The Annals of Probability
1973, Vol. 1, No. 1, 43–65

SPECIAL INVITED PAPERS

AN APPLICATION OF ERGODIC THEORY TO PROBABILITY THEORY

BY DONALD S. ORNSTEIN

Stanford University

1. Introduction. The purpose of this paper is to try to introduce some structure in the class of stationary processes, and to take a step towards their classification. The main results of this paper are restatements of or minor modifications of some recent results in ergodic theory ([3], [5]–[10], [16], [17]).

The mathematical model for a stationary process[1]. A stationary process can be thought of as a box that prints out one letter each unit of time, where the probability of printing out a given letter may depend on the letters already printed out but is independent of the time (that is, the mechanism in the box does not change).

EXAMPLE 1. The box contains a roulette wheel. We spin the wheel once each unit of time, and print out the result. (We call such a process an independent process.)

EXAMPLE 2. The box contains a roulette wheel. We look at all possibilities for three consecutive spins of the wheel, and divide these into two classes. Each time we spin the wheel, we look at the last three spins and print out 1 if they fall in the first class, and 2 if they fall in the second class.

EXAMPLE 3. The box contains two coins, one of which is biased so that the probability of heads is not $\frac{1}{2}$. We divide all sequences of heads and tails of length three into two classes. At each unit of time, we look at the sequence of heads and tails which the box has printed out in the last three times. If the sequence lies in the first class, we flip the first coin and print out heads; if it comes up heads and tails, it comes up tails. If the previous three print-outs were in the second class, then we would use the second coin. This is an example of a three-step Markov process; that is, the last three print-outs determine the probability of printing out "heads," but if we know the last three print-outs, the conditional probability of "heads" is unaffected by an additional knowledge of what the process printed out in the past. (Note that Example 2 need not be an n-step Markov process for any n.)

EXAMPLE 4. The box contains a mechanical system such as a gas. At each unit of time, we make a fixed measurement on the system which has only a finite number of possible outcomes, and print out the outcome of the measurement. (If the outcome of the measurement were real-valued, we could divide the line into a finite number of sets, and print out which set the measurement fell into.)

EXAMPLE 5. A teleprinter. This prints out letters where the probability of printing out a given letter depends on what has already been printed. (Many possibilities will have probability 0 because they will not make sense.)

The mathematical model for a process is an invertible, measure preserving transformation T, acting on a space X of total measure 1, *and a partition* $\mathcal{P}$ *of* X *into a*

Received April 12, 1972; revised August 24, 1972.

[1] For most of this paper, we will restrict ourselves to discrete-time processes, with only a finite number of possible outputs. The theory will work for continuous time and more general state spaces, but I think it worthwhile studying the simplest case first.

finite number of disjoint sets. We shall denote this process by $(\mathcal{P}, T)$. The reason for this is as follows: Starting with a process, we get a measure on finite sequences; that is, if we fix K_1 and K_2, we get a probability measure on sequences $\{\alpha_i\}$, $K_1 \leqq i \leqq K_2$, where the α_1 are possible outcomes of the process at time i. The measures must be consistent, in the sense that the probability of a fixed sequence $\{\alpha_1\}$, $K_1 \leqq i \leqq K_2$, obtained from the measure on a sequence $\{\beta_i\}$, $\bar{K}_1 \leqq i \leqq \bar{K}_2$ where $\bar{K}_2 \geqq K_2$ and $\bar{K}_1 \leqq K_1$, should be independent of the choice of $\bar{K}_1$ and $\bar{K}_2$. A simple special case of the Kolmogorov extension theorem says that the above measure extends uniquely to a measure on all doubly-infinite sequences. Thus we get a measure space X whose points are sequences $\{\alpha_i\}_{-\infty}^{\infty}$. (Each point can thus be thought of as one possible print-out of the process when the process runs forever in both directions.) Let T be the transformation that shifts each sequence one to the left. ($T\{\alpha_i\} = \{\beta_i\}$ where $\beta_i = \alpha_{i+1}$.) The assumption that the process is stationary is equivalent to assuming that T is measure-preserving. Let $\mathcal{P}$ be the partition of X, which partitions the sequences $\{\alpha_i\}$ in X according to their 0th coordinate α_0. (We can recover α_k by seeing which atom of $\mathcal{P}$ contains $T^k\{\alpha_i\}$.)

If we start with a transformation T and a partition $\mathcal{P}$, then to each point x we associate the sequence $\{\alpha_i\}$, where α_i is the atom of $\mathcal{P}$ containing $T^i(x)$. We thus have a measure on all finite sequences $\{\alpha_i\}$, $K_1 \leqq i \leqq K_2$ (or on all infinite sequences $\{\alpha_i\}$), and this uniquely determines a stationary process.

We say that $\mathcal{P}$, T and $\bar{\mathcal{P}}$, $\bar{T}$ give the same process (or $\mathcal{P}, T \sim \bar{\mathcal{P}}, \bar{T}$) if they give the same measure to all finite sequences.

The physical meaning of the transformation T is especially clear if the $\mathcal{P}$, T arises from a measurement on a mechanical system (as in Example 4). In that case, each configuration[2] of the system can be represented as a point in a measure space X (called the phase space), where the measure of a set E of configurations corresponds to the probability that the system is in a configuration in E. If we know the configuration of the system at time 0, then the configuration of the system is determined at time 1. We thus get a transformation T of X. If the system is in an equilibrium state, then the transformation will be measure-preserving. Any measurement on the system with a finite number of outcomes corresponds to a partition $\mathcal{P}$ of X (the parts where the measurement has a fixed outcome). The model for the process arising from this measurement is then $\mathcal{P}$, T.

In general, if we start with a process and construct its model $\mathcal{P}$, T, then the space X could be thought of as some "system" evolving according to T for some $\mathcal{P}$-measurement. (The infinite sequence of $\mathcal{P}$-measurements will determine the point in X uniquely, and vice versa.)

The Birkoff ergodic theorem implies that, if we look at a process long enough, any fixed string will occur with a fixed frequency. (More precisely, let B be a measurable set. Let $F(n, x)$ be the number of $0 \leqq i \leqq n$ such that $T^i x \in B$. Then $\lim_{n\to\infty} 1/nf(n, x)$ exists for a.e. x. Now let B be an atom in $\bigvee_0^K T^i\mathcal{P}$, and we get the above assertion.) If T is ergodic (that is, the only sets invariant under T have measure 0 or 1), then the limiting frequency of a string will be the same as the probability of that string. (More precisely, $\lim_{n\to\infty} 1/nf(n, x) = m(B)$ for a.e. x.) This

[2] For example, if the system is a hard sphere gas, then a configuration is determined by specifying the position and momentum of each molecule.

means that, if T is ergodic, we can recover all the information about a process from almost every instance of the process, and that probabilities can be identified with frequencies. For the rest of this paper, we will consider only processes $\mathcal{P}$, T where T is ergodic. (Any other process will be some sort of average of ergodic processes. See the lemma in Appendix 1.)

2. A notion of distance between processes. We will now introduce a metric on processes. Let $\mathcal{P}$, T and $\mathcal{Q}$, S be processes where $\mathcal{P}$ and $\mathcal{Q}$ are ordered partitions having the same number of atoms (we will assume S and T to be ergodic. $\bar{d}((\mathcal{P}, T), (\mathcal{Q}, S))$ will be the infimum of the α such that we can find an ergodic transformation U acting on a space X, and partitions $\bar{\mathcal{P}}$ and $\bar{\mathcal{Q}}$ of X such that:

(1) $|\bar{\mathcal{P}} - \bar{\mathcal{Q}}| \leqq \alpha$ (here $|\bar{\mathcal{P}} - \bar{\mathcal{Q}}|$ denotes the measure of $X - \bigcup_i (\bar{\mathcal{P}}^i \cap \bar{\mathcal{Q}}^i)$; all partitions are ordered, and $\bar{\mathcal{P}}^i$ and $\bar{\mathcal{Q}}^i$ are the ith atom of $\bar{\mathcal{P}}$ and $\bar{\mathcal{Q}}$);

(2) $\bar{\mathcal{P}}$, $U \sim \mathcal{P}$, T and $\bar{\mathcal{Q}}$, $U \sim \mathcal{Q}$, S (i.e., $\bar{\mathcal{P}}$, U and $\mathcal{P}$, T assign the same measure to all finite (or infinite) sequences; in other words, they represent the same process, and $\bar{\mathcal{Q}}$, U and $\mathcal{Q}$, S also assign the same measure to sequences).

The idea of the above definition is the following: We have one process U that prints out two letters each unit of time. If we look at the first letter only, we get the $\mathcal{P}$, T process, and if we look at the second letter, we get the $\mathcal{Q}$, S process. Furthermore, the probability that the two letters are different is $\leqq \alpha$.

