Hidden Markov Processes

Yariv Ephraim, *Fellow, IEEE*, and Neri Merhav, *Fellow, IEEE*

Invited Paper

***Abstract*—An overview of statistical and information-theoretic aspects of hidden Markov processes (HMPs) is presented. An HMP is a discrete-time finite-state homogeneous Markov chain observed through a discrete-time memoryless invariant channel. In recent years, the work of Baum and Petrie on finite-state finite-alphabet HMPs was expanded to HMPs with finite as well as continuous state spaces and a general alphabet. In particular, statistical properties and ergodic theorems for relative entropy densities of HMPs were developed. Consistency and asymptotic normality of the maximum-likelihood (ML) parameter estimator were proved under some mild conditions. Similar results were established for switching autoregressive processes. These processes generalize HMPs. New algorithms were developed for estimating the state, parameter, and order of an HMP, for universal coding and classification of HMPs, and for universal decoding of hidden Markov channels. These and other related topics are reviewed in this paper.**

***Index Terms*—Baum–Petrie algorithm, entropy ergodic theorems, finite-state channels, hidden Markov models, identifiability, Kalman filter, maximum-likelihood (ML) estimation, order estimation, recursive parameter estimation, switching autoregressive processes, Ziv inequality.**

I. Introduction

A *hidden Markov process* (HMP) is a discrete-time finite-state homogeneous Markov chain observed through a discrete-time memoryless invariant channel. The channel is characterized by a finite set of transition densities indexed by the states of the Markov chain. These densities may be members of any parametric family such as Gaussian, Poisson, etc. The initial distribution of the Markov chain, the transition matrix, and the densities of the channel depend on some parameter that characterizes the HMP. The process is said to be a *finite-alphabet* HMP if the output alphabet of the channel is finite. It is said to be a *general* HMP when the output alphabet of the channel is not necessarily finite.

HMPs are more commonly referred to as *hidden Markov models*. The term HMP was chosen since it emphasizes the process itself rather than its use as a model. HMPs comprise a rich family of parametric processes that was found useful in many applications. HMPs are closely related to mixture processes, switching autoregressive processes, dynamical systems in the sense of control theory, Markov-modulated Poisson processes, composite sources, and unifilar sources. HMPs are fairly general processes that are amenable to mathematical analysis.

HMPs have been widely studied in statistics. An HMP is viewed as a discrete-time bivariate parametric process. The underlying process is a finite-state homogeneous Markov chain. This process is not observable and is often referred to as the *regime*. The second process is a sequence of conditionally independent random variables given the Markov chain. At any given time, the distribution of each random variable depends on the Markov chain only through its value at that time. This distribution is time-invariant and it may be a member of any parametric family. The sequence of conditionally independent random variables is often referred to as the *observation* sequence.

HMPs are commonly encountered in information theory. Markov chains are common models for information sources with memory, and memoryless invariant channels are among the simplest models for communication channels. The hookup of Markov chains with memoryless channels yields a family of processes that are far more complex than the Markov chain sources. For example, there is no closed-form single-letter expression for the entropy rate of an HMP. Also, the method of types does not apply to HMPs unless they are unifilar sources. The state sequence of a unifilar source depends deterministically on the observation sequence and the initial state.

In recent years, the theory of HMPs has been substantially advanced and a wealth of new results was developed. In addition, numerous new applications have emerged. In the statistical literature, the main focus has been on HMPs with finite- as well as continuous-state spaces and a general alphabet. Identifiability of an HMP, consistency and asymptotic normality of the maximum likelihood (ML) parameter estimator, as well as algorithms for estimating the state, parameter, number of states, and the Fisher information matrix, were developed. The number of states of an HMP is called the *order*. In information theory, the main focus has been on finite-state finite-alphabet HMPs where order estimation, universal coding and classification of HMPs, and universal decoding of finite-state channels, which are hidden Markov channels, were studied.

Our goal is to present an overview of HMPs from the statistical and information-theoretic viewpoints. Our primary focus is on the theory of HMPs as it evolved in recent years. We also provide a brief survey of many applications of HMPs. Some

Manuscript received July 20, 2000; revised December 30, 2001.

