5

Which Processes Satisfy the Second Law?

Thomas M. Cover†
Durand Bldg Rm 121
Stanford University
Stanford, CA 94301, USA

5.1 Introduction

The second law of thermodynamics states that entropy is a nondecreasing function of time. One wonders whether this law is built into the physics of the universe or whether it is simply a common property of most stochastic processes. If the latter is the case, we should be able to prove the second law under mild conditions.

Thus motivated, we will reverse the usual physical development and put the emphasis on stochastic processes, physically generated or otherwise, and attempt to determine the family of processes for which the second law holds. In the course of this treatment we will suggest that relative entropy and conditional entropy are natural notions of what is meant by entropy in the second law. Certainly, the second law is true under milder conditions as we shift to these definitions.

We shall concern ourselves here, primarily, with discrete time finite state Markov processes. To the extent that the physical universe is Markovian, our comments will apply to physics. Here we should be aware that coarse graining (lumping of states) of a Markov chain may destroy Markovity. Also, while the Schrödinger wave function seems to evolve in a Markovian manner, the associated probabilities do not. Thus Markovity is a strong assumption.

We shall use Shannon entropy throughout. We ask whether the second law of thermodynamics is true of all finite state Markov processes. We shall find, somewhat surprisingly, that it is only true of doubly stochastic Markov processes. Equivalently, the second law is only true of Markov processes for which the equilibrium distribution is uniform over the finite state space. We will find that a slight change in the statement of the second law suffices to cover all Markov chains. Instead of the statement, "entropy always increases," we may substitute the more general statement that "relative entropy (of the current distribution with respect to the stationary distribution) decreases."

An interesting discussion of time symmetry and the second law can be found in Mackey (1992). The development of the second law from the physical standpoint is

† Email: cover@isl.stanford.edu.

argued in Van Kampen (1990), Wehrl (1978) and Tisza (1966), where good histories of the subject can be found. The consequences of the second law for Maxwell's Demon can be found in the collection of papers by Leff and Rex (1990). A probabilistic investigation of the behavior of entropy for stochastic processes can be found in Kullback (1959), Renyi (1961), Csiszar (1967), Fritz (1971), and Cover and Thomas (1991).

5.2 Entropy and its Interpretations

Let X be a random variable drawn according to a probability mass function $p(x)$ over the finite set of outcomes $\mathcal{X}$. Shannon entropy is defined as

$$H(X) = -\sum p(x) \log p(x).$$

We shall sometimes denote this as $H(p)$. Here the entropy H has the interpretation that it is the minimal expected number of yes-no questions required to determine the outcome X. It can also be shown that it is the minimum expected number of fair coin flips required by a random number generator to generate an outcome X with the desired distribution. Also of importance is the conditional entropy $H(X|Y)$ which can be written as

$$H(X|Y) = -\sum_{x,y} p(y)p(x|y) \log p(x|y) = \sum_{y} p(y)H(X|Y = y).$$

It can be shown, by writing $\log p(x, y) = \log p(x) + \log p(y|x)$, that $H(X, Y) = H(Y) + H(X|Y)$. The strict concavity of the logarithm and Jensen's inequality immediately yield the result

$$H(X|Y) \leq H(X),$$

with equality if and only if X and Y are independent. Thus, conditioning always reduces entropy. In fact, the reduction in entropy is strict unless the conditioning random variable Y is independent of X.

Some other interpretations of the entropy H are as follows:

Descriptive complexity:

$$H(X) \leq El(X) < H(X) + 1$$

where $El(X)$ is the minimum expected number of bits in the description of X.

Asymptotic equipartition theorem:
If $X_1, X_2, \ldots$ be a discrete valued ergodic random process. Then the actual probability of the sequence of outcomes $X_1, \ldots, X_n$ is close to 2^{-nH} where H is

the entropy rate of the process defined by $H = \lim_{n\to\infty} H(X_1,\ldots,X_n)/n$. More precisely,

$$p(X_1,\ldots,X_n) = 2^{-nH+o(n)},$$

where $o(n)/n$ converges to 0 as $n \to \infty$, with probability 1. The number of such "typical" sequences is approximately 2^{nH}. This result allows one to interpret 2^H as the volume of the effective support set of X.

