Two Lectures on Computational Mechanics:
I. Mostly Leading Up to Causal States

Cosma Rohilla Shalizi
Physics Department, University of Wisconsin, Madison, WI 53706
and the Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501
Electronic address: shalizi@santafe.edu
(18 June 1998)

One man's rigor is another man's mortis.
— Craig Bohren and Bruce Albrecht, *Atmospheric Thermodynamics*, introduction

This document is on-line at http://www.santafe.edu/∼shalizi/comp-mech-lectures/.

Contents

I. INFORMATION THEORY

A. Entropy Defined

Given a random variable $S : \Omega \mapsto \mathcal{A}$, Ω a probability space and $\mathcal{A}$ a countable set, define the entropy of S to be

$$H[S] \equiv -\sum_{s\in\mathcal{A}} P(S=s)\log P(S=s) \tag{1}$$

with the convention that capital letters are random variables and lower-case letters their particular values. (Notice that $H[S]$ is the expectation value of $\log P(S = s)$.)

$H[S]$ is interpreted as the *uncertainty in* S, as the mean number of yes-or-no questions needed to pick out the value of S on repeated trials, if the questions are chosen as well as possible. This is perhaps dubious, but justified by working well.

Any well-behaved probability distribution has an associated entropy.

B. Joint and Conditional Entropies

We define the joint entropy of two variables X (taking values in $\mathcal{A}$) and Y (taking values in $\mathcal{B}$) in the obvious way,

$$H[X,Y] \equiv -\sum_{(x,y)\in(\mathcal{A}\times\mathcal{B})} P(X=x\cdot Y=y)\log P(X=x\cdot Y=y) \tag{2}$$

and define the conditional entropy of one random variable on another from their joint entropy,

$$H[X|Y] \equiv H[X,Y] - H[Y] \tag{3}$$

which follows naturally from the definition of conditional probability, $P(X=x|Y=y) \equiv \frac{P(X=x \cdot Y=y)}{P(Y=y)}$.

We interpret $H[X|Y]$ as the uncertainty remaining in X once we know Y.

C. Useful Inequalities

The following inequalities and implications will prove useful. They're all pretty intuitive, given our interpretation, and they can all be proved with little more than straight algebra; see [1, ch. 2]

$$H[f(X)] \leq H[X] \tag{4}$$
$$H[X,Y] \geq H[X] \tag{5}$$
$$H[X,Y] \geq H[Y] \tag{6}$$
$$H[X|Y] \leq H[X] \tag{7}$$
$$H[X|Y] = 0 \text{ iff } X = f(Y) \tag{8}$$
$$H[X|Y] = H[X] \text{ iff } X \text{is independent of } Y \tag{9}$$
$$H[f(X)|Y] \leq H[X|Y] \tag{10}$$
$$H[X|f(Y)] \geq H[X|Y] \tag{11}$$
$$H[X,Y|Z] \geq H[X|Z] \tag{12}$$
$$H[X|Y,Z] \leq H[X|Y] \tag{13}$$
$$H[X,Y] = H[X] + H[Y|X] \tag{14}$$
$$H[X,Y|Z] = H[X|Z] + H[Y|X,Z] \tag{15}$$

The last two formulæ, called the "chain rules for entropies," are not of course inequalities, but will be handy later on anyhow. (Strictly speaking, Ineq. 9 should have some "except for cases of measure 0" language tacked on.)

II. PATTERNS

The general idea is that some object $\mathcal{O}$ has a pattern $\mathcal{P}$ (or a pattern represented, described, captured, etc. by $\mathcal{P}$ — insert your favorite metaphysics here) iff we can use $\mathcal{P}$ to predict or compress $\mathcal{O}$. (Ability to predict implies ability to compress, but not vice-versa; we will mostly worry about prediction.)

A. Ancestor-Worship

This general notion goes back to Kolmogorov, who was interested in the *exact* reproduction of fixed objects, in particular of binary numbers. His candidates for $\mathcal{P}$ were universal Turing machine programs; in particular, the shortest program which can produce $\mathcal{O}$. (Anything which is Turing-equivalent and where a notion of "length" makes good sense will do, since we can convert from one such system to another — say, from C to Turing machines — with only a finite description of the second system, and such constants will be assimilated in a moment.)