We will now define the distance in a different way. We will show in Appendix 1 that this definition is equivalent to the first. $\bar{d}((\mathcal{P}, T), (\mathcal{Q}, S))$ will be the sup of the α such that: given ϵ, there is an N such that, if $K > N$, we can find two collections C_1, C_2 of sequences of length K such that C_1 has measure $> 1 - \epsilon$ under $\mathcal{P}$, T, and C_2 has measure $> 1 - \epsilon$ under $\mathcal{Q}$, S, and any sequence in C_1 differs from any sequence in C_2 in more than αK places. (The reason for introducing the second definition of $\bar{d}$ is the following: It is clear from the first definition that, if the $\bar{d}$-distance is small, the two processes are in some sense close to each other. However, it is not clear what it means for $\bar{d}$ to be large. The second definition says that, if $\bar{d} > \alpha$, one could tell the two processes apart even if each of the processes were changed $< \frac{1}{2}\alpha$ percentage of the time.)

There is another topology on processes called the "vague" topology. Let $\mathcal{P}_i$, T_i and $\mathcal{P}$, T be ergodic processes, and assume that all of the $\mathcal{P}_i$ and $\mathcal{P}$ are ordered partitions with the same number of atoms. We say that $\mathcal{P}_i$, T_i converges to $\mathcal{P}$, T in the vague topology if, for each fixed K, the measure $u_{i,K}$, that $\mathcal{P}_i$, T_i puts on sequences of length K, converges to u_K, the measure that $\mathcal{P}$, T puts on sequences of length K. (Note that convergence in $\bar{d}$ implies convergence in the vague topology, but not conversely. A simple way to see that the converse is false is the following: We can pick a sequence of periodic processes that converge in the vague topology to an independent process. If we had convergence in $\bar{d}$, then the entropies would converge, but periodic processes have entropy 0.)

3. Isomorphism. Given T, we say that the *partition* $\mathcal{P}$ *generates* if any measurable set can be approximated arbitrarily well by a collection of atoms in $\bigvee_{-K}^{K} T^i \mathcal{P}$ for sufficiently large K. If we start with a process (i.e., a measure on finite sequences) and form its model $\mathcal{P}$, T, then $\mathcal{P}$ will automatically generate.

Suppose $\mathcal{P}$, T and $\mathcal{Q}$, S are processes where $\mathcal{P}$ and $\mathcal{Q}$ both generate. We say that

T and S (the transformations they induce) are isomorphic[3] if there is a transformation U and partions $\bar{\mathcal{P}}$ and $\bar{\mathcal{Q}}$ such that:

(1) $\bar{\mathcal{P}}$ and $\bar{\mathcal{Q}}$ each generate under U;

(2) $\bar{\mathcal{P}}$, $U \sim \mathcal{P}$, T and $\bar{\mathcal{Q}}$, $U \sim \mathcal{Q}$, S (i.e., $\bar{\mathcal{P}}$, U induces the same measure on finite sequences as $\mathcal{P}$, T does, and $\bar{\mathcal{Q}}$, U induces the same measure on finite sequences as $\mathcal{Q}$, S does. We introduce U for symmetry. U could be taken to be T or S).

If we think of U as a "system," and $\mathcal{P}$ and $\mathcal{Q}$ as measurements on U, then the condition that $\bar{\mathcal{P}}$ and $\bar{\mathcal{Q}}$ generate can be interpreted as saying that, with probability 1, the (doubly-infinite) sequence of $\bar{\mathcal{P}}$-measurements determines the sequence of $\bar{\mathcal{Q}}$-measurements, and the $\bar{\mathcal{Q}}$-measurements determine the $\bar{\mathcal{P}}$-measurements. (We can see this as follows: We can find a sequence of partitions $\bar{\mathcal{P}}_n$, converging to $\bar{\mathcal{P}}$ such that each atom of $\bar{\mathcal{P}}_n$ is made up of atoms of $\vee_{-n}^{n} T^i\bar{\mathcal{Q}}$. Thus for a.e. x, if we know which atom of $T^i\bar{\mathcal{Q}}$ contains x for all i (or, equivalently, which atom of $\bar{\mathcal{Q}}$ in which T^ix lies for all i), then we know which atom of $\bar{\mathcal{P}}$ (or $T^i\bar{\mathcal{P}}$) contains x.)

Thus if $\mathcal{P}$, T and $\mathcal{Q}$, S are isomorphic, then they can be realized as measurements on a "system," each of which determines the other. If $\mathcal{P}$, T is a measurment on a real mechanical system, then $\mathcal{Q}$, S can be realized as a measurement on the same mechanical system, which determines and is determined by the $\mathcal{P}$-measurments.

There is another interpretation of isomorphism in terms of "codes." A finite code c_l of length l will be a mapping from sequences $\{\alpha_1\}$ to sequences $\{\beta_i\} = c_l\{\alpha_i\}$, obtained as follows: Suppose there are k possibilities for each β_i. We then divide all possible sequences of length $2l + 1$ of α's into k classes. We then look at $\{\alpha_i\}_{i-l}^{i+l}$, and see which class it belongs to, letting that be β_i. We say that a sequence of finite codes c_n converges on a process $\mathcal{P}$, T if, for each sequence $\{\alpha_i\}$ generated by $\mathcal{P}$, T (except for a collection of sequences $\{\alpha_i\}$ which have measure 0 under $\mathcal{P}$, T), we have that the jth coordinate of $c_n\{\alpha_i\}$ will be the same for all n large enough. (The above holds for each integer j.) If a sequence of codes c_n converges on a process $\mathcal{P}$, T, we will call their limit a coding of the process $\mathcal{P}$, T. A coding of the process $\mathcal{P}$, T gives us another process.

It is easy to see that T and S are isomorphic if and only if there are codes c and $\bar{c}$ on $\mathcal{P}$, T and $\mathcal{Q}$, S, respectively, such that c maps almost every sequence $\{\alpha_i\}$ generated by $\mathcal{P}$, T to a sequence $\{\beta_i\}$ generated by $\mathcal{Q}$, S, and $\bar{c}$ maps a.e. sequence $\{\beta_i\}$ to a sequence $\{\alpha_i\}$. For a.e. $\{\alpha_i\}$, $\bar{c}c\{\alpha_i\} = \{\alpha_i\}$, and for a.e. $\{\beta_i\}$, $c\bar{c}\{\beta_i\} = \{\beta_i\}$.

4. Entropy. The Shannon-McMillan-Breiman theorem says that, if T is ergodic and $\mathcal{P}$ a finite partition, there is a number associated with $\mathcal{P}$, T, denoted by $E(\mathcal{P}, T)$, the "entropy of $\mathcal{P}$ relative to T," or the "entropy of the process $\mathcal{P}$, T," with the following property: Given $\epsilon > 0$, there is an N such that, if $n > N$, the measure of each atom in $\vee_0^{n-1} T^i\mathcal{P}$ is between $\frac{1}{2}^{(E(\mathcal{P},T)+\epsilon)n}$ and $\frac{1}{2}^{(E(\mathcal{P},T)-\epsilon)n}$, except for a collection of atoms the measure of whose union is $<\epsilon$.

The Shannon-McMillan-Breiman theorem says that most strings of length n put out by the process have roughly the same probability, namely $\frac{1}{2}^{E(\mathcal{P},T)n}$

[3] If we assume that all our measure spaces are non-pathological, that is, Lebesgue spaces ([14]) (Lebesgue spaces are spaces which are measure-theoretically the same as the unit interval, examples being the unit square, and the space X which we constructed, starting with a process), then the above definition is equivalent to the usual definition which says that T and S are isomorphic if there is a 1–1 invertible measure-preserving map of X (the space on which T acts) onto Y (the space on which S acts), such that $\varphi T(x) = S\varphi(x)$ for all x in X.

This theorem originally came out of attempts to code messages efficiently. It is useful because it tells us that, if we are willing to ignore messages which will arise with small probability, then we only have to code $2^{E(\mathcal{P},T)n}$ messages of length n.

Kolmogorov showed that, if $\mathcal{P}$ generates, then $E(\mathcal{P}, T) = \sup E(\mathcal{Q}, T)$ when sup is taken over all finite partitions $\mathcal{Q}$. He then defined $E(T) = \sup E(\mathcal{Q}, T)$.

If $\mathcal{Q}, T$ is an independent process, arising from spinning a roulette wheel whose slots have probabilities $\mathcal{P}^i$, $1 \leqq i \leqq k$, then $E(\mathcal{P}, T) = \sum_{i=1}^{k} \mathcal{P}^i \log \mathcal{P}^i$.