Y. Ephraim is with the Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA 22030 USA (e-mail: yephraim@gmu.edu).

N. Merhav is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: merhav@ee.technion.ac.il).

Communicated by S. Shamai, Guest Editor

Publisher Item Identifier S 0018-9448(02)04014-2.

sections of the paper require some background in probability theory. To facilitate reading, we have collected preliminary measure-theoretic material in one section. This manuscript is divided into fifteen sections. The plan for each of the remaining sections is outlined below.

II. *A Brief History*: Provides a brief history of HMPs and a review of the main theoretical results developed in recent years.

III. *Preliminaries*: Sets up the notation and provides some preliminary background material.

IV. *Statistical Properties*: Defines HMPs and their relations to mixture processes, switching autoregressive processes, dynamical systems, Markov-modulated Poisson processes, composite sources, and unifilar sources. Also defines hidden Markov channels. Summarizes statistical properties of HMPs such as stationary, mixing, and ergodic properties. These properties are inherited from the Markov chains. Provides ergodic theorems for the sample entropy and relative entropy densities of HMPs.

V. *State Estimation*: Presents numerically stable and computationally efficient recursions for prediction, filtering, and fixed-interval smoothing of the state sequence of the HMP. The recursions coincide with the Kalman filter and smoother, respectively, under linear Gaussian assumptions. The recursions are naturally stable, and they differ from those traditionally used in signal processing and communication applications such as automatic speech recognition and decoding of turbo codes, respectively.

VI. *ML Parameter Estimation*: Deals with several aspects of ML parameter estimation. Provides conditions for identifiability of an HMP. States theorems for consistency and asymptotic normality of the ML parameter estimator of an HMP with a finite as well as continuous-state space and a general alphabet. Provides similar theorems for switching autoregressive processes. Outlines the principles of the Baum algorithm for local ML parameter estimation, and Louis's formula for estimating the Fisher information matrix. States the Ziv inequality which provides a tight upper bound on the maximum value of the likelihood function for *any* finite-alphabet HMP.

VII. *Joint State and Parameter Estimation*: Focuses on joint estimation of the state sequence and parameter of an HMP. Presents the Baum–Viterbi algorithm and its relations to the Baum algorithm and to the generalized Lloyd algorithm for designing vector quantizers. The algorithm is useful when a sufficiently long vector of observations is generated from each state. Otherwise, it does not provide a consistent estimate of either the parameter or the state sequence. Describes a noniterative algorithm for global maximization of the joint likelihood function of states and observations of a left–right HMP. Discusses Bayesian estimation of the state sequence and parameter, and asymptotic properties of the estimator.

VIII. *Order Estimation*: Presents consistent estimators for a finite-alphabet HMP, and an estimator which does not underestimate the order of a general HMP.

IX. *Dynamical System Approach*: The HMP is seen as a dynamical system in the sense of control theory, and its parameter is estimated using the expectation–maximization algorithm. Conditional mean estimators of several statistics of the HMP, required by the expectation–maximization algorithm, are developed using the generalized Bayes rule. The approach is demonstrated for HMPs with Gaussian densities. The approach is particularly useful for continuous-time HMPs but this extension is not reviewed here.

X. *Recursive Parameter Estimation*: Describes algorithms for recursive estimation of the parameter of an HMP. A consistent asymptotically normal estimator is provided.

XI. *Signal Classification*: Deals with several classification problems involving HMPs including universal classification.

XII. *Signal Estimation*: The HMP is seen as a desired signal and its estimation from a noisy signal is discussed.

XIII. *Hidden Markov Channels*: Reviews some properties of finite-state channels such as capacity and the channel coding theorem. Presents the Lapidoth–Ziv asymptotically optimal universal decoding algorithm for finite-state channels.

XIV. *Selected Applications*: Briefly describes selected applications in communications, information theory, and signal processing. Also presents special forms of HMPs and non-ML parameter estimation procedures which were found useful in practice.