In particular, look at the first n digits of $\mathcal{O}$, $\mathcal{O}_n$, and the shortest program $\mathcal{P}_n$ to produce them. What happens to the limit

$$\lim_{n\to\infty} \frac{|\mathcal{P}_n|}{n} \tag{16}$$

If there is a fixed-length program which can generate arbitrarily many digits of the number, then this limit goes to 0. Most of our interesting numbers, rational or irrational (e.g. $\pi, e, \sqrt{2}$) are of this sort. These numbers are eminently compressible: the program is the compressed description, the pattern which the sequence obeys. If the limit goes to 1, on the other hand, we have a completely incompressible sequence.

There are many problems with the Kolmogorov complexity. It's uncomputable in general (owing to the halting problem); it's maximal for random sequences; it only applies to a single sequence; it makes no allowance for noise, demanding exact reproduction.

B. Patterns with Noise

An obvious next step is to allow some noise, in exchange for shorter descriptions. Some work along these lines has been done by philosophers like Dennett [2]. The intuition is that we should be able to get very simple models of really random processes — "to model coin-tossing, toss a coin." But this is an unstable intermediate position, because allowing noise brings in probability, and the natural setting for probability is ensembles.

C. Patterns in Ensembles

Here a pattern $\mathcal{P}$ is something which lets us predict the future of sequences drawn from the ensemble $\mathcal{O}$ at better than chance rates: it has to be statistically accurate, and confer some leverage, some advantage, as well. Let's fix some notation, and make some assumptions that will later let us prove neat theorems.

Let S_i by a stationary stochastic process which, as before, takes values from a countable set $\mathcal{A}$. We let i range over all the integers, so we get bi-infinite sequence, which we can break at any point we choose into a semi-infinite past, $\overleftarrow{S}$ and a semi-infinite future $\overrightarrow{S}$. (The last L values of the process are $\overleftarrow{S}^L$, the next $\overrightarrow{S}^L$.) We want to predict all or part of $\overrightarrow{S}$ from some function of some part of $\overleftarrow{S}$.

Call the set of all pasts $\overleftarrow{\mathbf{S}}$. Essentially our job is to divide this up into *equivalence classes*, classes of pasts which are equivalent for purposes of predicting the future. Any function on the pasts will induce such classes — simply say two pasts are equivalent if they give the same value for the function — so we'll talk indifferently about classes of pasts, indices for classes of pasts, and states. In general, the function from histories to states will be written $\eta(\overleftarrow{S})$ and will take us to a state $\mathcal{R}$: $\mathcal{R} = \eta(\overleftarrow{S})$. These collections of states are *partitions* of $\overleftarrow{\mathbf{S}}$ in the technical set-theoretic sense, a fact which will be useful latter. The class of all these partitions is sometimes called "Occam's pool" — the joke will make more sense in a little bit.

We say that η captures a pattern iff, for some L,

$$H\left[\overrightarrow{S}^L|\mathcal{R}\right] < H\left[\overrightarrow{S}^L\right] \tag{17}$$

This says that η captures a pattern only if it tells us something about futures of length L. (You can use the inequalities to show that it then tells us something about futures of any length $L' > L$.) Weaker notions of pattern are of course possible, but the reasons for our using this one (essentially, we can prove theorems about it) will be apparent shortly. Note that sticking to finite-lenngth futures means that we only have to deal with finite entropies.

Since $H[X|Y,Z] \leq H[X|Y]$ (Ineq. 14), it follows that $\forall n, H\left[\overrightarrow{S}^L|\overleftarrow{S}^n\right] \geq H\left[\overrightarrow{S}^L|\overleftarrow{S}^{n+1}\right]$, and so $\forall n, H\left[\overrightarrow{S}^L|\overleftarrow{S}^n\right] \geq H\left[\overrightarrow{S}^L|\overleftarrow{S}\right]$. Since $H[X|f(Y)] \geq H[X|Y]$ (Ineq. 12), for any η,

$$H\left[\overrightarrow{S}^L|\mathcal{R}\right] = H\left[\overrightarrow{S}^L|\eta(\overleftarrow{S})\right] \geq H\left[\overrightarrow{S}^L|\overleftarrow{S}\right]. \tag{18}$$

That is, conditioning on the whole of the past reduces the uncertainty in the future to as small a value as possible; but carrying the whole semi-infinite past around is a rather bulky and uncomfortable prospect. (Put a bit differently: We're Americans and would like to forget as much of the past as we possibly can.)