Kolmogorov complexity:
Let $K(x) = \min_{p:U(p)=x} l(p)$ be the minimum program length for a computer $\mathcal{U}$, which causes the computer to print x and halt. Let $X_1, X_2, \ldots$ be an ergodic process with entropy rate H. Then

$$E\frac{K(X_1, X_2, \ldots, X_n)}{n} \to H, \quad \text{as } n \to \infty.$$

Thus Kolmogorov complexity and Shannon entropy are asymptotically equal for ergodic processes. See Cover and Thomas (1991) for a proof for independent identically distributed processes.

Number of microstates:
For rough statements of the second law, it often suffices to take the logarithm of the number of microstates corresponding to a given macrostate in order to characterize the entropy of that macrostate. Implicit in this is that the probability is uniformly distributed over the microstates. (See our remarks about the asymptotic equipartition theorem.) In any case, the critical calculation of the number of microstates usually involves a multinomial coefficient which can be shown to be equal to

$$\binom{n}{np_1, np_2, \ldots, np_m} = 2^{nH(p_1,p_2,\ldots,p_m)+o(n)}.$$

Another important quantity for this discussion will be the relative entropy $D(p||r)$, sometimes known as the Kullback-Leibler information, or the information for discrimination. The relative entropy $D(p \parallel r)$ between two probability mass functions $p(x)$ and $r(x)$, $p(x) \geq 0$, $\sum p(x) = 1$, $r(x) \geq 0$, $\sum r(x) = 1$, is defined by

$$D(p \parallel r) = \sum_x p(x) \log \frac{p(x)}{r(x)}.$$

The relative entropy is always nonnegative, as shown in the following theorem:

Theorem 1. *$D(p \parallel r) \geq 0$ with equality if and only if $p(x) = r(x)$ for all x.*

Proof. Let A be the support set of $p(x)$. We use Jensen's inquality and the strict

concavity of the logarithm to show

$$-D(p \parallel r) = \sum_A p(x) \log \frac{r(x)}{p(x)} \leq \log \sum_A p(x) \frac{r(x)}{p(x)} = \log \sum_A r(x) \leq \log 1 = 0.$$

The interpretations of relative entropy are as follows:

Likelihood ratio:
The relative entropy is the expected log likelihood ratio between distributions p and r.

Hypothesis testing exponent:
The probability of error in a hypothesis test between distribution p and distribution r for independent identically distributed observations drawn according to one of these, has a probability of error given to first order in the exponent by $P_e = e^{-nD}$. Thus D is the degree of difficulty in distinguishing two distributions.

Redundancy:
If one designs an optimal description for distribution r when in fact distribution p is true, then instead of requiring $H(p)$ bits for the description, the random variable requires $H(p) + D(p \parallel r)$ bits, as given in the following expression:

$$H + D \leq E_p l(X) < H + D + 1.$$

Large deviation theory:
Also, relative entropy arises in large deviation theory. The probability of physical data appearing to have macrostate r when in fact observations are drawn according to p is $e^{-nD(r\|p)+o(n)}$.

5.3 General Results about Increase in Entropy

Although we shall eventually argue that the entropy increase is not true for general Markov chains, there are a number of preliminary general results about the increase of entropy which agree with intuition.

First, it makes sense that for any stochastic process whatsoever, Markov or not, in equilibrium or not, the entropy of the present state given the past increases as the amount of information about the past decreases. This is due to the fact that conditioning always reduces entropy. This is embodied in the following theorem.

This theorem proves that the conditional entropy of the present given the far past increases as the past recedes, but simple examples exist for which the conditional entropy of the process at time n given the fixed past up to time 0 may actually decrease.