The above two facts show that, if we have two independent processes with different $\sum \mathcal{P}^i \log \mathcal{P}^i$ (or different entropies), then they cannot be obtained as measurements on a system, each of which determines the other. (If each measurement determined the other, they would both generate the same σ-algebra, and hence have the same entropy by Kolmogorov's theorem.)

It was shown in [5] that, if two independent processes have the same entropy, they can be obtained as measurements on the same system, each of which determines the other (i.e., there is then one transformation and two generating partitions, each representing one of the original two processes). This question was open for many years, and stimulated important and deep results by Sinai, Mešalkin, Adler and Weiss, and others. Its solution is the starting point for the main results of this paper.

It is not hard to see that the processes $\mathcal{P}, T$ of 0-entropy are exactly those processes that are deterministic in the following sense: If we know the entire past history of the process, then we can predict the next output with probability 1 (that is, $\mathcal{P} \subset \bigvee_1^\infty T^i\mathcal{P}$). This implies the stronger statement that, if we know the entire history of the process before time $-n$, then we can predict the output at time 0 with probability 1. (Similarly, if we know enough of the history before time $-n$, then we can predict the output at time 0 with arbitrarily high probability.)

5. Classes of processes—*B*-processes. We will now divide the stationary processes into several classes. The first class that we will consider will be the class of "*B*-processes" defined as follows: We say that T is a Bernoulli shift if T is the transformation arising from an independent process (that is, there is a finite partition $\mathcal{P}$ that generates under T, and the $T^i\mathcal{P}$ are independent; the last condition says that, if A_i is an atom in $T^i\mathcal{P}$, $-n \leqq i \leqq n$, then the measure of the intersection of the A_i is the product of their measures). We say that $\mathcal{P}, T$ is a *B*-process if T is a Bernoulli shift, and $\mathcal{P}$ a partition that generates under T (but is not necessarily independent).

THEOREM 1. *If T is a Bernoulli shift, and $\mathcal{P}$ any finite partition (not necessarily a generator), then $\mathcal{P}, T$ is a B-process (that is, there is a B-process that gives the same measure on sequences as does $\mathcal{P}, T$).*

This theorem follows immediately from [6], where we prove that, if T is a Bernoulli shift and $\mathcal{P}$ any finite partition, then there is a partition $\mathcal{Q}$ such that $T^i\mathcal{Q}$ are independent, and the $T^i\mathcal{Q}$ generate exactly the same σ-algebra as does $T^i\mathcal{P}$.

If we state the definition of *B*-process in terms of codes, we have that the *B*-processes are exactly those processes that arise from a coding of an independent process. Thus, in a certain sense, these processes can be thought of as those processes whose random mechanism can be taken to be a roulette wheel.

THEOREM 2. *The class of B-processes is closed in the $\bar{d}$-metric.*

The proof of this theorem will be given in Appendix 2. It is a modification of the proof that two Bernoulli shifts with the same entropy are isomorphic.

COROLLARY. *The B-processes are exactly those processes that can be approximated arbitrarily well in the $\bar{d}$-metric by finite codings of roulette wheels.*

THEOREM 3. *If a sequence of (ergodic) processes $\mathcal{P}_i$, T_i converges to a B-process $\mathcal{P}$, T in the vague topology, and $E(\mathcal{P}_i, T_i) \to E(\mathcal{P}, T)$ and $\mathcal{P}$, $\mathcal{P}_i$ all have the same number of atoms, then the sequence converges in the $\bar{d}$-metric. Furthermore, the B-processes are the only process for which the above holds.*

This follows immediately from Lemma 3′ ([6])[4].

Theorem 3 says that a B-process $\mathcal{P}$, T is determined to within ϵ in the $\bar{d}$-metric by a finite number of parameters. That is, given ϵ, there is a K and δ such that, if $\mathcal{Q}$, S is any other ergodic process whose entropy is within δ of the entropy of $\mathcal{P}$, T, and such that the measure which $\mathcal{Q}$, S gives to sequences of length K is within δ of the measure which $\mathcal{P}$, T gives to sequences of length K, then $\bar{d}((\mathcal{P}, T), (\mathcal{Q}, S)) < \epsilon$.

Theorem 3 can be viewed as a measure-theoretic analog of structural stability.

A process is said to be an n-step Markov process (or a multi-step Markov process) if the probability of printing out a given letter, given the previous k print-outs for $k > n$, depends only on the previous n print-outs. An example was given in Example 3. A process is said to be mixing if, given $\{\alpha_i\}_1^m$ and $\{\beta_i\}_1^m$, the probability that $\gamma_i = \alpha_i$, $1 \leqq i \leqq m$, and $\gamma_{K+i} = \beta_i$, $1 \leqq i \leqq m$, tend to the product of the probabilities of $\{\alpha_i\}_1^m$ and $\{\beta_i\}_1^m$ as $K \to \infty$. Here γ_i is the output of the process at time i.

THEOREM 4. *The B-processes are the closure in the $\bar{d}$-metric of the multi-step, mixing Markov processes.*

Theorem 4 follows easily from [3] and Theorems 2 and 3. We can find a sequence of multi-step Markov processes that converge to any process in entropy and in the vague topology. If the process is a B-process, then Theorem 3 implies that we get convergence in the $\bar{d}$-metric. Because of [3], we have that multi-step Markov processes are B-processes, and by Theorem 2 the closure of these processes in the $\bar{d}$-metric contains only B-processes. (More details will be given in the appendix.)

Theorem 4 can be interpreted as saying that the B-processes are exactly those processes where the influence of the sufficiently distant past becomes negligible. (We can see this as follows: Using the first definition of $\bar{d}$, we can start with a B-process and, by modifying it with probability $< \epsilon$, we can obtain a multi-step Markov process. On the other hand, if a process is not a B-process, then it has distance $> \alpha > 0$ from all n-step Markov processes. Thus, if we modify the process by ignoring what it did at times more than n units in the past, we must get a process which is at distance α or more from our original process. Therefore, information about what our original process did at times more than n units in the past must have influenced the print-out at least α percentage of the time.)

THEOREM 5. *Two B-processes induce isomorphic transformations if and only if they have the same entropy.*

This is simply a restatement of the isomorphism theorem for Bernoulli shifts ([5]). If we start with a process and form its model $\mathcal{P}$, T, then $\mathcal{P}$ will automatically generate. Thus, the entropy of $\mathcal{P}$, T will equal the entropy of T. If $\mathcal{P}$, T is a B-process, then by definition T is isomorphic to some Bernoulli shift.

[4] Here we know only the theorem for mixing processes, but the proof can be modified to work for ergodic processes.

Mechanical systems. The proof of the isomorphism theorem for Bernoulli shifts gives a criterion that allows one to show that certain specific transformations are isomorphic to Bernoulli shifts. (In the case of a flow S_t, we will say that S_t is Bernoulli if, for each fixed t, the transformation S_t is a Bernoulli shift.)

One mechanical system for which the above works is geodesic flow on a surface of negative curvature. B. Weiss and I were able to show [18] (by using the above-mentioned criterion, together with results of Sinai and Anosov) that:

THEOREM 6. *Geodesic flow on a surface of negative curvature is Bernoulli.*

The above are called mechanical systems because, as noted by Kolmogorov ([4]), there are geodesic flows on surfaces of negative curvature which can be physically realized by a particle constrained to move on a surface in 3-space, subject to external forces produced by a finite number of centers of attraction or repulsion, which are placed near the surface.

What are the implications of showing that a mechanical system is Bernoulli? First of all, it means that any measurement (with a finite number of possible outcomes) performed on the system (at multiples of a fixed unit of time) gives rise to a B-process. Because there is always some experimental error, we really know the processes only to within ϵ in the $\bar{d}$-metric. Thus Theorem 4 tells us that any measurement on a Bernoulli system is indistinguishable from a multi-step Markov process (or a finite coding of a roulette wheel).

A picturesque consequence of a system's being Bernoulli is the following: Suppose we take a movie of such a mechanical system, and watch the movie on a TV screen. (This has the effect of looking at the system at discrete times, and watching some measurement with only a finite number of possibilities.) This process will be "essentially" indistinguishable (i.e., it could be approximated arbitrarily well in the $\bar{d}$-metric) from a process which arises from a finite coding of a roulette wheel.

I think it is clear that Bernoulli systems are the most "random" possible, and thus results of the type of Theorem 6 tend to explain why statistical methods are relevant in the study of mechanical systems (whose laws of motion are deterministic).

