Ergodicity of Markov Channels

ROBERT M. GRAY, FELLOW, IEEE, MARI O. DUNHAM, MEMBER, IEEE, AND R. L. GOBBI

Abstract—**A Markov channel is a discrete information channel that includes as special cases the finite state channels and finite state codes of information theory. Kieffer and Rahe proved that one-sided and two-sided Markov channels have the following property: If the input source to a Markov channel is asymptotically mean stationary (AMS), then so is the resulting input–output process and hence the ergodic theorem and the Shannon–McMillan–Breiman theorem hold for the input–output process. Kieffer and Rahe also provided a sufficient condition for any AMS ergodic source to yield an AMS ergodic input–output process. New conditions for a Markov channel to have this ergodicity property are presented and discussed here. Several relations are developed among various classes of channels, including weakly ergodic, indecomposable, and strongly mixing channels. Some connections between Markov channels and the theory of nonhomogeneous Markov chains are also discussed.**

Introduction

GIVEN AN INFORMATION SOURCE (a discrete-time random process) and a noisy channel (essentially a regular conditional probability measure describing a probability measure on output sequences given an input sequence), information about the source can be communicated to a receiver by first encoding the source sequence into a channel input sequence and decoding the channel output sequence into a reproduction sequence observed by the receiver. Assume that we have some measure of the quality of the reproduction sequence, that is, how well it approximates the original source sequence. The coding theorems of information theory quantify the theoretically optimum performance that can be achieved using the given source and given channel with any encoder and decoder within some constrained class, where "optimum" means that the system has the minimum possible average distortion. The design algorithms of information theory are methods for actually designing codes that work well, ideally not too badly in comparison with the theoretical optimum. The proofs of coding theorems rest primarily on the ergodic theorem and on the Shannon–McMillan–Breiman theorem. They also generally require that the appropriate sample averages converge to constants and hence that the underlying system be ergodic. (The proofs for nonergodic systems generally combine the ergodic proof with the ergodic decomposition). In addition, proofs of the convergence of some code design algorithms based on long training sequences of actual data also rest on the ergodic theorem. Hence it is of interest to know when communication system models satisfy the conditions for an ergodic theorem and the Shannon–McMillan–Breiman theorem.

Stationarity and block stationarity (stationarity of successive blocks of fixed size) have long been known to be a sufficient condition for these results, but the finite state channels and the finite state codes introduced by Shannon [14] do not generally meet these conditions. For example, if the process begins at time 0 in a particular state, then the channel or code may exhibit initial transients and hence not be stationary. Partially as a result, coding theorems for finite state channels have proved difficult, relying on the special properties of Markov chains, and the stationarity and ergodicity properties of finite state codes have been little developed. Gray and Kieffer [6] showed that a necessary and sufficient condition for a process to have an ergodic theorem is that it be asymptotically mean stationary (AMS) and that an AMS process satisfies the Shannon–McMillan–Breiman theorem. A channel is said to be AMS if connecting an AMS source or input process to the channel results in an AMS input–output process. Kieffer and Rahe [9] introduced a generalization of both finite state channels and finite state codes called a Markov channel and showed that such channels are AMS. An AMS channel is said to be ergodic if connecting any AMS ergodic source to the channel yields an AMS ergodic input–output process. Kieffer and Rahe showed that a sufficient condition for a Markov channel to be ergodic is that it be indecomposable in a sense similar to that of Blackwell, Breiman, and Thomasian [2]. Unfortunately, however, this condition is too strong for some applications. For example, in design studies of finite state codes many examples have been found that are not indecomposable, yet they appear to yield ergodic processes.

In this paper we develop and compare several sufficient conditions for Markov channels to be ergodic. The principal result focuses on channels whose output forms a weakly ergodic nonhomogeneous Markov chain as in Hajnal [7]. A superficially similar result was recently obtained for a very different application—proving exponential convergence of adaptive algorithms—by Shi and Kozin [15] using results of Furstenberg and Kesten [4] on products of random sequences of matrices. Additional results focus on the relations among weakly ergodic channels and

Manuscript received April 17, 1986; revised December 8, 1986. This paper was presented in part at the IEEE International Symposium on Information Theory, Brighton, England, June 23–28, 1985. This research was partially supported by the National Science Foundation under Grant ECS83-17981.

R. M. Gray is with the Information Systems Laboratory, Stanford University, Stanford, CA 94305, USA.

M. O. Dunham is with the Department of Computer Science and Electrical Engineering, Boston University, Boston, MA 02215, USA.

R. L. Gobbi is with Lockheed Missiles and Space Co., Sunnyvale, CA 94806, USA.

IEEE Log Number 8613567.

various types of indecomposable channels. Most notably, we show that weakly ergodic Markov channels in the sense of Hajnal are equivalent to strongly mixing channels in the sense of Adler [1].

The results developed here are easy consequences of known results for products of stochastic matrices. This paper provides, however, the first collection and comparison of the numerous ergodicity conditions for Markov channels and for products of stochastic matrices and it provides simple and direct proofs of the equivalence of many of the conditions.

Preliminaries

For convenience we mostly follow the notation of Kieffer and Rahe [9] for the basic definitions. Let $(\Omega, \mathbf{F})$ be a measurable space and T: $\Omega \to \Omega$ be a measurable but not necessarily invertible transformation. Following Gray and Kieffer [6], we say a probability measure μ on $(\Omega, \mathbf{F})$ is *asymptotically mean stationary* (AMS) with respect to T if $\lim_{n \to \infty} n^{-1} \Sigma_{i=0}^{n-1} \mu(T^{-i}F)$ exists for all $F \in \mathbf{F}$. In the special case where $\mu(T^{-1}F) = \mu(F)$; for all $F \in \mathbf{F}$, the measure μ is said to be *stationary* with respect to T. μ is *ergodic* with respect to T if $\mu(F)$ is 0 or 1 for all invariant $F \in \mathbf{F}$. We will omit the modifying phrase "with respect to T" when the transformation is clear from the context.

We say that a dynamical system $(\Omega, \mathbf{F}, \mu, T)$ is AMS, stationary, or ergodic if the measure μ on $(\Omega, \mathbf{F})$ is. The most important properties of AMS measures and systems are summarized below for reference. Proofs may be found in Gray and Kieffer [6].

Properties of AMS Measures

Fix a dynamical system $(\Omega, \mathbf{F}, \mu, T)$.