Let X_m^n denote $(X_m, X_{m+1}, X_{m+2}, \ldots, X_n)$ throughout this discussion.

Theorem 2. *For all stochastic processes, $H(X_0|X_{-\infty}^{-n})$ is monotonically nondecreasing.*

The apparently similar quantity $H(X_n|X_{-\infty}^0)$ does not generally increase with n, but it does increase if the process is stationary.

Proof. Conditioning reduces entropy. Thus

$$H(X_0|X_{-\infty}^{-n}) = H(X_0|X_{-\infty}^{-(n+1)}, X_{-n}) \leq H(X_0|X_{-\infty}^{-(n+1)}),$$

proving the first statement. In the second statement, periodic processes provide counterexamples to the monotonicity of $H(X_n|X_{-\infty}^0)$, but the additional assumption of stationarity yields

$$H(X_{n+1}|X_{-\infty}^0) = H(X_n|X_{-\infty}^{-1}) \geq H(X_n|X_{-\infty}^0),$$

establishing the increase of $H(X_n|X_{-\infty}^0)$ for stationary processes.

We can now demonstrate a nice symmetry property of the conditional entropy of the present, given the past and given the future, for all stationary processes, Markov or otherwise. (A stationary process is in equilibrium.)

Theorem 3. *$H(X_0|X_{-1}, X_{-2}, \ldots, X_{-n}) = H(X_0|X_1, X_2, \ldots, X_n)$ for all stationary processes, Markov or otherwise. Also, for all stationary processes,*

$$H(X_{-n}^{-1}|X_0) = H(X_1^n|X_0).$$

Proof. By stationarity, $H(X_{-n}, \ldots, X_{-1}, X_0) = H(X_0, X_1, \ldots, X_n)$. Then the chain rule yields

$$H(X_{-n}^{-1}|X_0) + H(X_0) = H(X_1^n|X_0) + H(X_0),$$

thus proving the second assertion. The first assertion is proved similarly.

Remark. The fact that the entropy of the present given the n-past is equal to the entropy of the present given the n-future is somewhat surprising in light of the fact that the statement is true even for time-irreversible processes. Consider, for example, a Markov chain with transition matrix

$$P = \begin{bmatrix} .1 & .9 & 0 \\ 0 & .1 & .9 \\ .8 & 0 & .2 \end{bmatrix}.$$

Here it is clear that one can determine the direction of time by looking at the sample path. Nonetheless, the entropy of the present given a chunk of the future is equal to the entropy of the present given the corresponding chunk of the past.

5.4 Relative Entropy Always Decreases

We now state a theorem about relative entropy which shows the monotonic increase of relative entropy for all Markov chains, stationary or not. From this we will derive, by application, the second law of thermodynamics, which holds for doubly stochastic Markov chains. Versions of the following theorem appear in Kullback

(1959), Cover and Thomas (1991), Van Kampen (1990), Fritz (1973), Csiszár (1967), Renyi (1961), and the survey by Wehrl (1978).

Theorem 4. *Let μ_n and μ'_n be two probability mass functions on the state space of a finite state Markov chain at time n. Then $D(\mu_n \parallel \mu'_n)$ is monotonically decreasing. In particular, if μ is the unique stationary distribution,*

$$D(\mu_n \parallel \mu) \searrow 0.$$

Before proving this we need a definition and a lemma. We first define a conditional version of the relative entropy.

Definition: Given two joint probability mass functions $p(x, y)$ and $q(x, y)$, the *conditional relative entropy* $D(p(y|x) \parallel q(y|x))$ is the expected value of the relative entropies between the conditional probability mass functions $p(y|x)$ and $q(y|x)$ averaged over the probability mass function $p(x)$. More precisely,

$$D(p(y|x) \parallel q(y|x)) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)}.$$

Lemma: *(Chain rule for relative entropy.)*

$$D(p(x, y) \parallel q(x, y)) = D(p(x) \parallel q(x)) + D(p(y|x) \parallel q(y|x)).$$

Proof: Write $p(x, y)/q(x, y) = p(x)p(y|x)/q(x)q(y|x)$ and expand $D(p(x, y) \parallel q(x, y))$.