As pointed out previously, a mechanical system gives rise to a measure-preserving transformation on a measure space (called its phase space). If we show that the system is Bernoulli, then, given its entropy, we know exactly what measure-preserving transformation the system gives rise to. (Previous results of Hedlund, Hopf, Anosov, and Sinai show that certain mechanical systems give rise to transformations with various properties, such as ergodicity, mixing, and the Kolmogorov property.) It is somewhat surprising that the transformation arising from a geodesic flow is in some sense the simplest possible.

Actually, there are even stronger results along these lines. In [10], it is shown that, if S_t and S_t are Bernoulli flows, and if $E(S_1) = E(S_1)$, then S_t and S_t are isomorphic. Thus Theorem 6 says that we know exactly (measure-theoretically) what the flow is that arises from a geodesic flow on a surface of negative curvature. Also in [10] is given a simple description of a Bernoulli flow.

I think that there is a great deal of work yet to be done along the line of finding out in various specific cases whether a process is a B-process. In particular, it would be very interesting to see which mechanical systems are Bernoulli flows.

The preceding results have analogs which will hold in the case of continuous-time, where the model for a continuous-time B-process is a Bernoulli flow and a finite

partition. For example, continuous-time Markov processes on a finite state space or Brownian motion contained in a rectangular box give rise to Bernoulli flows (of infinite entropy) ([12]). I think that there is much work to be done in this direction also.

6. Classes of processes—*K*-processes. The *K*-processes or Kolmogorov processes are those processes which satisfy the 0–1 law. We say that a process $\mathcal{P}$, T satisfies the 0–1 law if $\bigcap_{n=1}^{\infty} \bigvee_{j=n}^{n} T^j\mathcal{P}$ is trivial[5] (that is, $\bigvee_{i=n}^{\infty} T^i\mathcal{P}$ is the class of measurable sets generated by the $T^i\mathcal{P}$ for $i \geqq n$; the only sets that are contained in the above classes for all n have measure either 0 or 1).

Probabilistically, this says that any event which can be determined by what the process does in the arbitrarily distant past has probability 0 or 1. This means that the process is completely nondeterministic in the sense that no event can be predicted from the arbitrarily distant past. If a process is not a *K*-process, then we can find an N and divide the sequences of length N into two classes, each having probability $\geqq \frac{1}{4}$, and such that we can predict which class the next N outputs will belong to with probability $>99/100$ by knowing the history of the process before time $-l$, no matter how large l is (all we need to know is a finite number of outputs before time $-l$, but as l gets larger, this number will in general increase). The simplest example of a *K*-process is an independent process.

There are some beautiful results about *K*-processes, due to Sinai and Rochlin which I will now describe. The main result is the following: If T is any ergodic transformation, and if $\mathcal{P}$ is a finite partition which generates, then the σ-algebra $\bigcap_{n=1}^{\infty} \bigvee_{j=n}^{\infty} T^j\mathcal{P}$ consists of exactly those sets E, such that the entropy of the partition consisting of E and its complement relative to T is 0. This result has several consequences:

The class of K-processes contains the class of B-processes. We can see this as follows: If $\mathcal{P}$, T is a *B*-process, then there is a partition $\mathcal{Q}$ such that $\mathcal{Q}$ generates, and $T^i\mathcal{Q}$ are independent. Therefore, $\bigcap_{n=1}^{\infty} \bigvee_{j=n}^{\infty} T^i\mathcal{Q}$ is trivial. Therefore, there are no partitions of entropy 0. Therefore, $\bigcap_{n=1}^{\infty} \bigvee_{j=n}^{\infty} T^i\mathcal{P}$ is trivial.

Another consequence is the following: We say that T is a *K*-automorphism if there is a generator $\mathcal{P}$ such that $\mathcal{P}$, T is a *K*-process. The Sinai-Rochlin theorem implies that T is a *K*-automorphism if and only if there are no partitions of entropy 0 (if T is thought of as a "system," then no measurements made on this system are deterministic).

Kolmogorov conjectured that all *K*-processes were *B*-processes. In [8], we show this to be false.

THEOREM 7. *There is a K-process that is not a B-process.*

Because of our results on *B*-processes, there is an $\alpha > 0$ such that the process in the above theorem is at distance α from any *B*-process (and, in particular, at distance α from any n-step Markov process or any finite coding of a roulette wheel).

In a joint paper with P. Shields ([11]), it is shown that there is an uncountable class of *K*-processes that are not *B*-processes. All the processes in this class have the same entropy, no two are isomorphic, and any two are at distance $\geqq \alpha$ in the $\bar{d}$-metric.

---

[5] The usual definition involves countable partitions. Because of a theorem of Krieger (*Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 1971), the two definitions are equivalent.

Furthermore, any neighborhood in the vague topology of any process in the above class contains uncountably many members of the class (this shows that K-processes strongly violate Theorem 3 of the previous section). In addition, the transformation arising from these processes is not isomorphic to their inverses.

The class of K-processes is closed in the $\bar{d}$-metric. (This is not hard to see, and we will sketch a proof. We will argue by contradiction. Suppose $\mathcal{P}$, T is not a K-process. Then there would be $\mathcal{Q}$ such that $E(\mathcal{Q}, T) = 0$, and a $\mathcal{Q}_1$ close to $\mathcal{Q}$ such that $E(\mathcal{Q}_1, T^n)$ is small for all n, and $\mathcal{Q}_1 \in \bigvee_{-K}^{K} T^i\mathcal{P}$ for some K. Now choose a K-process $\bar{\mathcal{P}}$, $\bar{T}$ which approximates $\mathcal{P}$, T so well in the $\bar{d}$-metric that, if we define $\bar{\mathcal{Q}}_1$ to be the partition in $\bigvee_{-K}^{K} \bar{T}^i\bar{\mathcal{P}}$ corresponding to $\mathcal{Q}_1$, then $\bar{d}[(\mathcal{Q}_1, T^n), (\bar{\mathcal{Q}}_1, \bar{T}^n)]$ is small for all n. But since $\bar{T}$ is a K-automorphism, $(\bar{T}^n)^i\bar{\mathcal{Q}}_1$ will be ϵ-independent for large enough n, and hence $E(\bar{\mathcal{Q}}_1, \bar{T}^n)$ is close to $E(\bar{\mathcal{Q}}_1)$. This gives a contradiction.)

7. Classes of processes—*0*-entropy. As we have already pointed out, the processes of 0-entropy are exactly those processes that are deterministic. This class is also closed in the $\bar{d}$-metric.

8. Classes of processes—those remaining. Is every process in some sense obtained from a process of 0-entropy and a K-process? There was a conjecture along these lines, due to Pinsker, which can best be stated in terms of transformations: Is every ergodic transformation the direct product of a K-automorphism and a transformation of 0-entropy? An equivalent question is the following: Can every process be represented as a measurement on a "system" which is the direct product of a K-system and a system of 0-entropy? This turns out to be false, and in [9] we show

THEOREM 8. *There is a mixing transformation (of non-0-entropy, and not a K-automorphism] which is not the direct product of a transformation of 0-entropy and a K-automorphism.*

APPENDIX 1

We will prove here the equivalence of the two definitions of the $\bar{d}$-metric. Before doing this, however, we shall introduce a third definition.

Third definition of $\bar{d}$. Let $\mathcal{P}_i$, $1 \leqq i \leqq n$, be a sequence of ordered partitions of X, each having k atoms. Let $\mathcal{Q}_i$, $1 \leqq i \leqq n$, be a sequence of ordered partitions of Y, each of which has k atoms. We define $\bar{d}(\{\mathcal{P}_i\}_1^n, \{\mathcal{Q}_i\}_1^n)$ as follows: Let φ and ψ be 1–1 measure-preserving maps of X and Y, respectively, onto Z. Let $\bar{d}_{\varphi,\psi}(\{\mathcal{P}_i\}_1^n, \{\mathcal{Q}_1^n) = 1/n \sum_{i=1}^n |\varphi(\mathcal{P}_i) - \psi(\mathcal{Q}_i)|$, where $|\varphi(\mathcal{P}_i) - \psi(\mathcal{Q}_i)|$ denotes the measure of the set of points in Z which lie in the image under φ of the rth atom of $\mathcal{P}_i$, and the image under ψ of the sth atom of $\mathcal{Q}_i$ where $r \neq s$. We define $\bar{d}(\{\mathcal{P}_i\}_1^n, \{\mathcal{Q}_i\}_1^n)$ as $\inf_{\varphi,\psi} \bar{d}_{\varphi,\psi}(\{\mathcal{P}_i\}_1^n, \{\mathcal{Q}_i\}_1^n)$.

We now define $\bar{d}((\mathcal{P}, T), (\mathcal{Q}, S))$ as $\sup_n \bar{d}(\{T^i\mathcal{P}\}_1^n, \{S^i\mathcal{Q}\}_1^n)$.