Let's invoke Occam's razor: "It is vain to do with more what can be done with less." To use it we need to fix what is to be done, and what "more" and "less" mean. Now, the job we want done is getting $H\left[\overrightarrow{S}^L|\mathcal{R}\right]$ down as far as possible, for all lengths L at once, to the just-established limit of $H\left[\overrightarrow{S}^L|\overleftarrow{S}\right]$. But we want to do this as simply as possible, with as little as possible. Now, because $P(\overleftarrow{S}=\overleftarrow{s})$ is well-defined, there's an induced measure on the η-states, i.e. $P(\mathcal{R}=r)$ is well-defined. But this means we can calculate the entropy of the distribution over states, $H[\mathcal{R}]$. This uncertainty in our current state is the average amount of memory, in bits, that we retain about the past; we would like to do with as little of this as possible. (Again, this is America.) The quantity $H[\mathcal{R}]$ is called the "statistical complexity" or "machine complexity" of the set of states $\{r\}$; so we want to minimize statistical complexity, subject to the constraint of maximally accurate prediction. (The idea behind calling the class of all partitions of $\overleftarrow{\mathbf{S}}$ Occam's pool should now be clear: we want to find the shallowest point in the pool.)

III. CAUSAL STATES

It would be nice if we could take our (maximally accurate prediction and minimal statistical complexity), make them axioms and show that there's only one set of states (and so one function from histories to those states) which satisfy them; at least, I suppose it would be nice if you like axiomatic systems. No such proof is known. What we can prove is that states defined in a certain way satisfy those constraints. (I suspect they're the *only* set of states which satisfy those constraints, but haven't shown it yet.)

A. Definition of Causal States

We define the function ϵ from histories to classes of histories thus:

$$\epsilon(\overleftarrow{s}) \equiv \left\{ \overleftarrow{s}' | \forall \overrightarrow{s} \left(P\left(\overrightarrow{S} = \overrightarrow{s} \mid \overleftarrow{S} = \overleftarrow{s}\right) = P\left(\overrightarrow{S} = \overrightarrow{s} \mid \overleftarrow{S} = \overleftarrow{s}'\right)\right)\right\} \tag{19}$$

The range of ϵ consists of the *causal states* of the process. Alternately and equivalently (exercise!), we could define an equivalence relation $\sim$ such that two histories are equivalent iff they have the same conditional distribution of futures, and then define causal states as the equivalence classes generated by $\sim$. Either way, the divisions of this partition of $\overleftarrow{\mathbf{S}}$ are made between regions which leave us in different degrees of ignorance about the future.

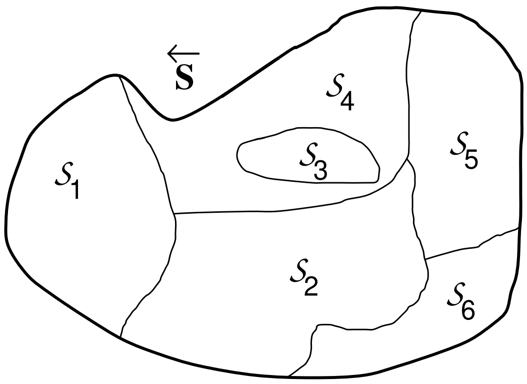

FIG. 1. A schematic representation of the partitioning of the set $\overleftarrow{\mathbf{S}}$ of all histories into causal states $\{\mathcal{S}_i\}$. Within each causal state all the individual histories $\overleftarrow{S}$ have the same conditional distribution $P(\overrightarrow{S} \mid \overleftarrow{S})$ for future observables. Note that the $\mathcal{S}_i$ need not form compact sets; we have simply drawn them that way here for clarity.

In the statistical inference and statistical explanation literature, they'd say that causal states are the "statistical-relevance basis for causal explanations," where the elements of the basis are maximal combinations of independent variables with statistically different distributions for the dependent variables. (See [3] for the gory details.)

More colloquially: The causal states record every distinction that makes a difference.

B. Proving that Causal States Are the Funk

I'm now going to try to convince you that causal states are the neatest thing since sliced bread, by proving three optimality results about them and some related lemmas. The first two results show that they satisfy the constraints we got from Occam, of accurate prediction and minimal complexity; the third shows that there's a sense in which they're maximally deterministic. All of these results involve comparing causal states, generated by ϵ, with other sets of states, generated by some other function η, and showing that none of these other sets of states — none of the other patterns — can out-perform the causal states.

