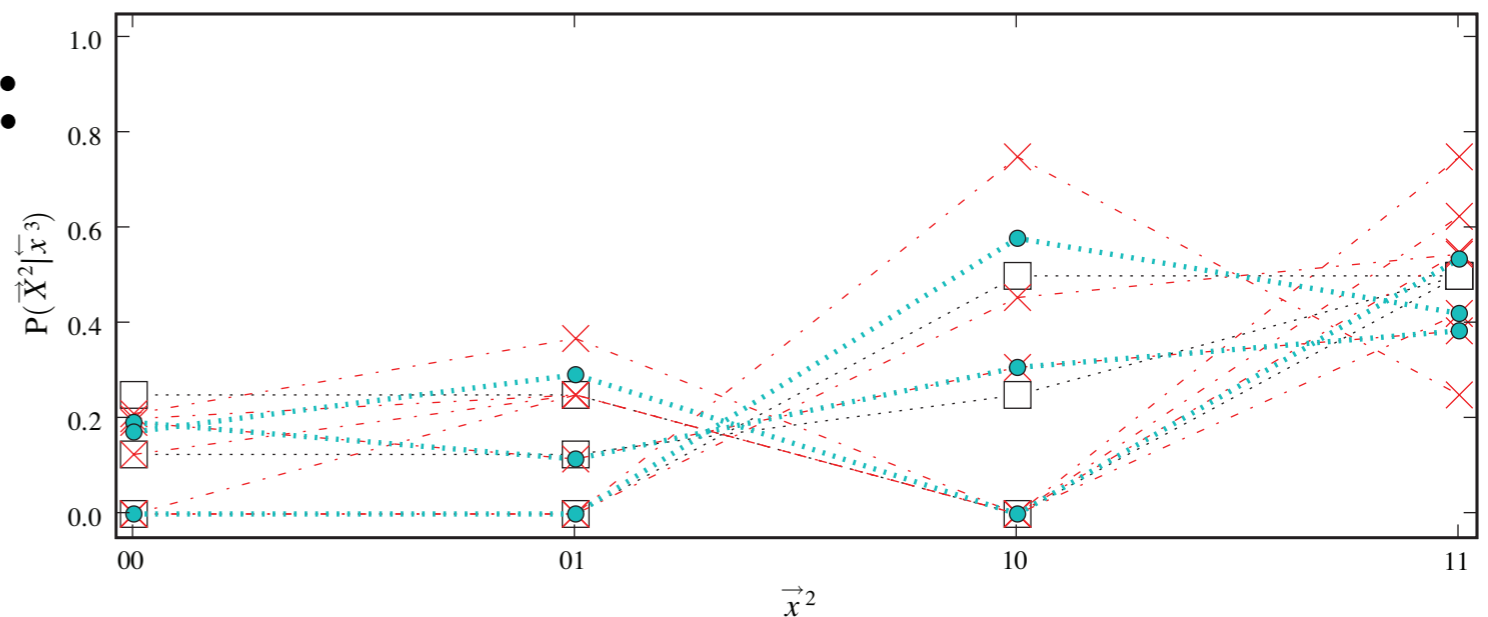
Even Process



Correction finds
optimal number
of states (3)

Future distributions:

- ✗ Finite Data
- ideal (truth)
- result (algorithm)



Conclusion

Simple and intuitive principles allow us to:

1. Find optimal abstractions.

- construct maximally predictive models at fixed model complexity
- in the limit of full prediction we find the causal states (which constitute unique minimal sufficient statistics).
- correct for sampling errors due to finite data set size.

2. characterize a process in terms of its causal compressibility by studying the full range of optimal models.

Extensions

- Online learning: Reconstructing the “epsilon machine”, including transition probabilities.
- Active learning.