1) The dynamical system is AMS if and only if there exists a probability measure $\bar{\mu}$ on $(\Omega, \mathbf{F})$ that is stationary with respect to T such that $\bar{\mu}(F) = \mu(F)$ for every invariant set $F \in \mathbf{F}$.
2) The dynamical system is AMS if there exists a probability measure $\bar{\mu}$ on $(\Omega, \mathbf{F})$ that is stationary with respect to T such that if $F \in \mathbf{F}$ is invariant and $\bar{\mu}(F) = 0$, then also $\mu(F) = 0$.
3) If T is invertible (as is the two-sided shift, for example), the dynamical system is AMS if and only if there exists a probability measure $\bar{\mu}$ on $(\Omega, \mathbf{F})$, stationary with respect to T, such that μ is absolutely continuous with respect to $\bar{\mu}$.
4) We say that the dynamical system $(\Omega, \mathbf{F}, \mu, T)$ has an (individual) ergodic theorem if for any bounded measurable function f: $\Omega \to (-\infty, \infty), \lim_{n \to \infty} n^{-1} \Sigma_{i=0}^{n-1} f T^i$ exists almost everywhere (a.e.) $[\mu]$. A dynamical system $(\Omega, \mathbf{F}, \mu, T)$ has an ergodic theorem if and only if it is AMS.
5) The Shannon–McMillan–Breiman theorem holds for AMS systems.

The fourth and fifth properties justify the importance of asymptotic mean stationarity: it is a necessary and sufficient condition for the ergodic theorem to hold and a sufficient condition for the Shannon–McMillan–Breiman theorem.

In information theory, the most important dynamical systems are sources and source–channel hookups. These systems and cascades of such systems model communication systems consisting of information sources, coders, and noisy channels. Let $(A, \mathbf{A})$ be a measurable space which can be thought of as modeling the output of a random process at one particular time. We wish to consider sequences of such outputs, but we will need to consider both one-sided and two-sided sequences. We consider the one-sided case first. Let $(A_1^\infty, \mathbf{A}_1^\infty)$ be the measurable space of one-sided sequences from A, that is, the set of all sequences $(x_1, x_2, \cdots)$ from A and $\mathbf{A}_1^\infty$ is the usual product σ-field of subsets of A_1^∞. Let T denote the left shift on A_1^∞, that is, the measurable map T: $A_1^\infty \to A_1^\infty$ defined by $T(x_1, x_2, \cdots) = (x_2, x_3 \cdots)$. A dynamical system $(A_1^\infty, \mathbf{A}_1^\infty, \mu, T)$ of this form is called a one-sided source or process and is abbreviated to $[A, \mu]$. A is called the alphabet of the source. We also use the words source or process to refer to the discrete-time random process defined by the sequence of coordinate random variables on the product space, e.g., $\{X_n; n = 1, 2, \cdots\}$ defined by the maps X_n: $A_1^\infty \to A$ where $X_n(x) = x_n$ if $x = (x_1, x_2, \cdots)$.

We define a two-sided source in a similar manner: Let A^∞ denote the set of all two-sided sequences drawn from A, that is, all sequences of the form $x = \{x_i\}_{-\infty}^{\infty}$, and let $\mathbf{A}^\infty$ denote the corresponding product σ-field. Let T again denote the shift, now defined by $(Tx)_i = x_{i+1}$, all i. We use the same notation for the shift on different spaces; context should make clear what the underlying space is. A one-sided or two-sided source $[A, \mu]$ is AMS, stationary, or ergodic if the corresponding dynamical system $(A_1^\infty, \mathbf{A}_1^\infty, \mu, T)$ (one-sided) or $(A^\infty, \mathbf{A}^\infty, \mu, T)$ (two-sided) is.

Rather than continue separate development for the one-sided and two-sided cases, we introduce notation which allows us to treat the two simultaneously when convenient. Given a measurable space $(A, \mathbf{A})$, let $(\Sigma_A, \mathbf{\Sigma}_A)$ denote the corresponding one-sided or two-sided sequence space, that is, either $(A_1^\infty, \mathbf{A}_1^\infty)$ or $(A^\infty, \mathbf{A}^\infty)$, as appropriate. Let $\mathbf{I}$ denote the time index set; $\mathbf{I}$ is $\{1, 2, 3, \cdots\}$ for the one-sided case and all integers for the two-sided case.

A *channel* is a triple $[A, \nu, B]$ with input alphabet A and output alphabet B and a family of probability measures $\{\nu_x; x \in \Sigma_A\}$ on $(\Sigma_B, \mathbf{\Sigma}_B)$ such that for each $F \in \mathbf{\Sigma}_B$ the map $x \to \nu_x(F)$ is a measurable map from $(\Sigma_A, \mathbf{\Sigma}_A)$ into $[0,1]$ and its Borel field. A channel is called one-sided or two-sided if the sequence spaces are one-sided or two-sided, respectively. Given a source $[A, \mu]$ and a channel $[A, \nu, B]$, then the source–channel hookup or input–output process $\mu\nu$ is the process $[A \times B, \mu\nu]$ where the measure $\mu\nu$ is defined by

$$\mu\nu(F) = \int_{\Sigma_A} \nu_x(F_x)\, d\mu(x), \qquad F \in \mathbf{\Sigma}_{A \times B},$$

where $F_x = \{y \in B_1^\infty : (x, y) \in F\}$. We also denote by T the shift on $\Sigma_A \times \Sigma_B$, that is, $T(x, y) = (Tx, Ty)$. Define for all integers $n \in I$ the following coordinate random variables:

$$X_n: \Sigma_A \to A \text{ by } X_n(x) = x_n,$$
$$Y_n: \Sigma_B \to B \text{ by } Y_n(y) = y_n,$$
$$(X_n, Y_n)(x, y) = (X_n(x), Y_n(y)) = (x_n, y_n);$$

then we also use the words input–output process to mean the random process $\{(X_n, Y_n); n = 1, 2, \cdots\}$ in the one-sided case and $\{(X_n, Y_n); n = \cdots, -1, 0, 1, 2, \cdots\}$ in the two-sided case. As with any source, an input–output process is stationary, AMS, or ergodic if the corresponding dynamical system $(\Sigma_{A \times B}, \boldsymbol{\Sigma}_{A \times B}, \mu\nu, T)$ is.