Proof of Theorem 4: Let μ_n and μ'_n be two probability mass functions on the state space of a Markov chain at time n, and let μ_{n+1} and μ'_{n+1} be the corresponding distributions at time $n+1$. Let the corresponding joint mass functions be denoted by p and q. Thus $p(x_n, x_{n+1}) = p(x_n)r(x_{n+1}|x_n)$ and $q(x_n, x_{n+1}) = q(x_n)r(x_{n+1}|x_n)$, where $r(\cdot|\cdot)$ is the probability transition function for the Markov chain. Then by the chain rule for relative entropy, we have two expansions:

$$\begin{aligned} D(p(x_n, x_{n+1}) \parallel q(x_n, x_{n+1})) &= D(p(x_n) \parallel q(x_n)) + D(p(x_{n+1}|x_n) \parallel q(x_{n+1}|x_n)) \\ &= D(p(x_{n+1}) \parallel q(x_{n+1})) + D(p(x_n|x_{n+1}) \parallel q(x_n|x_{n+1})). \end{aligned}$$

Since both p and q are derived from the Markov chain, the conditional probability mass functions $p(x_{n+1}|x_n)$ and $q(x_{n+1}|x_n)$ are both equal to $r(x_{n+1}|x_n)$ and hence $D(p(x_{n+1}|x_n) \parallel q(x_{n+1}|x_n)) = 0$. Now using the non-negativity of $D(p(x_n|x_{n+1}) \parallel q(x_n|x_{n+1}))$, we have

$$D(p(x_n) \parallel q(x_n)) \geq D(p(x_{n+1}) \parallel q(x_{n+1}))$$

or

$$D(\mu_n \parallel \mu'_n) \geq D(\mu_{n+1} \parallel \mu'_{n+1}).$$

Consequently, the distance between the probability mass functions is decreasing with time n for any Markov chain.

Finally, if we let μ_n' be any stationary distribution μ, then $\mu_{n+1}' = \mu_n' = \mu$. Hence

$$D(\mu_n \parallel \mu) \geq D(\mu_{n+1} \parallel \mu),$$

which implies that any state distribution approaches the stationary distribution as time passes. The sequence $D(\mu_n \parallel \mu)$ is a monotonically non-increasing non-negative sequence and must therefore have a limit. It can be shown that the limit is actually 0 if the stationary distribution is unique.

We now specialize this result to obtain the result for entropy increase for Markov chains.

Theorem 5. *Consider a finite state Markov chain. Then $H(X_n) \nearrow$ for any initial distribution on X_0 if and only if the Markov transition matrix is doubly stochastic,* i.e., *if and only if the stationary distribution for the Markov chain is uniform.*

Proof. Let m denote the number of states. We note that

$$D(\mu_n \parallel \mu) = \sum_x \mu_n(x) \log \frac{\mu_n(x)}{(1/m)} = -H(\mu_n) + \log m.$$

Thus monotonic decrease in D induces a monotonic increase in H. Moreover, if the stationary distribution is unique, then $D \searrow 0$ by Theorem 4, and $H(X_n) \nearrow \log m$.

To see that there are initial distributions for which the entropy decreases when the doubly stochastic conditions are not satisfied, let the initial state X_0 have the uniform distribution. Then, $H(X_0) = \log m$, which is the maximum possible entropy. As time goes on, $H(X_n)$ will converge to $H(\mu)$, the entropy of the stationary distribution. Since the stationary distribution is not uniform for this example, the entropy must decrease at some time.

5.5 Stationary Markov Chains

A number of entropy-increase or second law theorems are true if the process is already in equilibrium. This seems strange since a process in equilibrium is stationary and the entropy will remain constant. However, the appropriate entropy is the conditional entropy of the future given the present. That is, if one cuts into a process in equilibrium and observes its state, the conditional uncertainty of the future will grow with time.