Actually, it is not hard to see that the above sequence has a limit which is equal to the sup, because if $\bar{d}(\{T^i\mathcal{P}\}_1^r, \{S^i\mathcal{Q}\}_1^r) = \alpha$, then $\bar{d}(\{T^i\mathcal{P}\}_1^{sr}, \{S^i\mathcal{Q}\}_1^{sr}) > \alpha$ for all s, and hence, for n large enough, $\bar{d}(\{T^i\mathcal{P}\}_1^n, \{S^i\mathcal{Q}\}_1^n) > \alpha - \epsilon$. It is the existence of the above limit which shows that the above definition of $\bar{d}$ satisfies the triangle inequality, and hence yields a metric.

We will now show that the three definitions of $\bar{d}$ are equivalent. For this purpose, we will call the first definition in the paper $\bar{d}_1$, the second $\bar{d}_2$, and the one given in this appendix $\bar{d}_3$.

$d_1 = d_3$. Suppose that $d_3((\mathcal{P}, T), (\mathcal{Q}, S)) = \gamma$. We will show that

$$d_1((\mathcal{P}, \tau), (\mathcal{Q}, S)) = \gamma$$

(it is obvious that $d_1((\mathcal{P}, T), (\mathcal{Q}, S))$ cannot be $<\gamma$). To do this, we will put a shift invariant measure on sequences of pairs $\{\alpha_i, \beta_i\}$, where each α_i and β_i is an integer between 1 and k, and such that the measure induced on the sequences $\{\alpha_i\}$ is the same as that induced by $\mathcal{P}$, T, and the measure induced on the sequence $\{\beta_i\}$ is the same as that induced by $\mathcal{Q}$, S, and, furthermore, the measure of all sequences such that $\alpha_0 \neq \beta_0$ is $\leqq \gamma$. U will be the shift on $\{\alpha_i, \beta_i\}$. $\bar{\mathcal{P}}$ will partition the sequences according to α_0, and $\bar{\mathcal{Q}}$ will partition the sequences according to β_0. (This U may not be ergodic, but we will worry about that later.)

To get the measure on sequences $\{\alpha_i, \beta_i\}$, we will first define, for each n and m, a measure on sequences $\{\alpha_i, \beta_i\}_{-m}^{n}$ in such a way that the measure assigned to a particular sequence α_i, β_i, $-m \leqq i \leqq m$, is the same as the measure assigned to the union of all sequences $\{\alpha_i, \beta_i\}_{-\bar{m}}^{\bar{n}}$, $\bar{n} \geqq n$, $\bar{m} \geqq m$, which agree with our original sequence for $-m \leqq i \leqq n$. (We will then get a measure on all infinite sequences by a simple and obvious case of the Kolmogorov extension theorem.)

We will get the measure on finite sequences as follows: For each ψ, φ, and n, $\{\psi(T^i\mathcal{P})\}_1^n$, $\{\varphi(S^i\mathcal{Q})\}_1^n$ gives us a measure on sequences of α_i, β_i of length n. It also gives us a measure on sequences of length r for $r \leqq n$. (Simply average the measures given by $\{\psi(T^i\mathcal{P})\}_l^{l+r}$, $\{\varphi(S^i\mathcal{Q})\}_l^{l+r}$ as l goes from 1 to $n - r$, giving each of the above measures weight $1/n - r$.) We will now choose a sequence $\varphi_{n_i}, \psi_{n_i}, n_i$ such that, for each r, the measure on sequences of length r, defined as above by the $\varphi_{n_i}, \psi_{n_i}, n_i$, converges. (This can be done by a standard diagonalization.) Furthermore, $\lim_{i\to\infty} 1/n_i \sum_{j=1}^{n_i} |\psi_{n_i}(T^j(\mathcal{P})) - \varphi_{n_i}(S^j(\mathcal{Q}))| = \gamma$. The limit gives us a measure on sequences of length r for all r, and, in particular, gives us a measure on $\{\alpha_i, \beta_i\}_{-m}^{n}$. It is easy to see that the consistency conditions are satisfied, and that the measures therefore extend to a measure on all infinite sequences $\{\alpha_i, \beta_i\}$. It is also easy to see that the measure is invariant under the shift U. It is also easy to see that the measure of the set of sequences, such that $\alpha_0 \neq \beta_0$, is γ.

We will now show that U could be chosen to be ergodic. We will do this as follows: Each x corresponds (under U and $\bar{\mathcal{P}} \vee \bar{\mathcal{Q}}$) to a doubly-infinite sequence $\{{}_x\alpha_i, {}_x\beta_i\}$ (${}_x\alpha_i$ being the atom of $\bar{\mathcal{P}}$ which contains $U^i x$). By the ergodic theorem, for a.e. x, each sequence of length r will occur with a limiting frequency (i.e., fix $\{\bar{\alpha}_i, \bar{\beta}_i\}_{i=1}^{r}$; then the percentage of l, $|l| < n$, such that ${}_x\alpha_{l+i} = \bar{\alpha}_i$ and ${}_x\beta_{l+i} = \bar{\beta}_i$ for $1 \leqq i \leqq r$, will tend to a limit as n tends to ∞). These frequencies will give a measure on finite sequences which extend (as before) to a shift-invariant measure on doubly-infinite sequences. Thus, to a.e. x corresponds a measure-preserving transformation U_x acting on X_x, and a partition $\bar{\mathcal{P}}_x \vee \bar{\mathcal{Q}}_x$ of X_x which partitions the sequences according to their first coordinate. We will show that U_x is ergodic for a.e. x. We will then be finished because, for a.e. x, $\bar{\mathcal{P}}_x$, U_x will be the same as $\mathcal{P}$, T, and $\bar{\mathcal{Q}}_x$, U_x the same as $\mathcal{Q}$, S. For any $\epsilon > 0$, the set of x where $|\mathcal{P}_x - \mathcal{Q}_x| > \gamma + \epsilon$ will have zero measure. We could therefore find a U_x with all the described properties.

LEMMA. *U_x is ergodic for a.e. x.*

PROOF. This follows easily from theorems of Rochlin, but we will sketch a proof for the sake of completeness.

It will simplify notation if we let $\mathcal{R} = \bar{\mathcal{P}} \vee \bar{\mathcal{Q}}$. Then to each x corresponds a sequence which we will call $\{_x\xi_i\}_{-\infty}^{\infty}$. The only properties of U that we will use (or that we have) is that U is measure-preserving and invertible.

As we saw before for a.e. x, $\{_x\xi_i\}$ gives us a measure on finite sequences of length n for all n. We will say that a measure on finite sequences has property A if: given a sequence $\{\gamma_i\}_1^l$ and $\epsilon > 0$, there is an N such that, if $n > N$, we can find a collection C of sequences of length n of measure $> 1 - \epsilon$, and such that the frequency of $\{\gamma_i\}_i^l$ in any two sequences in C differs by less than ϵ. (By frequency of $\{\gamma_i\}_1^l$ in $\{\xi_i\}_1^n$, we mean the fraction of l such that $\xi_{l,i} = \gamma_i$, $1 \leqq i \leqq l$.)

We will now show that, for a.e. x, the measure on finite sequences induced by $_x\xi_i$ satisfies property A. We first note that X (the space on which U acts) breaks up into invariant sets such that, if y and $\bar{y}$ belong to the same set, the frequency of $\{\gamma_i\}_1^l$ in $\{_y\xi_i\}_{-\infty}^{\infty}$ and $\{_{\bar{y}}\xi_i\}_{-\infty}^{\infty}$ differ by less than $1/10\epsilon$. We can now consider these sets separately. Call one of them $\bar{X}$, with the frequency of $\{\gamma_i\}_1^l$ of one of the points in $\bar{X}$ being called f. Applying the ergodic theorem, we can find a set $\bar{X}_1 \subset \bar{X}$ and an N, such that $m(\bar{X}_1) > (1 - \epsilon^2)m(\bar{X})$, and, if $y \in \bar{X}_1$ and $n > N$, the frequency of $\{\gamma_i\}_1^l$ in $\{_y\xi_i\}_1^n$ is within $\frac{1}{2}\epsilon$ of f. One more application of the ergodic theorem gives us that there is a part B of $\bar{X}$ having measure greater than $(1 - \epsilon)m(\bar{X})$, and such that, if $x \in B$, the frequency that $U^i x$ is in $\bar{X}_1$ tends to a limit which is greater than $1 - \epsilon$. Thus, for any x in B, the measures induced by $\{_x\xi_i\}_{-\infty}^{\infty}$ satisfy property A.