A channel $[A, \nu, B]$ is said to be *stationary* if

$$\nu_{Tx}(F) = \nu_x(T^{-1}F); \qquad \text{all } x \in \Sigma_A, \text{ all } F \in \boldsymbol{\Sigma}_B.$$

The channel is said to be AMS if for every AMS source $[A, \mu]$, the input–output process is AMS. An AMS channel is said to be *ergodic* if $\mu\nu$ is ergodic whenever μ is AMS and ergodic. Note that an AMS channel ν may yield an ergodic $\mu\nu$ for some AMS ergodic sources and not for others. This raises an issue that often occurs when studying channels. In some applications it suffices to know that given a channel ν and a *specific* AMS ergodic input source μ, the resulting pair process $\mu\nu$ is ergodic. In other cases one requires the stronger result that $\mu\nu$ will be ergodic for *all* such input sources. For example, information theoretic quantities such as the channel capacity can be defined as a maximization or minimization over all AMS ergodic sources and the proofs of coding theorems require ergodicity of the input–output process. We will consider both types of results here and this will require different types of properties of channels: those which hold almost everywhere for a particular source, those which hold almost everywhere for all sources within some given class (such as all stationary sources), and those which hold everywhere.

Markov Channels

The most general known class of AMS channels are the Markov channels introduced by Kieffer and Rahe [9]. For a fixed positive integer K, let $\boldsymbol{P}$ denote the space of all $K \times K$ stochastic matrices $P = \{P(i, j); i, j = 1, 2, \cdots, K\}$. Using the Euclidean metric on this space we can construct the Borel field $\mathscr{P}$ of subsets of $\boldsymbol{P}$ generated by the open sets to form a measurable space $(\boldsymbol{P}, \mathscr{P})$. This, in turn, gives a one-sided or two-sided sequence space $(\Sigma_P, \boldsymbol{\Sigma}_P)$.

A map $\phi: \Sigma_A \to \Sigma_P$ is said to be *stationary* if $\phi T = T\phi$. Given a sequence $P \in \Sigma_P$, let $\boldsymbol{M}(P)$ denote the set of all probability measures on $(\Sigma_B, \boldsymbol{\Sigma}_B)$ with respect to which $Y_m, Y_{m+1}, Y_{m+2}, \cdots$ forms a Markov chain with transition matrices $P_m, P_{m+1}, \cdots$ for any integer m, that is, $\lambda \in \boldsymbol{M}(P)$ if and only if for any m,

$$\lambda[Y_m = y_m, \cdots, Y_n = y_n] = \lambda[Y_m = y_m] \prod_{i=m}^{n-1} P_i(y_i, y_{i+1}),$$
$$n > m, y_m, \cdots, y_n \in B.$$

In the one-sided case only $m = 1$ need be verified. Observe that in general the Markov chain is nonhomogeneous.

Define a channel $[A, \nu, B]$ to be *Markov* if there exists a stationary measurable map $\phi: \Sigma_A \to \Sigma_P$ such that $\nu_x \in \boldsymbol{M}(\phi(x))$, $x \in \Sigma_A$. Kieffer and Rahe [9] proved that one-sided and two-sided Markov channels are AMS.

An important example is given by finite state channels and codes. Given a Markov channel with stationary mapping ϕ, the channel is said to be a *finite state channel* (FSC) if we have a collection of stochastic matrices $P_a \in \boldsymbol{P}$; $a \in A$ and that $\phi(x)_n = P_{x_n}$, that is, the matrix produced by ϕ at time n depends only on the input at that time, x_n. If the matrices P_a; $a \in A$ contain only 0's and 1's, the channel is called a *finite state code*. There are several equivalent models of finite state channels and we pause to consider an alternative form that is more common in information theory. (See Gallager [5, Ch. 4] for a discussion of equivalent models of FSC's and numerous physical examples.) An FSC converts an input sequence x into an output sequence y and a state sequence s according to a conditional probability

$$\Pr(Y_k = y_k, S_k = s_k; k = m, \cdots, n | X_i = x_i, S_i = s_i; i < m)$$
$$= \prod_{i=m}^{n} P(y_i, s_i | x_i, s_{i-1});$$

that is, conditioned on X_i, S_{i-1}, the pair Y_i, S_i is independent of all prior inputs, outputs, and states. This specifies an FSC defined as a special case of a Markov channel where the output sequence above is here the joint state-output sequence $\{y_i, s_i\}$. Note that with this setup, saying the Markov channel is AMS implies that the triple process of source, states, and outputs is AMS (and hence obviously so is the Gallager input–output process). We will adopt the Kieffer–Rahe viewpoint and call the outputs $\{Y_n\}$ of the Markov channel "states" even though they may correspond to state-output pairs for a specific physical model. We do not here treat the issue of when the process $\{X_n, Y_n\}$ in the Gallager FSC model might be AMS and ergodic without the process $\{X_n, S_n, Y_n\}$ sharing the property.

In the two-sided case, the Markov channel is significantly more general than the FSC because the choice of matrices $\phi(x)_i$ can depend on the past in a very complicated (but stationary) way. One might think that a Markov channel is not a significant generalization of an FSC in the one-sided case, however, because there stationarity of ϕ does not permit a dependence on past channel inputs, only on future inputs, which might seem physically unrealistic. Many practical communications systems do effectively depend on the future, however, by incorporating delay in the coding. The prime example of such "look-ahead" coders are trellis and tree codes used in an incremental fashion. Such codes investigate many possible output strings several steps into the future to determine the possible effect on the receiver and select the best path, often by a Viterbi algorithm. (See, e.g., Viterbi and Omura [16].) The encoder then outputs only the first symbol of the selected path. While clearly a finite state machine, this code does not fit the usual model of a finite state channel or code because of the dependence of the transition matrix on future inputs (unless, of course, one greatly expands the state space). It is, however, a Markov channel.

While the proof that Markov channels are AMS is difficult, especially for the one-sided case, we will need a few properties of the construction which we now summarize.

Let $[A, \mu]$ be an AMS source and $[A, \nu, B]$ a Markov channel. Let $\phi\colon \Sigma_A \to \Sigma_P$ be the stationary map such that $\nu_x \in \boldsymbol{M}(\phi(x))$, $x \in \Sigma_A$. Since $[A, \mu]$ is AMS, there is a stationary measure $\bar{\mu}$ such that $\bar{\mu}(F) = 0$ for an invariant $F \in \Sigma_A$ implies that $\mu(F) = 0$ also. For both the one-sided and two-sided cases the key construction takes place on a two-sided process. Consider a two-sided source $[A, \bar{\mu}^*]$ defined as follows: if the original source is two-sided, then $\bar{\mu}^* = \bar{\mu}$; if the original source is one-sided, then let $\bar{\mu}^*$ be the two-sided extension of the one-sided measure $\bar{\mu}$, that is,

$$\bar{\mu}^*((X_m, X_{m+1}, \cdots) \in F) = \bar{\mu}(F), \qquad F \in A_1^\infty, \quad (1)$$

which specifies $\bar{\mu}^*$. We also define a two-sided stationary map $\phi'\colon A^\infty \to \boldsymbol{P}^\infty$ by setting $\phi' = \phi$ if the original system is two-sided and defining $\phi'(x)_i = (\phi((x_i, x_{i+1}, \cdots)))_1$ if the original system is one-sided.