Theorem 6. *If X_n is stationary Markov chain, then the entropy $H(X_n)$ is constant, and*

$$H(X_n|X_1) \nearrow$$

with n.

Proof. Stationarity implies the marginal distributions are the same; thus $H(X_n)$ is constant. To prove monotonicity, we use conditioning and Markovity to show

$$H(X_{n+1}|X_1) \geq H(X_{n+1}|X_2, X_1) = H(X_{n+1}|X_2) = H(X_n|X_1),$$

where the first inequality follows from conditioning, the second from Markovity, and the last from stationarity.

5.6 Time Asymmetry

It is intriguing that a time asymmetric law like the second law of thermodynamics arises from a time symmetric physical process. This is not so puzzling if one believes that the initial conditions are extraordinary – for example, if one starts in a low entropy state. Thus, the time asymmetry comes from the asymmetry between the initial and final conditions.

However, if the process is in equilibrium, then the entropy is constant. Nothing could be more time symmetric than that. However, the conditional entropy $H(X_n|X_0)$ of a state at time n given the present, is monotonically increasing as observed in Theorem 6. There is no asymmetry in this because the conditional uncertainty $H(X_{-n}|X_0)$ of the past given the present is also monotonically increasing. In short, the observation of the state of a process in equilibrium at time 0 yields an amount of information about the past and about the future which monotonically dissipates with time. Thus there is symmetry: conditional entropy increases in both directions of time.

A true time asymmetry arises when we consider relative entropy. We have observed that the relative entropy distance $D(\mu_n \parallel \mu'_n)$ between two probability mass functions on the state space decreases with time for Markov chains. This is true even for transition matrices $r(x_{n+1}|x_n)$ that generate time reversible Markov chains. Here, then, is an apparent paradox. Why can't we reverse time, argue that the reversal of a Markov process is also Markov, and conclude that $D(\mu_n \parallel \mu'_n)$ decreases?

The answer is that the time reversed processes, although Markov, do not have the same transition matrices, *i.e.* $p(x_n|x_{n+1}) \neq q(x_n|x_{n+1})$, so the argument in the proof of Theorem 4 does not apply. We conclude that there is indeed a time-asymmetric behavior ($D(\mu_n \parallel \mu'_n) \searrow$) even for Markov processes generated from time symmetric physical laws.

5.7 The Relation of Time-Discrete and Time-Continuous Markov Chains

It should be pointed out that the study of time-continuous and time-discrete Markov chains may lead to different statements about the second law of thermodynamics. In a time-continuous Markov chain, one has intensities λ_{ij} for the Poisson rate at which transitions take place from state i to state j. A typical condition (Yourgrau et al. 1982) for the Ḣ theorem to hold, for example, would be the microscopic reversibility condition $\lambda_{ij} = \lambda_{ji}$.

A time-discrete Markov chain can be thought of as being generated by a time-continuous Markov process where the states are labeled $X_1, X_2, X_3, \ldots$ as the transi-

tions occur. Thus in our formalism

$$H(X_n) = H(X(t)|N(t) = n),$$

where $N(t)$ denotes the number of transitions that have taken place in the continuous-time Markov chain. It may well be that $H(X(t))$ increases while

$$H(X_n) = H(X(t)|N(t) = n)$$

does not increase with n. Thus, conditioning on the number of events may change the qualitative statement of the second law for such processes.

The discrete-time analysis in this paper deals with the event-driven rather than absolute time driven idea of time.

5.8 Summary of Relevant Results

We now gather these results in increasing order of generality.

The entropy $H(X_n)$ increases for a finite state Markov chain with an arbitrary initial distribution if the stationary distribution is uniform.

The conditional entropies $H(X_n|X_1)$ and $H(X_{-n}|X_0)$ increase with time for a stationary Markov chain.

The conditional entropy $H(X_0|X_n)$ of the initial condition X_0 increases for any Markov chain.