We will now show that, if the measure on finite sequences induced by $\{_x\xi_i\}_{-\infty}^{\infty}$ satisfies property A, then U_x will be ergodic. We first note that, if property A holds for each sequence of length l, it holds for any collection C of strings of length l (we look at the frequency of occurrence of strings in C). We will now argue by contradiction. Assume U_x is not ergodic. Then there would be a set E, and $U_x(E) = E$. We could then approximate E by a collection C of sequences of length l (that is, we could find a set E' approximating E, and consisting of those y in X_x such that $\{_y\gamma_i\}_1^l$ is in C, where $\{_y\gamma_i\}_1^l$ is the sequence corresponding to U_x, y and $\mathcal{P}_x \vee \mathcal{Q}_x$; we could have $|E - E'| < \epsilon^2 m(E)$ and $|E - E'| < \epsilon^2 m(X_x - E)$). Then for any n, and for most $(> (1 - \epsilon)m(E))$ of the y in E, $\{_y\gamma_i\}_1^n$ will have a high $(> 1 - \epsilon)$ frequency of C; for most of the y in $(X_x - E)$, $\{_y\gamma_i\}_1^n$ will have a low $(< \epsilon)$ frequency of C.

$d_3 = d_2$. We will now show the equivalence of the second and third definitions of $\bar{d}$. Suppose $\bar{d}_3((\mathcal{P}, T), (\mathcal{Q}, S)) = \alpha$. Then $\bar{d}_2((\mathcal{P}, T), (\mathcal{Q}, S))$ must be $\geqq \alpha - \gamma$ for all $\gamma > 0$. Suppose this were not so for some $\gamma > 0$. Then for fixed l and $\epsilon > 0$, we could find an N such that, if $n > N$, there are two classes C_1^n and C_2^n of sequences of length n, having measure $> 1 - \epsilon$ under $\mathcal{P}$, T and $\mathcal{Q}$, S, respectively, and such that for any sequence in C_1 (or C_2), the l-sequences in it have distribution within $1/10\gamma$ of the distribution of l-sequences under $\mathcal{P}$, T (or $\mathcal{Q}$, S). If (as we assumed) $\bar{d}_2((\mathcal{P}, T), (\mathcal{Q}, S))$ were not $\geqq \alpha - \gamma$, then ϵ could be chosen so that, for infinitely many n, there is a sequence in C_1^n and a sequence in C_2^n which differ in $< (\alpha - \gamma)n$ places. This implies that the distribution of l-sequences in the above two n-sequences is within $\alpha - \gamma$ in the $\bar{d}_3$-metric; hence the distribution of l-sequences under $\mathcal{P}$, T and $\mathcal{Q}$, S differs by $< \alpha - \frac{1}{2}\gamma$. The above would hold for all l, giving a contradiction.

On the other hand, it is clear $\bar{d}_3((\mathcal{P}, T), (\mathcal{Q}, T))$ cannot be $> \alpha + \gamma$ for $\gamma > 0$ because, for each n, $\bar{d}_3(\{T^i\mathcal{P}\}_1^n, \{S^i\mathcal{Q}\}_1^n) \leqq \alpha$, and, by removing two collections of

sequences of measure ϵ (in the z-space), there are still sequences which match to within $(\alpha + \frac{1}{2}\gamma)n$ terms.

APPENDIX 2

THEOREM. *If S_i are Bernoulli shifts (of finite entropy), and*

$$\lim \bar{d}((\mathcal{R}_i, S_i), (\mathcal{R}, S)) = 0,$$

then S is a Bernoulli shift (on $\vee_{-\infty}^{\infty} S^i\mathcal{R}$).

We will give a proof of this theorem. This involves putting together some theorems from [6], and slightly modifying one of the main arguments in [5] or [17].

We will first recall some notation. All partitions are ordered. $\vee_{i=0}^{n} \mathcal{R}_i$ is a partition with a canonical ordering (given the ordering on the $\mathcal{R}_i$). By $d(\mathcal{P})$, we mean the vector $m(\mathcal{P}_1), \cdots, m(\mathcal{P}_k)$, where $\mathcal{P}_i$ are the atoms of $\mathcal{P}$. By $D(\mathcal{P}, \mathcal{Q})$, we mean the sum of the measures of the symmetric differences of $\mathcal{P}_i$ and $\mathcal{Q}_i$. By $d(\mathcal{P}, \mathcal{Q})$, we mean $\sum |m(\mathcal{P}_i) - m(\mathcal{Q}_i)|$ (d is not to be confused with $\bar{d}$). If $\mathcal{P}$ is a partition, the $\mathcal{P} - n$-name of x is the sequence (of length n) whose ith term is the index of the atom of $T^{-i}\mathcal{P}$ containing x. The $\mathcal{P} - n$-name of an atom in $\vee_0^n T^{-i}\mathcal{P}$ is the $\mathcal{P} - n$-name of any one of its points. If $\mathcal{R}_i$ is a sequence of n partitions, we define the $\mathcal{R}$-name of an atom r of $\vee_0^n \mathcal{R}_i$ to be the sequence whose ith term is the index of the atom of $\mathcal{R}_i$ containing r.

We will talk about the distribution of u-names in an n-name. By this, we will mean the following: we take the collection of all u-consecutive terms in the n-name. The percentage of times each particular sequence of length u occurs gives us a measure on sequences of length u, which we call the distribution of u-names in the n-name.

DEFINITION. We will say that a process $(\mathcal{R}, S)$ is *finitely determined*, or f.d., if $\mathcal{R}, S$ has the following property: Suppose that $\mathcal{R}_i, S_i$ is a sequence of ergodic processes converging to $\mathcal{R}, S$ in the vague topology and $E(\mathcal{R}_i, S_i) \to E(\mathcal{R}, S)$. Then $\mathcal{R}_i, S_i$ converges to $\mathcal{R}, S$ in the $\bar{d}$-metric. (We assume that all the partitions $\mathcal{R}$ and $\mathcal{R}_i$ have the same (finite) number of atoms.)

LEMMA 1. *Let $\mathcal{R}, S$ and $\mathcal{P}, T$ satisfy the following: $\mathcal{R}$ and $\mathcal{P}$ have the same number of atoms, T is mixing, $\mathcal{R}, S$ is* f.d., $E(\mathcal{P}, T) < E(\mathcal{R}, S) \leqq E(T)$, *and* (1) $\bar{d}[(\mathcal{R}, S), (\mathcal{P}, T)] < \epsilon^2/100$. *Then, given $\delta > 0$, we can find $\mathcal{P}'$ such that* (1′) $\bar{d}[(\mathcal{R}, S), (\mathcal{P}', T)] < \varphi$, $|\mathcal{P} - \mathcal{P}'| < \epsilon$.

The point of this lemma is the following: We are trying to find a model of $\mathcal{R}, S$ in T. We have already found a model $\mathcal{P}$ which is $\epsilon^2/100$-good (in the $\bar{d}$-metric). We can then change $\mathcal{P}$ to $\mathcal{P}'$ to make it arbitrarily good in the $\bar{d}$-metric.

PROOF OF LEMMA 1. Because $S, \mathcal{R}$ is f.d., it will be enough to show that, given $\bar{\delta}$ and u, we can find $\mathcal{P}'$ satisfying: (1) $d(\vee_0^u T^i\mathcal{P}', \vee_0^u S^i\mathcal{R}) < \bar{\delta}$, and (2) $|E(\mathcal{P}', T) - E(\mathcal{R}, S)| < \bar{\delta}$. (This is the only place in which we use the fact that $(S, \mathcal{R})$ is f.d.)

Choose a refinement $\mathcal{Q}$ of $\mathcal{P}$ such that $E(\mathcal{R}, S) - E(\mathcal{Q}, \tau) = \beta > 0$ where $\beta < \bar{\delta}/100$.

Choose $\gamma < \min(\bar{\delta}, \epsilon)$ such that $D(\mathcal{Q}', \mathcal{Q}) < \gamma$ implies that $E(\mathcal{Q}', T) \geqq E(\mathcal{Q}, T) - \bar{\delta}/100$.

We will call an atom in $\vee_0^n T^{-i}\mathcal{Q}$ good (or good $\mathcal{Q}_n$-atoms) if its measure is between

$(\frac{1}{2})^{[E(\mathcal{Q},T)\pm\beta/100]n}$. We will call an atom in $\bigvee_0^n S^{-i}\mathcal{R}$ good if its measure is between $(\frac{1}{2})^{[E(\mathcal{R},S)\pm\beta/100]n}$, and if its n-name has the property that the distribution of u-names in it is within $\bar{\delta}$ of $d(\bigvee_0^u S^{-i}\mathcal{R})$.