Kieffer and Rahe construct a two-sided channel $[A, \hat{\nu}, B]$ with the following properties.

a) The channel is stationary and hence so is the input–output process $\bar{\mu}^*\hat{\nu}$.

b) $\hat{\nu}_x \in \boldsymbol{M}(\phi'(x))$ and hence $\hat{\nu}$ has the same transition structure as ν. In particular, for any $a, b \in B$ and any integers $n > m$,

$$\hat{\nu}_x(Y_n = b | Y_m = a) = \nu_x(Y_n = b | Y_m = a).$$

c) If the original system is two-sided, then $\mu\nu$ is absolutely continuous with respect to $\bar{\mu}^*\hat{\nu}$ and hence $\mu\nu$ is AMS. If the original system is one-sided, then let $(\bar{\mu}^*\hat{\nu})'$ denote the restriction of the two-sided stationary measure $\bar{\mu}^*\hat{\nu}^*$ to a one-sided process, that is

$$(\bar{\mu}^*\hat{\nu})'(F) = \bar{\mu}^*\hat{\nu}((X_n, Y_n), (X_{n+1}, Y_{n+1}), \cdots \in F), \qquad F \in (\boldsymbol{A} \cdot \boldsymbol{B})_1^\infty.$$

Then if $(\bar{\mu}^*\hat{\nu})'(F) = 0$ for an invariant set $F \in (\boldsymbol{A} \cdot \boldsymbol{B})_1^\infty$, then also $\mu\nu(F) = 0$ and hence again $\mu\nu$ is AMS.

Indecomposable Markov Channels

Our development is based on the following observation of Kieffer and Rahe: if the two-sided stationary process $\bar{\mu}^*\hat{\nu}$ is also ergodic, then so is the original AMS process $\mu\nu$ from property c) above. Thus in order to prove that a Markov channel is ergodic, we must show that connecting a two-sided stationary ergodic source to a special two-sided stationary Markov channel yields a stationary and ergodic input–output process. Kieffer and Rahe used this approach to prove the result we describe next.

Given a stochastic matrix $P \in \boldsymbol{P}$, a nonempty set $B' \subset B$ is a *closed* set of states for P if

$$\sum_{b \in B'} P(a, b) = 1, \qquad a \in B'.$$

P is *decomposable* if there exist two disjoint closed sets of states. Otherwise P is *indecomposable*. A one-sided Markov channel $[A, \nu, B]$ such that $\nu_x \in \boldsymbol{M}(\phi(x))$ for a stationary map ϕ is said to be *indecomposable* if for every $x \in A_1^\infty$ and every positive integer n the product $\phi(x)_1\phi(x)_2 \cdots \phi(x)_n$ is an indecomposable stochastic matrix. (One can define indecomposability for a two-sided channel in a similar manner.) Kieffer and Rahe [9] proved that an indecomposable Markov channel is ergodic. It follows easily from their proof that if we only require that ν_x be indecomposable μ-a.e. for an ergodic stationary source μ in the sense that with μ probability one we get an x for which $\phi(x)_1\phi(x)_2 \cdots \phi(x)_n$ is an indecomposable stochastic matrix for all n, then $\mu\nu$ is also ergodic. For an AMS source a similar conclusion can be drawn by requiring that the channel be indecomposable almost everywhere with respect to the stationary mean.

Observe that stationarity of ϕ implies that if the channel is indecomposable, then the matrices $\phi(T^m x)_1\phi(T^m x)_2 \cdots \phi(T^m x)_n = \phi(x)_{m+1}\phi(x)_{m+2} \cdots \phi(x)_{m+n}$ are also indecomposable for all appropriate m and n.

It is easy to find important examples of Markov channels that are not indecomposable. In particular, finite state codes have matrices $\phi(x)_i$ that have only 1's and 0's and are often decomposable for a fixed x. This is a common occurrence, for example, in finite state vector quantizers designed by clustering algorithms based on a training sequence (Dunham and Gray [3]). As another example, a Markov channel could be such that for each x in a set of positive measure, there exists an n for which $\phi(x)_1, \cdots, \phi(x)_n$ is decomposable, but after some $N = N(x)$ all products from 1 to n are indecomposable. For example, a finite state channel in which the transition matrix is scrambling (as defined later) with positive probability has this property. In the second example the techniques of Kieffer and Rahe should still apply to yield an ergodic channel. Kieffer and Rahe make explicit use of the indecomposability property in their ergodicity proof. We here bypass this strong condition and a simpler proof of ergodicity is given using the strongly mixing condition of Adler [1].

The original definition of an indecomposable FSC of Blackwell, Breiman, and Thomasian [2] was different. They defined an FSC to be indecomposable if connecting any indecomposable Markov source to the FSC yielded an indecomposable input–output Markov chain. They proved, however, that for FSC's their definition and the above definition, in terms of products of transition matrices, were equivalent. While their proof of this equivalence does not appear to generalize to Markov channels, we will adopt the Kieffer–Rahe definition of indecomposable as the more natural for our purposes. In particular, it is not desirable to require Markov sources for determining the properties of a channel.

Gallager [5] gave another definition of an indecomposable FSC and stated that his definition produced the same class of channels as the Blackwell, Breiman, and Thomasian definition. The definition is quite different, however, and Gallager did not provide a proof. Furthermore, unpublished proofs are based on the construction of periodic input sequences to force certain channel behavior and hence the equivalence of these two definitions of indecomposability does not appear to extend to the more general μ-a.e. indecomposability definitions (since such periodic sequences will in general have total probability zero). In addition, the proof does not appear to generalize to all Markov channels. Hence we shall refer to this property as *indecomposability in the Gallager sense* or *Gallager indecomposability* and consider it separately. For any integer $m \in \boldsymbol{I}$, any integer $n \geq m$, and any $x \in \Sigma_A$, the stochastic matrix describing the output transition probabilities from time m to time n is given by

$$H_{mn}(x) = \{\nu_x(Y_n = k | Y_m = j);\ k, j = 1, 2, \cdots, K\} = \prod_{i=m}^{n} \phi(x)_i. \qquad (2)$$