More notation fixing: $\mathcal{S}$ is the random variable for the current causal state, S_1 is the next "observable" we get from the original stochastic process, $\mathcal{S}'$ is the next causal state, $\mathcal{R}$ is the current state according to η, $\mathcal{R}'$ the next η-state. Since LaTeXdoesn't have lower-case calligraphic characters, s will stand for a particular value of $\mathcal{S}$, r for a particular value of $\mathcal{R}$. When I need to quantify over alternatives to the causal states, I will quantify over $\mathcal{R}$.

Theorem 1 (Causal States Are Maximally Prescient) $\forall \mathcal{R}$, $H\left[\overrightarrow{S}^L|\mathcal{R}\right] \geq H\left[\overrightarrow{S}^L|\mathcal{S}\right]$.

Proof. We've already seen that $H\left[\overrightarrow{S}^L|\mathcal{R}\right] \geq H\left[\overrightarrow{S}^L \mid \overleftarrow{S}\right]$. But by construction (Eq. 19), $P(\overrightarrow{S}=\overrightarrow{s} \mid \overleftarrow{S}=\overleftarrow{s}) = P(\overrightarrow{S}=\overrightarrow{s} \mid \mathcal{S} = \epsilon(\overleftarrow{s}))$. Since entropies depend only on the probability distribution, $H\left[\overrightarrow{S}^L|\mathcal{S}\right] = H\left[\overrightarrow{S}^L \mid \overleftarrow{S}\right]$. Thus $H\left[\overrightarrow{S}^L|\mathcal{R}\right] \geq H\left[\overrightarrow{S}^L|\mathcal{S}\right]$.

That is to say, causal states are as good at predicting the future — are as *prescient* — as complete histories; they satisfy the first requirement we took from Occam.

All are subsequent results will concern rivals which are as prescient as the causal states.

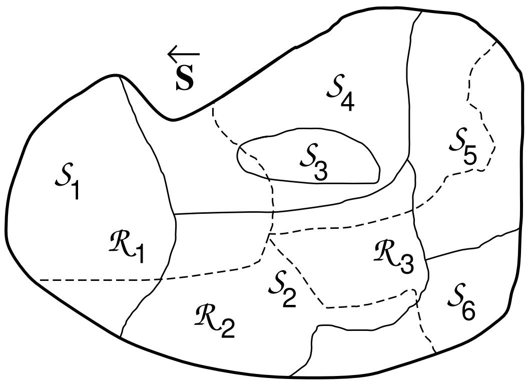

FIG. 2. An alternative set $\{\mathcal{R}_i\}$ of states that partition $\overleftarrow{\mathbf{S}}$ overlaid on the causal states. The collection of all such alternative partitions form Occam's "pool". Note again that the $\mathcal{R}_i$ need not be compact.

Lemma 1 (Rivals of Causal States Are Refinements) $\forall \mathcal{R}$, *if* $H\left[\overrightarrow{S}^L|\mathcal{R}\right] = H\left[\overrightarrow{S}^L|\mathcal{S}\right]$, *then* $\forall r \exists s$ *such that* $r \subseteq s$.

Proof. We invoke a trivial extension of theorem 2.7.3 of [1]: If X_1, X_2, $\ldots X_n$ are random variables over the same set $\mathcal{A}$ with distinct probability distributions, Θ is a random variable over the integers from 1 to n such that $P(\Theta = i) = \lambda_i$, and Z is a random variable over $\mathcal{A}$ such that $Z = X_\Theta$, then $H[Z] = H\left[\sum_i \lambda_i X_i\right] \geq \sum_i \lambda_i H[X_i]$. (The first sum is purely symbolic; the second is for real.) This becomes an equality iff all the λ_i are either 0 or 1, since H is strictly concave, because $x \log x$ is strictly convex for $x \geq 0$.

Define $\phi_{sr} \equiv P(\mathcal{S} = s|\mathcal{R} = r)$.