Because of the Shannon-McMillan-Breiman theorem and the ergodic theorem, we can choose n so large that the measure of the union of atoms in $\bigvee_0^n T^{-i}\mathcal{Q}$ and $\bigvee_0^n S^{-i}\mathcal{R}$ which are not good is less than $\gamma/100$. We can also choose n so large that $u_n < \bar{\delta}/100$, $n\beta > 100$, and $m(A) < 1/n$ implies that $-[m(A)\log m(A) + (1 - m(A))\cdot \log(1 - m(A))] < \delta/100$.

Because of hypothesis (1), we can choose partitions $\mathcal{R}_i$, $0 \leqq i \leqq n$, such that $d(\bigvee_0^n \mathcal{R}_i) = d(\bigvee_0^n S^{-i}\mathcal{R})$, and $(n+1)^{-1}\sum_{i=0}^n D(\mathcal{R}_i, T^{-i}\mathcal{P}) < \epsilon^2/100$. The good atoms in $\bigvee_0^n \mathcal{R}_i$ will be those which correspond to the good atoms in $\bigvee_0^n S^{-i}\mathcal{R}$. We will call these good $\mathcal{R}_n$-atoms. The $\mathcal{R} - n$-name of a point will be defined with respect to $\bigvee_0^n \mathcal{R}_i$.

We will now pick out a subset of the good $\mathcal{Q}_n$-atoms, which we will call very good $\mathcal{Q}_n$-atoms. They are defined as follows: We take those good $\mathcal{Q}_n$-atoms g, which are more than half covered by good $\mathcal{R}_n$-atoms, whose n-name differs from the n-name of the atom in $\bigvee_0^n T^{-i}\mathcal{P}$ containing g, in less than $\epsilon/2$ n-places. Because $(n+1)^{-1}\sum_{i=0}^n D(\mathcal{R}_i, T^{-i}\mathcal{P}) < \epsilon^2/100$, we get that the measure of the union of very good $\mathcal{Q}_n$-atoms is $> 1 - \epsilon/2$. Let E be the set of points whose $\mathcal{P} - n$-name differs from its $\mathcal{R} - n$-name in less than $\epsilon/2$ n-places. (More than $\frac{1}{2}$ of a very good $\mathcal{Q}_n$-atom lies in E.)

Because the measure of a good $\mathcal{Q}_n$-atom is greater than two times the measure of a good $\mathcal{R}_n$-atom, we have that any l very good $\mathcal{Q}_n$-atoms intersect at least l good $\mathcal{R}_n$-atoms in E. We can therefore apply the marriage lemma to assign to each very good $\mathcal{Q}_n$-atom a good $\mathcal{R}_n$-atom which intersects it in E. Thus we have assigned to each very good $\mathcal{Q}_n$-atom g a good $\mathcal{R}_n$-atom, whose $\mathcal{R} - n$-name agrees with the $\mathcal{P} - n$-name of the atom of $\bigvee_0^n T^{-i}\mathcal{P}$, containing g in more than $(1 - \epsilon/2)$ n-places.

Because there are more good $\mathcal{R}_n$-atoms than there are good $\mathcal{Q}_n$-atoms, we can extend the above assignment so that each good $\mathcal{Q}_n$-atom is assigned a good $\mathcal{R}_n$-atom.

We can now apply Rochlin's theorem to obtain a set F'' such that T^iF'', $0 \leqq i \leqq n$, are disjoint, and $m(\bigcup_0^n T^iF'') > 1 - \gamma/100$. Let $F' = T^kF''$. Since T is mixing, we can assume that $|d(\bigvee_0^n T^{-i}\mathcal{Q}/F') - d(\bigvee_0^n T^{-i}\mathcal{Q})|$ is so small that, by removing a set from F' of measure $< (\gamma/100)m(F')$ and calling what is left F, we get $d(\bigvee_0^n T^{-i}\mathcal{Q}/F) = d(\bigvee_0^n T^{-i}\mathcal{Q})$. Call those atoms in $\bigvee_0^n T^{-i}\mathcal{Q}/F$, which are contained in good and very good $\mathcal{Q}_n$-atoms, good and very good $\mathcal{Q}_n - F$-atoms, and carry over our assignment of good $\mathcal{R}_n$-atoms to $\mathcal{Q}_n - F$-atoms.

We will now define $\mathcal{P}'$ on $\bigcup_0^n T^i\bar{F}$ is the union of the good $\mathcal{Q}_n - F$-atoms. If g is a good $\mathcal{Q} - F$-atom, then T^jg will lie in $\mathcal{P}_i'$, where i is the jth term in the $\mathcal{R} - n$-name of the $\mathcal{R}_n$-atom assigned to g. It does not matter how we define $\mathcal{P}'$ on the rest of X.

Because each $\mathcal{Q}_n - F$-atom in $\bar{F}$ was assigned to a different $\mathcal{R}_n$-atom, we have that $\bigvee_{-n}^n T^i(\mathcal{P}' \vee \mathcal{F})$[6] restricted to $\bigcup_0^n T^i\bar{F}$ refines $\mathcal{Q}$. Because the measure of the union of good $\mathcal{Q}_n$-atoms is $> 1 - \gamma/100$, and because of our choice of γ, we get that $E(\mathcal{P}', T) > E(\mathcal{R}, S) - \bar{\delta}$.

Because the name of the good $\mathcal{R}_n$-atoms have the property that the u-names in them are distributed well, we get $d(\bigvee_0^u T^i\mathcal{P}', \bigvee_0^u S^i\mathcal{R}) < \bar{\delta}$. Lastly, since each very

[6] $\mathcal{F}$ is the partition consisting of $\bar{F}$ and its complement.

good $\mathcal{Q}_n$-atom has the property that the $\mathcal{P}$ — n-name of the atoms in $\bigvee_0^n T^{-i}\mathcal{P}$ containing it, and the $\mathcal{R}$ — n-name of the $\mathcal{R}_n$-atom assigned to it, differ in less than $\epsilon/2$ n-places, and since the measure of the union of the very good $\mathcal{Q}_n$-atoms is $> 1 - \epsilon/2$, we have $\bar{d}(\mathcal{P}', \mathcal{P}) < \epsilon$.

LEMMA 2. *If $\mathcal{R}_i S_i \to \mathcal{R}$, S (in the $\bar{d}$-metric), and if the $\mathcal{R}_i$, S_i are Bernoulli shifts, and $\mathcal{R}_i$ are arbitrary partitions which generate, and $E(\mathcal{R}_i, S_i) < E(\mathcal{R}_{i+1}, S_{i+1}) < E(\mathcal{R}, S)$, then S, acting on $\bigvee_{-\infty}^{\infty}$, is a Bernoulli shift.*

PROOF OF LEMMA 2. Let $\gamma_i = \bar{d}[(\mathcal{R}_i, S_i), (\mathcal{R}_{i+1}, S_{i+1})]$. We can assume that $\sum_{i=1}^{\infty} (\gamma_i)^{\frac{1}{2}} < \infty$. Let T be a Bernoulli shift such that $E(T) = E(\mathcal{R}, S)$. By Sinai's theorem, we can assume that there is a partition $\mathcal{P}_1$ such that $\bar{d}((\mathcal{P}_1, T), (\mathcal{R}_1, S_1)) = 0$.

We will prove Lemma 2 by constructing a sequence $\mathcal{P}_i$ such that $\sum_{i=1}^{\infty} D(\mathcal{P}_i, \mathcal{P}_{i+1}) < \infty$ and $\bar{d}[(\mathcal{P}_i, T), (\mathcal{R}_i, S_i)] < \xi_i$, where $\xi_i < \gamma_i$, and ξ_i is so small that it forces $E(\mathcal{P}_i, T) < E(\mathcal{R}_{i+1}, S_{i+1})$. We will then be finished because $\mathcal{P}_i \to \mathcal{P}$ and $\mathcal{P}$, T will be equivalent to $\mathcal{R}$, S. We will then use the theorem on factors of Bernoulli shifts to get Lemma 2.

We will use Lemma 1 of this paper to get the above sequence as follows: First note that, because of Lemma 3′ of [6], the S_i, $\mathcal{R}_j$ are f.d. Suppose we already have $\mathcal{P}_i$, T. Applying Lemma 1 to $\mathcal{P}_i$, T and $\mathcal{R}_{i+1}$, S_{i+1}, and noting that $\bar{d}[(\mathcal{P}_i, T), (\mathcal{R}_{i+1}, S_{i+1})] < \gamma_i + \xi_i$, we get $\mathcal{P}_{i+1}$ such that $D(\mathcal{P}_{i+1}, \mathcal{P}_i) < 200\,(\gamma_i)^{\frac{1}{2}}$ and $\bar{d}[(\mathcal{P}_{i+1}, T), (\mathcal{R}_{i+1}, S_{i+1})] < \xi_{i+1}$.