Define a Markov channel to be *indecomposable in the Gallager sense* if for every $\epsilon > 0$ there is an N such that for all x, final states b, and initial states a, a'

$$|(H_{1n}(x))_{ab} - (H_{1n}(x))_{a'b}| < \epsilon, \qquad \text{if } n \geq N; \qquad (3)$$

that is, the channel is indecomposable if

$$\lim_{n \to \infty} |\nu_x(Y_n = b | Y_1 = a) - \nu_x(Y_n = b | Y_1 = a')| = 0 \quad (4)$$

uniformly in x, b, a, and a'. This is essentially a condition that the rows of the product matrices asymptotically become more alike uniformly in x. Gallager [5] proved that a finite state channel is indecomposable if and only if for some n and each x

$$(H_{1n}(x))_{ab} > 0, \qquad \text{all } a \qquad (5)$$

for some b; that is, $H_{1n}(x)$ must have for each x at least one column that has no zero entries. Furthermore, n can always be taken to be less than 2^{K^2}, where K is the number of channel states. In the general case of a Markov channel, Gallager indecomposability still implies the existence of a positive column since if the rows are the same in the limit, at least one column must be positive. Gallager's proof of the converse, however, does not immediately generalize and hence these conditions may not be equivalent in the general case. Hence when considering general Markov channels we will refer to (5) as the *strong positive column property* rather than as indecomposability. The adjective "strong" is because one n must serve for all x. For later use we generalize this property by dropping the uniformity requirement. A Markov channel is said to have the *positive column property* if for every x there is an n for which $H_{1n}(x)$ has a positive column.

As discussed by Kieffer and Rahe, Pfaffelhuber [13] showed that if a finite state channel is indecomposable in the Gallager sense, then a two-sided stationary channel constructed in a manner similar to Kieffer and Rahe yields an ergodic source when driven by a stationary and ergodic source, but Phaffelhuber did not show that this meant the original channel was ergodic.

Weakly Ergodic Markov Channels

In this section we introduce a new class of Markov channels by simply adopting a definition from the theory of nonhomogeneous Markov chains. We begin with a Markov channel $[A, \nu, B]$ and an AMS ergodic source μ. Let $\bar{\mu}^*$ and $\hat{\nu}$ be the induced two-sided stationary source and channel of the first section and let ϕ' be the two-sided map such that $\hat{\nu}_x \in \boldsymbol{M}(\phi'(x))$. As above, for any integer m, any integer $n \geq m$, and any $x \in A^\infty$, the stochastic matrix describing the output transition probabilities from time m to time n is given by

$$H^*_{mn}(x) = \{\hat{\nu}_x(Y_n = k | Y_m = j);\ k, j = 1,, \cdots, K\} = \prod_{i=m}^{n} \phi'(x)_i. \qquad (6)$$

Note that (2) and (6) are the same for the two-sided case because then $\phi' = \phi$. In particular, given x, $Y_1, Y_2, \cdots$ is a nonhomogeneous Markov chain with the transition probabilities given by either (2) (for the original channel) or (6) (for the induced two-sided stationary channel).

A nonhomogeneous Markov chain described by the transition matrices H_{mn}; $m = 1, 2, \cdots$; $n = m, \cdots$ is said to be *weakly ergodic* (Hajnal [7]) if

$$\lim_{n \to \infty} |(H_{mn})_{ij} - (H_{mn})_{kj}| = 0; \quad \text{all } m = 1, 2, \cdots; i, j, k = 1, 2, \cdots, K.$$

The theory of weakly ergodic nonhomogeneous Markov chains has been extensively studied in the literature (see, e.g., Hajnal [7], Paz [12], Kingman [10]). The definition immediately suggests a corresponding definition of a class of Markov channels: we shall say that a Markov channel $[A, \nu, B]$ is *weakly ergodic* if for all positive integers m (one-sided) or if for all integers m (two-sided)

$$\lim_{n \to \infty} |(H_{mn}(x))_{ij} - (H_{mn}(x))_{kj}| = 0; \quad \text{all } i, j, k = 1, 2, \cdots, K, \text{ all } x. \quad (7)$$

As usual, we shall say that a channel is weakly ergodic μ-a.e. for a source μ if (7) holds for all x in a set of μ probability one. Clearly with this definition we inherit all of the properties of weakly ergodic nonhomogeneous Markov chains, but that is not our goal; our focus is to randomly select one of these chains and determine when the joint input–output process is ergodic in the usual sense. Observe the strong resemblance between (7) and (3); a weakly ergodic Markov channel is a natural generalization of a channel that is indecomposable in the Gallager sense where we have simply dropped the requirement that the limits be uniform in x. Both conditions require that asymptotically the rows of a matrix become more alike, but the above condition permits different input sequences x to have different convergence rates.

Since the mapping ϕ is stationary, it follows that $H_{mn}(x)=H_{1(n-m)}(T^m x)$ and hence we need only verify (7) for the special case $m=1$ to prove a channel is weakly ergodic. The next lemma shows that a similar simplification holds for the almost everywhere definition if the source is stationary.

Lemma 1: Suppose that μ is a stationary source. Then a Markov channel ν is weakly ergodic a.e. if and only if with μ probability one,

$$\lim_{n\to\infty}|(H_{1n}(x))_{ij}-(H_{1n}(x))_{kj}|=0;\qquad \text{all } i,j,k=1,2,\cdots,K.$$

Proof: Define the sets F_m by

$$F_m=\Big\{x:\lim_{n\to\infty}|(H_{mn}(x))_{ij}-(H_{mn}(x))_{kj}|=0;\qquad \text{all } i,j,k=1,2,\cdots,K\Big\}.$$

Using the stationarity of ϕ,

$$\begin{aligned}F_m&=\Big\{x:\lim_{n\to\infty}|(H_{1n}(T^m x))_{ij}-(H_{1n}(T^m x))_{kj}|=0;\\ &\qquad \text{all } i,j,k=1,2,\cdots,K\Big\}\\ &=T^{-m}F_1\end{aligned}$$

and hence since μ is stationary $\mu(F_m)=\mu(T^{-m}F_1)=1$ all m and therefore $\mu(\cap_{m\in I}F^m)=1$ and the channel is weakly ergodic μ-a.e.