The relative entropy $D(\mu_n||\mu)$ between a distribution μ_n and the stationary distribution μ decreases with time for any Markov chain.

The conditional entropy $H(X_n|X_0, X_{-1}, \ldots)$ of a process at time n given the past up to time zero increases for any stationary process.

The conditional entropy $H(X_0|X_{-n}, X_{-(n+1)}, \ldots)$ of the present given the past increases as the past recedes for *all* processes.

5.9 Conclusions

While the second law of thermodynamics is only true for special finite-state Markov chains, it is universally true for Markov chains that the relative entropy distance of a given distribution from the stationary distribution decreases with time. This leads to a natural general statement of the second law for Markov chains: relative entropy decreases. To specialize this to the statement that entropy increases, one must restrict oneself to Markov chains with a uniform stationary distribution or to certain Markov chains with a suitable low entropy initial condition.

Finally, the second law of thermodynamics says that uncertainty increases in closed physical systems and that the availability of useful energy decreases. If one can make the concept of "physical information" meaningful, it should be possible to augment the statement of the second law of thermodynamics with the statement, "useful information becomes less available." Thus the ability of a physical system to

act as a computer should slowly degenerate as the system becomes more amorphous and closer to equilibrium. A perpetual computer should be impossible.

Discussion

Albrecht Relative entropy sounds like it has something to do with coarse graining. Can one think of it as being relative to a particular description or parameterization of phase space?

Cover Yes. Although the coarse graining is arbitrary, the natural distribution to place on the coarse graining is the stationary or equilibrium distribution. Then, the relative entropy distance of the current distribution from this stationary distribution is monotonically decreasing in time for any Markovian process. Consequently, the difficulty in measuring the difference of the current distribution from the equilibrium distribution increases with time. One should be aware, however, that coarse graining can destroy Markovity unless the partition is adroitly chosen.

Lloyd It is a commonly claimed feature of the psychological arrow of time that we know more about the past than we do about the future. How do you square this with the result that the amounts of information about the past and future that a system has in the present are equal in a stationary process?

Cover The key ingredient is that the statement holds in general only for processes in equilibrium. The theorem you are referring to states that the future is as uncertain as the past, conditioned on the present, for stationary Markov processes. That is to say, where we are going is conditionally as uncertain as how we got to where we are, at least for Markov processes in equilibrium. This intriguing symmetry is true even for time-asymmetric Markov chains. Apparently, this symmetry in information about past and future given the present follows entirely from the stationary Markovian assumption of the process and not from the time symmetry. That's the main point.

References

Cover, T., and Thomas, J. (1991) *Elements of Information Theory*, Wiley, New York.

Csiszar, I. (1967) Information type measures of difference of probability distributions. *Studia Sci. Math. Hung.*, **2**, 299-318.

Fritz, J. (1973) An information-theoretical proof of limit theorems for reversible Markov processes. *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, Verlag Dokumentation, Munich, 183-197.

Kullback, S. (1959) *Information Theory and Statistics*, Wiley, New York.

Leff, H., and Rex, A., eds., (1990) *Maxwell's Demon: Entropy, Information, Computing*, Princeton University Press, Princeton, New Jersey.

Mackey, M. (1992) *Time's Arrow: The Origins of Thermodynamic Behavior*, Springer-Verlag, New York.

Renyi, A. (1961) On measures of entropy and information. *Proc. 4th Berkeley Symposium on Math. Stat. and Probability*, **1**, 547-561, Berkeley.

Tisza, L. (1966) *Generalized Thermodynamics*, MIT Press, Cambridge, Mass.

Van Kampen, N.G., (1990) *Stochastic Processes and Chemistry*, North Holland, 1990.

Wehrl, A. (1978) General properties of entropy. *Rev. Mod. Phys.*, **50**, 221.

Yourgrau, W., Van der Merwe, A., Raw, G. (1982) *Treatise on Irreversible and Statistical Thermophysics*, Dover Publications, New York.