Then

$$H\left[\overrightarrow{S}^L|\mathcal{R}=r\right] = H\left[\sum_s \phi_{sr} P(\overrightarrow{S}^L|\mathcal{S}=s)\right] \tag{20}$$

$$\geq \sum_s \phi_{sr} H\left[\overrightarrow{S}^L|\mathcal{S}=s\right] \tag{21}$$

$$H\left[\overrightarrow{S}^L|\mathcal{R}\right] = \sum_r P(\mathcal{R}=r) H\left[\overrightarrow{S}^L|\mathcal{R}=r\right] \tag{22}$$

$$\geq \sum_r P(\mathcal{R}=r) \sum_s \phi_{sr} H\left[\overrightarrow{S}^L|\mathcal{S}=s\right] \tag{23}$$

$$= \sum_{sr} P(\mathcal{R}=r)\phi_{sr} H\left[\overrightarrow{S}^L|\mathcal{S}=s\right] \tag{24}$$

$$= \sum_{sr} P(\mathcal{S}=s \cdot \mathcal{R}=r) H\left[\overrightarrow{S}^L|\mathcal{S}=s\right] \tag{25}$$

$$= \sum_i P(\mathcal{S} = s) H\left[\overrightarrow{S}^L | \mathcal{S} = s\right] \quad (26)$$

$$= H\left[\overrightarrow{S}^L | \mathcal{S}\right] \quad (27)$$

That is to say,

$$H\left[\overrightarrow{S}^L | \mathcal{R}\right] \geq H\left[\overrightarrow{S}^L | \mathcal{S}\right] \quad (28)$$

with equality if and only if every ϕ_{sr} is either 0 or 1. Thus, if $H\left[\overrightarrow{S}^L | \mathcal{R}\right] = H\left[\overrightarrow{S}^L | \mathcal{S}\right]$, every r is entirely contained within some s (except for possible sub-sets of measure 0).

(We cannot work the proof the other way around, and show that the causal states have to be a refinement of the equally-prescient η-states, because applying the theorem we stole from Cover and Thomas hinges on being able to reduce uncertainty by specifying *which* distribution we're picking from. Since the causal states are constructed so that the distribution of futures is the same for all their sub-sets, this isn't so; Eq. 19 and Theorem 1 together protect us.)

Theorem 2 (Causal States Have Minimal Complexity) *$\forall \mathcal{R}$, if $H\left[\overrightarrow{S}^L | \mathcal{R}\right] = H\left[\overrightarrow{S}^L | \mathcal{S}\right]$, then $H[R] \geq H[S]$.*

Proof. By Lemma 1, every $r \subseteq$ some s. This means that there is a many-to-one relation from the η-states to the causal states, i.e., that the causal state is a function of the η-state, $\mathcal{S} = g(\mathcal{R})$. But $H[f(X)] \leq H[X]$ (Ineq. 5) so $H[\mathcal{S}] = H[g(\mathcal{R})] \leq H[\mathcal{R}]$.

Recall that earlier we defined the entropy in a set of states used for prediction, for a pattern, to be the statistical complexity of that pattern. We have just established that no rival pattern, which is as good at predicting as the causal states, is any simpler than the causal states. Occam therefore tells us that we should use the causal states.

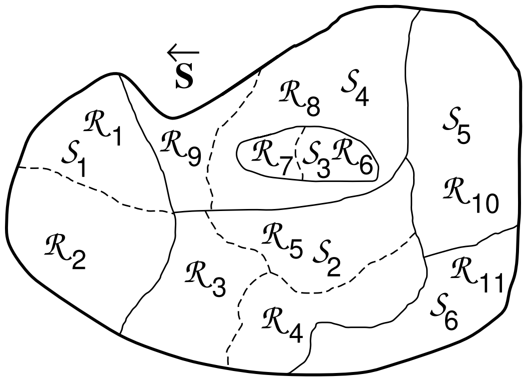

FIG. 3. Any alternative partition $\{\mathcal{R}\}$ that is as prescient as the causal states must be a refinement of the causal-state partition. That is, each $\mathcal{R}_i$ must be a (possibly improper) subset of some $\mathcal{S}_j$. Otherwise, at least one $\mathcal{R}_i$ would have to contain parts of at least two causal states. And so using this $\mathcal{R}_i$ to predict the future observables leads to more uncertainty about $\overrightarrow{S}$ than using the causal states.

The entropy of the causal states is also conventionally written C_μ, with the μ to remind us that it's a metric property, and depends on the distribution over states.