It will be convenient to use an alternative description of weakly ergodic Markov chains. Following Paz [12], given a stochastic matrix $P=\{P_{ts};t,s\in B\}$, define $\delta(P)$ by

$$\delta(P)=\max_{s,t}\sum_{k\in B}(P_{tk}-P_{sk})^+$$

where $(a)^+$ is a if a is positive and 0 otherwise. $\delta(P)$ is a measure of how unlike the rows of a stochastic matrix are. If, for example, the rows are all identical, $\delta(P)$ is 0. $\delta(P)$ is 1 if and only if at least two rows are orthogonal, that is, there exist two rows such that the nonzero entries of one are in a disjoint collection of columns from the nonzero entries of the other. Intuitively this means that there are two initial states which place all of their probability on disjoint collections of final states and hence we can distinguish between these two initial states by observing only the final state. These provide two extremes since $0\le\delta(P)\le 1$. We now collect some properties from Hajnal [7], Paz [12], and Wolfowitz [17]. A necessary and sufficient condition for $\delta(P)<1$ is that for any two rows i and k there is at least one column j for which both $P_{ij}>0$ and $P_{kj}>0$. A matrix with this property is said to be *scrambling*. A matrix is not scrambling if and only if at least two rows are orthogonal. A sufficient (but not necessary) condition for $\delta(P)<1$ is that P have at least one column with all entries nonzero. δ satisfies the following inequality:

$$\delta(PQ)\le\delta(P)\,\delta(Q).$$

Thus, for example, given a product of matrices, if any one of the matrices is scrambling, then the product is also. We note in passing that δ can be viewed as the maximum $\bar{d}$ distance between the rows of the stochastic matrix and this fact can be used to prove the above multiplicative inequality.

A nonhomogeneous Markov chain is weakly ergodic if and only if

$$\lim_{n\to\infty}\delta(H_{mn})=0;\qquad m=1,2,3,\cdots \tag{8}$$

(In fact, (8) is used by Paz as the definition for weakly ergodic.) The equivalence of the two properties is easily shown since both correspond to the rows of the sequence of matrices becoming more alike. Translating this into a statement for channels, and recalling the stationarity of ϕ and the equivalence of the transition matrices for the original channel and the induced stationary two-sided channel, we have the following lemma.

Lemma 2: A Markov channel is weakly ergodic if and only if

$$\lim_{n\to\infty}\delta(H_{1n}(x))=0;\qquad x\in\Sigma_A$$

or

$$\lim_{n\to\infty}\delta(H^*_{1n}(x))=0;\qquad x\in A^\infty.$$

If a Markov channel ν is weakly ergodic, then so is the induced stationary two-sided channel $\hat{\nu}$. Given a source $[A,\mu]$, a Markov channel is weakly ergodic μ-a.e. if

$$\mu\left(\left\{x:\lim_{n\to\infty}\delta\left(\prod_{i=m}^{n-1}\phi(x)_i\right)=0;m\in I\right\}\right)=1.$$

If the source is stationary, then from Lemma 1 only $m=1$ need be considered.

There is nothing "magic" about the particular choice of δ, any nonnegative function of a matrix with the properties that $\delta(PQ)\le\delta(P)\delta(Q)$ and the fact that $\delta(P_n)\to 0$ if and only if $|(P_n)_{ik}-(P_n)_{jk}|\to 0$ all i,j,k will work.

We now state and prove the first main result which provides an alternative characterization of Markov channels that are weakly ergodic a.e.

Theorem 1: A necessary condition for a Markov channel to be weakly ergodic a.e. $[\mu]$ for a stationary measure μ is that there exists an N such that

$$E\left[\ln\delta\left(\prod_{j=1}^{N}\phi(X)_j\right)\right]<0. \tag{9}$$

A sufficient condition for a Markov channel to be weakly ergodic a.e. $[\mu]$ for a stationary and ergodic measure μ is that there exists an N such that (9) holds.

Proof: The necessary condition is trivial. Let $M_i=\phi(X)_i$ denote the stationary and ergodic sequence of random stochastic matrices, and consider the limit of

$$\frac{1}{n}\ln\delta\left(\prod_{j=1}^{n}M_j\right).$$

Let N be such that (9) holds, let i be a nonnegative integer such that $i<N$, and let n be a large integer. We

can parse the product of n matrices into an initial piece of i matrices, a group of $K_n = \langle (n-i)/N \rangle$ blocks of N matrices each, where $\langle a \rangle$ denotes the largest integer contained in a, and a remaining group of $n - i - NK_n$ matrices and apply the product inequality for δ to write

$$\delta\left(\prod_{j=1}^{n} M_j\right) \le \delta\left(\prod_{j=1}^{i} M_j\right)\left(\prod_{j=0}^{K_n-1} \delta\left(\prod_{m=1}^{N} M_{i+jN+m}\right)\right) \cdot \delta\left(\prod_{j=n-i-K_nN}^{n} M_j\right) < \prod_{j=0}^{K_n-1} \delta\left(\prod_{m=1}^{N} M_{i+jN+m}\right).$$

Since $K_n \ge \langle (n-N)/N \rangle$, we have further that

$$\delta\left(\prod_{j=1}^{n} M_j\right) < \prod_{j=0}^{\langle (n-N)/N \rangle -1} \delta\left(\prod_{m=1}^{N} M_{i+jN+m}\right)$$

and hence that

$$\ln\delta\left(\prod_{j=1}^{n} M_j\right) \le \sum_{j=0}^{\langle (n-N)/N \rangle -1} \ln\delta\left(\prod_{m=1}^{N} M_{i+jN+m}\right).$$

Since this inequality is true for any i in the allowed range, we also have that

$$\frac{1}{n}\ln\delta\left(\prod_{j=1}^{n} M_j\right) \le \frac{1}{n}\frac{1}{N}\sum_{i=0}^{N-1}\sum_{j=0}^{\langle (n-N)/N \rangle -1} \ln\delta\left(\prod_{m=1}^{N} M_{i+jN+m}\right) = \frac{1}{n}\frac{1}{N}\sum_{j=0}^{N\langle (n-N)/N \rangle -1} \ln\delta\left(\prod_{m=1}^{N} M_{j+m}\right).$$

From the ergodic theorem, the right-hand sum converges as $n \to \infty$ with probability one to the expectation

$$\frac{1}{N} E \ln\delta\left(\prod_{m=1}^{N} M_m\right) < 0.$$

Thus we have shown that

$$\limsup_{n\to\infty} \frac{1}{n}\ln\delta\left(\prod_{i=1}^{n} M_i\right) < 0 \tag{10}$$

with probability one. This, however, implies that

$$\lim_{n\to\infty} \delta\left(\prod_{i=1}^{n} M_i\right) = 0.$$

To see this, assume the contrary. Since δ is nonnegative this would mean that a subsequence of $d_n = \delta(\prod_{i=1}^{N} M_i)$ would converge to some $d > 0$ which would imply from continuity of logarithms that a subsequence of $\ln d_n$ would converge to $\ln d$ and hence that a subsequence of $n^{-1}\ln d_n$ would converge to 0, violating (10) which says that all subsequences must converge to something strictly less than 0. This completes the proof.