We get Theorem 1 from Lemma 2 as follows: Choose Bernoulli shifts $\bar{S}_i$ and $\bar{S}$ with independent generators $\mathcal{B}_i$ and $\mathcal{B}$, such that $E(S_i \times \bar{S}_i) \leqq E(S_{i+1} \times \bar{S}_{i+1}) < \cdots < E(S \times \bar{S})$ and $\lim_{i \to \alpha} d(\mathcal{B}_i, \mathcal{B}) = 0$ have that $\bar{d}((\mathcal{B}_i \times \mathcal{R}_i, S_i \times \bar{S}_i), (\mathcal{B} \times \mathcal{R}, X \times \bar{S}))_1 \to 0$. Applying the above lemma, we get that $S \times \bar{S}$, acting on $\bigvee_{-\infty}^{\infty} (S \times S_1)^i (\mathcal{B} \times \mathcal{R})$, is a Bernoulli shift. Because factors of Bernoulli shifts are Bernoulli shifts, we have that $S \times \bar{S}$, acting on $\bigvee_{-\infty}^{\infty} (S \times \bar{S})^i (\mathcal{R} \times X)$, is a Bernoulli shift, but this is isomorphic to S acting on $\bigvee_{-\infty}^{\infty} S^i \mathcal{R}$.

APPENDIX 3

Any stationary process can be approximated arbitrarily well by n-step Markov processes in the vague topology; that is, given a process $\mathcal{P}, T$ and an n, we can find an n-step Markov process $\mathcal{P}_n$, T_n which gives the same measure to sequences of length n as does $\mathcal{P}$, T. We will now give an informal description of $\mathcal{P}_n$, T_n (it will be obvious how to make it rigorous). $\mathcal{P}_n$, T_n works as follows: We assume that the distribution of the last $(n-1)$-letters printed out is the same as the distribution of $(n-1)$-sequences under $\mathcal{P}$, T. The probability of the next letter will be such that the distribution of the last $(n-1)$-letters, together with the letter being printed, is the same as the distribution of n-sequences under $\mathcal{P}$, T. (It is obvious that this can be done; merely copy $\mathcal{P}$, T.) This can be continued because, to print out a letter the next time, we again look at the last $(n-1)$-letters printed out. Their distribution is the same as the distribution of the last $(n-1)$-letters in the previous step (because $\mathcal{P}$, T was stationary). It is thus obvious that this defines a stationary process which is an n-step Markov process, and which gives n-sequences the same measure as does $\mathcal{P}$, T.

We should point out that we have given a canonical description of $\mathcal{P}_n$, T_n. $\mathcal{P}_n$, T_n can be thought of roughly as the process $\mathcal{P}$, T modified by erasing its memory of what it did more than n-steps in the past.

Furthermore, it is obvious that $E(\mathcal{P}_n, T_n) \geqq E(\mathcal{P}, T)$.

If T is mixing, then so is T_n. We see this as follows: It is easy to see (and is well known) that T_n will be mixing if there is a K such that: let $\{\alpha_i\}_1^n$ and $\{\beta_i\}_1^n$ be two n-sequences which have probability >0 under $\mathcal{P}_n, T_n$. Then the probability that $\gamma_{K+i} = \beta_i$, $1 \leqq i \leqq n$, given that $\gamma_i = \alpha_i$, $1 \leqq i \leqq n$, is >0. The above statement is true for the process $\mathcal{P}, T$ (instead of $\mathcal{P}_n, T_n$) because T is mixing. This will imply that it holds for $\mathcal{P}_n, T_n$, because any finite sequence which has non-zero probability under $\mathcal{P}, T$ also has non-zero probability under $\mathcal{P}_n, T_n$. (This is obvious by induction.)

APPENDIX 4

In this appendix, I will describe another interpretation for a process which suggests a line of future research.

The following model was suggested to me by D. W. Robinson. Suppose we have an infinite number of molecules positioned at the integer points on the line, and that each molecule has two possible spins. Thus each configuration of the system can be described by a sequence of 0 and 1 (or -1 and $+1$). A "state" of the system will correspond to a probability measure on all configurations or all sequences of 0, 1. It will be assumed that the measure is translation-invariant. Thus the mathematical model for a "state" of the system is the same as the mathematical model for a process. (If each molecule is unaffected by the other molecules, then the state would correspond to an independent process.) It now turns out that the Kolmogorov entropy of the process is the same as the Gibbs' entropy of the "physical" system ([15]). Any classification of stationary processes gives a classification of states of these systems.

For each state of the system, we have a potential energy-per-unit volume, which depends on the forces of interaction between the molecules. (For each configuration of a finite number of molecules, we have a potential energy. For each configuration, we take the limit of the potential energy due to the molecules between $-N$ and N, and divide by $2N$, taking the limit as N tends to ∞. For a given state, a.e. configurations will have the same potential energy-per-unit volume.) By physical principles, we would expect the equilibrium state to be the state which maximizes the entropy for a fixed potential energy. Furthermore, we would expect that, as a state tends to equilibrium, its entropy increases to the entropy of the equilibrium state, and that we get convergence in the vague topology. (Thus if the equilibrium state is a B-state (or B-process), as is the case for finite range interaction, we could conclude that convergence to equilibrium must take place in the $\bar{d}$-metric.)

The above models become physically interesting if we replace the integer points on the line by the integer points in 3-space. To study the states of such a system, we would have to consider three commuting transformations (translations in the x-, y-, and z-directions), and a partition $\mathcal{P}$. B. Weiss and M. Smorodinsky have already suggested to me that a theory for three commuting transformations can be developed along the lines of the theory for one transformation.

REFERENCES

[1] ANOSOV, D. V. (1967). Geodessic flows on closed Riemann manifolds of negative curvature. *Trudy Mat. Inst. Akad. Nauk SSSR* **90**.

[2] ANOSOV, D. V. and SINAI, Y. G. (1967). Certain smooth ergodic systems. *Uspehi Mat. Nauk* **22** 107–172.

[3] FRIEDMAN, N. A. and ORNSTEIN, D. S. (1970). On isomorphism of weak Bernoulli transformations. *Advances in Math.* **5** 365–395.

[4] KOLMOGOROV, A. N. (1957). Théorie générale des systèms dynamiques et mécanique classique. *Proc. Internat. Congress Math. (Amsterdam)* **1**. North-Holland, Amsterdam. 315–333. MR 20 No. 4066.

[5] ORNSTEIN, D. S. (1970). Bernoulli shifts with the same entropy are isomorphic. *Advances in Math.* **4** 337–352.

[6] ORNSTEIN, D. S. (1970). Factors of Bernoulli shifts are Bernoulli shifts. *Advances in Math.* **5** 349–364.

[7] ORNSTEIN, D. S. (1970). Imbedding Bernoulli shifts in flows. *Lecture Notes in Mathematics.* Springer-Verlag, Berlin. (Proc. of Conf. on Ergodic Theory, Ohio State Univ., March 1970) 178–218.

[8] ORNSTEIN, D. S. A Kolmogorov automorphism that is not a Bernoulli shift. To appear in *Advances in Math.*

[9] ORNSTEIN, D. S. A K-automorphism with no square root and Pinsker's conjecture. To appear.

[10] ORNSTEIN, D. S. The isomorphism theorem for Bernoulli flows. To appear.

[11] ORNSTEIN, D. S. AND SHIELDS, P. C. An uncountable family of K-automorphisms. To appear.

[12] ORNSTEIN, D. S. AND SHIELDS, P. C. Mixing Markov shifts of kernel type are Bernoulli. To appear.

[13] ROCHLIN, V. A. (1964). Metric properties of endomorphisms of compact commutative groups. *Izv. Akad. Nauk SSSR Mat.* **28** 867–874. (English translation in *Trans. Amer. Math. Soc.* (1967) 244–252.) MR 29, No. 5955.

[14] ROCHLIN, V. A. (1969). Lectures on the entropy theory of transformations with invariant measure. *Russian Math. Surveys* **22** 1–52. MR 36 No. 349.

[15] RUELLE, D. (1969). *Statistical Mechanics Rigorous Results.* Benjamin, New York.

[16] SMORODINSKY, M. (1971a). Ergodic theory, entropy, entropy. *Lecture Notes in Mathematics.* Springer-Verlag, Berlin.

[17] SMORODINSKY, M. (1971b). An exposition of Ornstein's isomorphism theorem. To appear.

[18] ORNSTEIN, D. S. and WEISS, B. Geodesic Flows are Bernoullian. To appear, *Israel Journal of Mathematics.*

DEPARTMENT OF MATHEMATICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305