It is here that it becomes important that we're trying to predict $\overrightarrow{S}$ and not just some $\overrightarrow{S}^L$. Suppose two histories $\overleftarrow{s}$ and $\overleftarrow{s}'$ have the same conditional distribution for $\overrightarrow{S}^L$, but differ at some point after the next L steps into the future. They would then belong to different causal states. An η-state which merged those two causal states, however, would have just as much ability to predict $\overrightarrow{S}^L$ as the causal states would, but would be simpler (the uncertainty in the current state would be lower). Causal states are optimal — but for the hardest job.

Lemma 2 (Causal States Are Deterministic Automata) *There exists a function N such that $s' = N(s, s_1)$.*

In automata theory, a set of states is said to be *deterministic* if the current state and the next input (here, the next result from the original stochastic process) together fix the next state.

Proof. The lemma is equivalent to asserting that $\forall s_1, \overleftarrow{s}, \overleftarrow{s}'$, if $\overleftarrow{s} \sim \overleftarrow{s}'$, then $\overleftarrow{s}\, s_1 \sim \overleftarrow{s}' s_1$, where $\overleftarrow{s}\, s_1$ is to be understood as the semi-infinite sequence which we get from tacking s_1 to the end of $\overleftarrow{s}$. (This is a just another history, and belongs to some causal state or other.) Suppose this were not true. Then there would have to exist at least one future $\overrightarrow{s}$ such that

$$P(\overrightarrow{S}=\overrightarrow{s} \mid \overleftarrow{S}=\overleftarrow{s}\, s_1) \neq P(\overrightarrow{S}=\overrightarrow{s} \mid \overleftarrow{S}= \overleftarrow{s}' s_1) \tag{29}$$

But this would imply that

$$P(\overrightarrow{S}= s_1 \overrightarrow{s} \mid \overleftarrow{S}=\overleftarrow{s}) \neq P(\overrightarrow{S}= s_1 \overrightarrow{s} \mid \overleftarrow{S}= \overleftarrow{s}') \tag{30}$$

where we read $s_1 \overrightarrow{s}$ as the semi-infinite string which begins s_1 and continues as $\overrightarrow{s}$. (Remember, we assumed that the point at which we break the stochastic process into a past and a future is arbitrary.) But this is to say that there's a future which has different probabilities depending on whether we conditioned on $\overleftarrow{s}$ or on $\overleftarrow{s}'$, which is contrary to our assumption that the two histories belong to the same causal state. Therefore, there is no such future $\overrightarrow{s}$, and the alternative statement of the lemma is true, so the lemma is true.

Theorem 3 (Causal States Are Maximally Deterministic) $\forall \mathcal{R}$, *if* $H\left[\overrightarrow{S}^L|\mathcal{R}\right] = H\left[\overrightarrow{S}^L|\mathcal{S}\right]$, *then* $H[\mathcal{R}'|\mathcal{R}] \geq H[\mathcal{S}'|\mathcal{S}]$.

From lemma 2, $\mathcal{S}' = N(\mathcal{S}, S_1)$, therefore $H[\mathcal{S}'|\mathcal{S}, S_1] = 0$ (by Ineq. 9). Therefore, from the chain rule for entropies, Eq. 15,

$$H[S_1|\mathcal{S}] = H[\mathcal{S}', S_1|\mathcal{S}] \tag{31}$$

We have no result like the lemma for the rival states $\mathcal{R}$, but entropies are always non-negative, $H[\mathcal{R}'|\mathcal{R}, S_1] \geq 0$. Since $H\left[\overrightarrow{S}^L|\mathcal{R}\right] = H\left[\overrightarrow{S}^L|\mathcal{S}\right]$ (by hypothesis), $H[S_1|\mathcal{R}] = H[S_1|\mathcal{S}]$ (by forming the appropriate marginal distributions and then taking their entropies). Now we apply the chain rule again,

$$\begin{aligned} H[\mathcal{R}', S_1|\mathcal{R}] &= H[S_1|\mathcal{R}] + H[\mathcal{R}'|S_1, \mathcal{R}] && (32)\\ &\geq H[S_1|\mathcal{R}] && (33)\\ &= H[S_1|\mathcal{S}] && (34)\\ &= H[\mathcal{S}', S_1|\mathcal{S}] && (35)\\ &= H[\mathcal{S}'|\mathcal{S}] + H[S_1|\mathcal{S}', \mathcal{S}] && (36) \end{aligned}$$

where in the last step we have used the chain rule once more.