The proof resembles the "easy half" of the subadditive ergodic theorem of Kingman [11]. In fact, the result can be proved by applying the subadditive ergodic theorem [11], [8] to the sequence $\ln\delta(\prod_{i=1}^{n}\phi(X)_i)$. The condition of the theorem is similar to conditions developed by Shi and Kozin [15] on matrices used in adaptive algorithms in order to prove the exponential convergence of those algorithms. They, however, used the results of Furstenberg and Kesten [4] on products of random matrices rather than the simpler tools used here.

The theorem provides easily several relations among the various classes of channels.

Corollary 1: Given a Markov channel and a stationary source μ the following conditions are equivalent (all statements are μ-a.e.).

a) The channel is weakly ergodic.
b) For each x there is an n such that no two rows of $H_{1n}(x)$ are orthogonal.
c) For each x there is an n such that H_{1n} is scrambling.
d) The channel has the positive column property.

Proof: The proof follows from the necessary condition of the theorem and hence does not require ergodicity of μ. Since $\delta(H_{1n}(x)) \le 1$ for all n, it can have an expectation of 1 for all n if and only if, for all x in a set of probability one, $\delta(H_{1n}(x)) = 1$ for every n. This can only be if for every n there are at least two orthogonal rows in the matrix product. Thus b) implies a). Obviously d) implies c) and c) implies b). That d) follows from a) follows analogously to the proof of Gallager's Theorem [5, th. 4.6.3]. Let K denote the size of the state alphabet and choose $\epsilon < 1/K$. Choose n so large that

$$|(H_{1n}(x))_{ik} - (H_{1n}(x))_{jk}| < \epsilon$$

for all k, j, i, and observe that one of the columns must have an element no smaller than $1/K$. Hence, all the entries in that column are positive.

John Kieffer has pointed out to the authors that the results of Theorem 1 and Corollary 1 can also be derived as consequences of results of Nawrotzki [18] (see also Cogburn [19]) which provide sufficient conditions for the maximal $\bar{d}$-distance between the rows of $M_1 M_2 \cdots M_n$ to go to zero almost surely and hence for condition c) of Corollary 1 to hold.

Corollary 2: A sufficient condition for a Markov channel to be weakly ergodic μ-a.e. for a stationary μ is that it be indecomposable in the Gallager sense μ-a.e.

Proof: This follows immediately from the fact that the strong positive column property implies the positive column property and that Gallager indecomposability implies the strong positive column property.

Corollary 3: A sufficient condition for a Markov channel to be weakly ergodic μ-a.e. for a stationary μ is that it be indecomposable μ-a.e.

Proof: From Wolfowitz [17, Lemma 4] there is a finite number $t = t(K)$ such that all products $H_{1n}(x) = \phi(x)_1 \cdots \phi(x)_n$ with $n > t$ have $\delta(H_{1n}(x)) < 1$ if all such products are indecomposable. Since this holds for a set of measure one, (9) must hold and the conclusion follows from the theorem.

The result of Wolfowitz used above suggests an interesting class of weakly ergodic channels that are not indecomposable. Consider a finite state channel with state transition probability matrices P_a; $a \in A$, where A is finite. Suppose that one of these matrices, say P_b, is bad in the sense that all finite products of the remaining matrices are indecomposable, but finite products including the bad matrix P_b may not be (e.g., P_b itself is decomposable). Suppose a source is connected to the channel that produces a sequence of independent identically distributed symbols according to a distribution which places a small nonzero probability on b. Clearly $\phi(x)_m \cdots \phi(x)_n = P_{x_m} \cdots P_{x_n}$ will not always be indecomposable with probability 1 since the matrices P_b can occur with nonzero probability; hence the channel is not indecomposable a.e. If the probability that P_b occurs is small enough, however, with probability one we will get an x such that decomposable matrices will occur sufficiently rarely to ensure that for some m and $n > m + t$ there will be a subsequence $P_{x_m} \cdots P_{x_n}$ having all subproducts indecomposable and hence having $\delta(P_{x_m} \cdots P_{x_m}) < 1$ and hence having $\delta(P_{x_1} \cdots P_{x_n}) < 1$. Since this happens with probability one, again (9) must hold and the channel must be weakly ergodic μ-a.e.

Mixing Markov Channels

Before stating the second principal result, we require yet another definition of a class of channels. Following Adler [1] we define a channel to be *strongly mixing* μ-a.e. (or *asymptotically independent of the remote past*) for a stationary measure μ if for any finite-dimensional cylinder sets F and G we have that

$$\lim_{n \to \infty} |\nu_x(F \cap T^{-i}G) - \nu_x(F)\nu_x(T^{-i}G)| = 0. \qquad (11)$$

Adler built the modifier "a.e." into the definition, but we make it explicit for consistency with the other definitions.

Lemma 3: Given a stationary measure μ, a Markov channel is weakly ergodic μ-a.e. if and only if it is strongly mixing μ-a.e.

Proof: Eq. (11) is trivially true if $\nu_x(F)$ is 0, so an equivalent definition of strongly mixing is to require (11) to hold for all F with $\nu_x(F) > 0$. In this case we can divide by $\nu_x(F)$ to obtain an equivalent condition using conditional probabilities. A channel is strongly mixing if and only if for all finite dimensional cylinders F and G

$$\lim_{n \to \infty} |\nu_x(T^{-n}G|F) - \nu_x(T^{-n}G)| = 0. \qquad (12)$$

This in turn is equivalent to showing that (12) holds for all thin cylinders of the form

$$F = \{ y \colon y_m = a_1, y_{m+1} = a_2, \cdots, y_{m+k-1} = a_k \}$$

and

$$G = \{ y \colon y_m = b_1, y_{m+1} = b_2, \cdots, y_{m+j-1} = b_j \}$$

since general cylinders are finite unions of disjoint thin cylinders and if (12) holds for, say, F_1, G_1 and F_2, G_2, then it also holds for $F_1 \cup F_2, G_1 \cup G_2$ using the triangle inequality. Consider first one-dimensional cylinders of the form $F = \{ y \colon y_m = a \}$, $G = \{ y \colon y_m = b \}$. If the channel is strongly mixing, then

$$|\nu_x(Y_{m+n} = b|Y_m = a) - \nu_x(Y_{m+n} = b)| \underset{n \to \infty}{\to} 0; \qquad \text{all } a \in B$$

and hence by the triangle inequality also

$$|\nu_x(Y_{m+n} = b|Y_m = a) - \nu_x(Y_{m+n} = b|Y_m = a')| \underset{n \to \infty}{\to} 0; \qquad \text{all } a, a' \in B;$$

that is, the channel must be weakly ergodic. Conversely, if the channel is weakly ergodic, then

$$|\nu_x(Y_{m+n} = b|Y_m = a) - \nu_x(Y_{m+n} = b|Y_m = a')| \underset{n \to \infty}{\to} 0$$

for any a, a', b. Thus we must also have that

$$\begin{aligned} &|\nu_x(Y_{m+n} = b|Y_m = a) - \nu_x(Y_{m+n} = b)| \\ &\quad = |\nu_x(Y_{m+n} = b|Y_m = a) - \sum_{a' \in B} \nu_x(Y_m = a') \\ &\quad \cdot \nu_x(Y_{m+n} = b|Y_m = a')| \\ &\quad \le \sum_{a' \in B} \nu_x(Y_m = a') \\ &\quad \cdot |\nu_x(Y_{m+n} = b|Y_m = a) - \nu_x(Y_{m+n} = b|Y_m = a')| \\ &\quad \cdot \underset{n \to \infty}{\to} 0 \end{aligned}$$

and hence (12) must hold for such sets. The result for general thin cylinders follows similarly since, given x, the Y_n form a Markov chain. In particular, define $Y_m^k = (Y_m, Y_{m+1}, \cdots, Y_{m+k-1})$, the random vector beginning with the sample at time m and having dimension k and $b^k = (b_1, b_2, \cdots, b_k)$. Let F be the event $Y_m^k = a^k$ and G be the event $Y_m^j = b^j$. Then once $n > k$ there is no overlap of the events F and $T^{-n}G$ and hence $\nu_x(Y_{n+m}^j = b^j | Y_m^k = b^k)$ depends on b^k only through $\nu_x(Y_{n+m} = b_1 | Y_{m+k-1} = b_k)$, which asymptotically does not depend on b^k by the one-dimensional result. This completes the proof of the lemma.

The lemma coupled with Adler [1] provides the final link for the second main result of this paper, which begins the next section.

Ergodic Markov Channels

Theorem 2: If a stationary Markov channel ν is weakly ergodic a.e. $[\mu]$ for a stationary and ergodic source μ, then $\mu\nu$ is stationary and ergodic. A Markov channel is ergodic if it is weakly ergodic μ-a.e. with respect to all stationary measures (e.g., if it is weakly ergodic everywhere).

Proof of Theorem 2: If ν is stationary and weakly ergodic, then from the previous lemma it is strongly mixing. Adler [1] proves that connecting a stationary ergodic source μ to ν will then provide a stationary and ergodic $\mu\nu$. This proves the first statement. If ν is Markov and hence

AMS, form the stationary two-sided channel $\hat{\nu}$ as described previously. If ν is weakly ergodic, then so is $\hat{\nu}$. Thus $\hat{\nu}$ is stationary and, from the lemma, strongly mixing. The second statement follows from the first by applying the first statement to the two-sided and stationary source $\bar{\mu}^*$ and channel $\hat{\nu}$ and recalling the observation of Kieffer and Rahe [9] that ergodicity of $\bar{\mu}^*\hat{\nu}$ implies that $\mu\nu$ is ergodic.

Theorem 2 provides the most general class of ergodic AMS channels currently known.

Acknowledgment

The authors gratefully acknowledge the helpful comments of Professors R. G. Gallager and John C. Kieffer.

References

[1] R. L. Adler, "Ergodic and mixing properties of infinite memory channels," *Proc. Amer. Math. Soc.*, 12, pp. 924–930, 1960.

[2] D. Blackwell, L. Breiman, and A. Thomasian, "Proof of Shannon's transmission theorem for finite-state indecomposable channels," *Ann. Math. Statist.*, vol. 18, pp. 1209–1220, 1958.

[3] M. O. Dunham and R. M. Gray, "An algorithm for the design of labeled-transition finite-state vector quantizers," *IEEE Trans. Commun.*, COM-33, pp. 83–89, 1985.

[4] H. Furstenberg and H. Kesten, "Products of random matrices," *Ann. Math. Statist.*, vol. 31, pp. 457–469, 1960.

[5] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[6] R. M. Gray and J. C. Kieffer, "Asymptotically mean stationary measures," *Ann. Prob.*, 8, pp. 962–973, 1980.

[7] J. Hajnal, "Weak ergodicity in non-homogeneous Markov chains," *Proc. Cambridge Philos. Soc.*, 54, pp. 233–246, 1958.

[8] Y. Katznelson and B. Weiss, "A simple proof of some ergodic theorems," *Israel Journ. Math.*, 42, pp. 291–296, 1982.

[9] J. C. Kieffer and M. Rahe, "Markov channels are asymptotically mean stationary," *Siam J. Math. Anal.*, 12, pp. 293–305, 1981.

[10] J. F. C. Kingman, "Geometrical aspects of the theory of non-homogeneous Markov chains," *Math. Proc. Camb. Phil. Soc.*, 77, pp. 171–183, 1975.

[11] ——, "Subadditive ergodic theory," *Ann. Prob.*, 1, pp. 883–909, 1973.

[12] A. Paz, *Stochastic Automata Theory*. New York: Academic, 1971.

[13] E. Pfaffelhuber, "Channels with asymptotically decreasing memory and anticipation," *IEEE Trans. Inform. Theory*, 17, pp. 379–385, 1971.

[14] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, 27, pp. 379–423, 623–656, 1948.

[15] D. H. Shi and F. Kozin, "On almost sure convergence of adaptive algorithms," preprint.

[16] A. Viterbi and J. Omura, *Principals of Digital Communication and Coding*. New York: McGraw-Hill, 1979.

[17] J. Wolfowitz, "Products of indecomposable, aperiodic, stochastic matrices," *Proc. Amer. Math. Soc.*, vol. 14, pp. 733–737, 1963.

[18] K. Nawrotzki, "Discrete open systems for Markov chains in a random environment I," *Elektron. Informationsverarb. Kybernet.*, vol. 17, pp. 569–599, 1981.

[19] R. Cogburn, "On products of random stochastic matrices and their applications," *Contemporary Mathematics*, vol. 50, pp. 199–213, 1984.