Using the chain rule one last time (only with feeling),

$$H[\mathcal{R}', S_1|\mathcal{R}] = H[\mathcal{R}'|\mathcal{R}] + H[S_1|\mathcal{R}', \mathcal{R}]. \tag{37}$$

Putting these two expansions together, we get

$$\begin{aligned} H[\mathcal{R}'|\mathcal{R}] + H[S_1|\mathcal{R}', \mathcal{R}] &\geq H[\mathcal{S}'|\mathcal{S}] + H[S_1|\mathcal{S}', \mathcal{S}] && (38)\\ H[\mathcal{R}'|\mathcal{R}] - H[\mathcal{S}'|\mathcal{S}] &\geq H[S_1|\mathcal{S}', \mathcal{S}] - H[S_1|\mathcal{R}', \mathcal{R}] && (39) \end{aligned}$$

From lemma 1, we know that $\mathcal{S} = g(\mathcal{R})$, which means there's another function g' from ordered pairs of η-states to ordered pairs of causal states, $(\mathcal{S}', \mathcal{S}) = g'(\mathcal{R}', \mathcal{R})$. Therefore, inequality 12 implies $H[S_1|\mathcal{S}', \mathcal{S}] \geq H[S_1|\mathcal{R}', \mathcal{R}]$ and so

$$\begin{aligned} H[S_1|\mathcal{S}', \mathcal{S}] - H[S_1|\mathcal{R}', \mathcal{R}] &\geq 0 && (40)\\ H[\mathcal{R}'|\mathcal{R}] - H[\mathcal{S}'|\mathcal{S}] &\geq 0 && (41)\\ H[\mathcal{R}'|\mathcal{R}] &\geq H[\mathcal{S}'|\mathcal{S}] && (42) \end{aligned}$$

Q.E.D.

What the theorem says is that there is no more uncertainty about the next causal state, given the current causal state, then there is about the next state given the current state for any other set of states which are as prescient. Or, perhaps slightly less of a mouthful: the causal states approach as closely to perfect determinism (in the usual sense!) as any rival which is as good at predicting the future.

C. Some Restrictions May Apply...

Let's catalogue all the restrictive assumptions we've made so far.

1. The observed process takes on discrete values.
2. The process is discrete in time.
3. The process is a pure time-series, without spatial extension.
4. The observed process is stationary.
5. Prediction can only be based on the past of the process, not on any outside source of information.

Can any of these be relaxed without much trouble?

The first probably can; the information-theoretic quantities we've been using are defined for continuous random variables; somebody (probably me...) will just have to grind through the math and make sure everything checks out. The second also looks solvable, since there's a lot of math on continuous stochastic processes, but it may involve some funky probability theory or functional analysis. There are already tricks to make spatially extended systems look like time-series (essentially, one looks at all the paths through space-time, treating each one like a time-series), which the lecture on CAs (if it happens) will cover.

We don't know how to relax the assumption of stationarity.

Finally, I'd say that the last restriction is a *feature* when it comes to thinking about patterns and the intrinsic structure of a process. "Pattern" is a vague word of course, but I *think* it's only supposed to involve things *inside* the process, not the rest of the universe. In any case, if we relax that assumption, lots of troubling sources of information present themselves. Imagine that one Sunday afternoon you wander over to your friend's house and find him watch a ball game on TV. He offers you a bet on the next play, which you take and lose; and he keeps on offering you bets all through the game, which you keep on losing. Is he unnaturally lucky or skilled? No; the game was really broadcast two hours earlier and you are watching a videotape. Your friend can obtain quite remarkable accuracy by using this source of information outside the current process; but that's cheating, and doesn't tell us very much about the pattern of the game. But this scruple is only appropriate if what you care about is the pattern; if you just want the best prediction you can possibly get, watch the videotape!

---

[1] Thomas M. Cover and Joy A. Thomas (1991), *Elements of Information Theory*. New York: Wiley.

[2] Daniel C. Dennett (1991), "Real Patterns," *Journal of Philosophy* **88 (1)**, 27–51. Reprinted in Daniel Dennett, *Brainchildren: Essays on Designing Minds*, Cambridge, Massachusetts: MIT Press, 1997.

[3] Wesley C. Salmon (1984), *Scientific Explanation and the Causal Structure of the World*. Princeton, New Jersey: Princeton University Press.