XV. *Concluding Remarks*.

II. A Brief History

HMPs were introduced in full generality in 1966 by Baum and Petrie [25] who referred to them as *probabilistic functions of Markov chains*. Indeed, the observation sequence depends probabilistically on the Markov chain. During 1966–1969, Baum and Petrie studied statistical properties of stationary ergodic finite-state finite-alphabet HMPs. They developed an ergodic theorem for almost-sure convergence of the relative entropy density of one HMP with respect to another. In addition, they proved consistency and asymptotic normality of the ML parameter estimator [25], [251]. In 1969, Petrie [251] provided sufficient conditions for identifiability of an HMP and relaxed some of the assumptions in [25]. In 1970, Baum, Petrie, Soules, and Weiss [28], [29] developed forward–backward recursions for calculating the conditional probability of a state given an observation sequence from a general HMP. They also developed a computationally efficient iterative procedure for ML estimation of the parameter of a general HMP using the forward–backward recursions. This procedure is the well-known *expectation–maximization* (EM) algorithm of Dempster, Laird, and Rubin [80] applied to HMPs. Local convergence of the algorithm was established in [28], [29]. The algorithm is often referred to as the Baum algorithm, or the Baum–Petrie algorithm, or the Baum–Welch algorithm in honor of Lloyd Welch [311]. Similar forward–backward recursions were developed earlier by Chang and Hancock [56] in their work on optimal decoding of intersymbol interference channels.

Prior to the introduction of probabilistic functions of Markov chains, *deterministic functions of Markov chains* were extensively studied. They are often referred to as *aggregated Markov processes* in the statistical literature since a function may collapse several states of the Markov chain onto a single letter. Deterministic and probabilistic functions of finite-state Markov chains are related when the alphabet of the HMP is finite. Any deterministic function of a Markov chain can be described as a trivial finite-alphabet HMP, and any finite-alphabet HMP can be described as a deterministic function of Markov chain with an augmented state space [25], [251], [116]. Deterministic functions of Markov chains were used by Shannon in 1948 [290] as models for information sources. Ash [14, p. 185] refers to them as *Markov sources* but the term has more often been associated with unifilar sources introduced by Gallager [133, Sec. 3.6]. Shannon developed the fundamental ergodic theorem for convergence in probability of the sample entropy of a stationary ergodic Markov chain [290]. The theorem was proved for stationary ergodic finite-alphabet processes, for L^1 and almost sure convergence, by McMillan and Breiman, respectively. It is commonly referred to as the Shannon–McMillan–Breiman theorem or as the *aymptotic equipartition* property [152, Ch. 3]. The theorem applies to any stationary ergodic finite-alphabet HMP. Deterministic functions of Markov chains were also intensively studied in the statistical literature, notably by Blackwell [41], Blackwell and Koopmans [42], Burke and Rosenblatt [52], Gilbert [136], Fox [125], Dharmadhikari [83]–[86], Heller [160], and Carlyle [54], who investigated identifiability and conditions for deterministic functions of Markov chains to be Markov chains.

HMPs comprise a rich family of parametric random processes. In the context of information theory, we have already seen that an HMP is a Markov chain observed through a memoryless channel. More generally, consider a finite-state channel [133, Sec. 4.6]. The transition density of the channel depends on a nonobservable Markov chain. This channel is sometimes called a *hidden Markov channel*. An HMP observed through a finite-state channel is an HMP with an augmented state space. The Gilbert–Elliott channel is an important example of a finite-state channel [137], [97], [14], [243], [204]. This channel introduces a binary additive hidden Markov noise process which is independent of the input process. The Gilbert–Elliott channel is a good model for fading channels. Finite-state channels are also known as *stochastic sequential machines* (SSMs) or *probabilistic automata* [250]. A subclass of SSMs is formed by *partially observable Markov decision processes* [242].

HMPs are also related to a number of random processes commonly encountered in engineering, statistics, and econometrics. We first point out the obvious relation to mixture processes [212], [109], [232], [266], [301]. Each observation of an HMP has a mixture distribution, but contrary to mixture processes, HMP observations need not be statistically independent. HMPs are special cases of switching autoregressive processes with Markov regimes [156, Ch. 22]. These are autoregressive processes whose dynamics at each time instant depend on the state of a Markov chain at that time. When the autoregressive order is zero, the switching autoregressive process degenerates to an HMP. HMPs may be cast as dynamical systems in the sense of control theory. When the state space is finite or countably infinite, each state is represented by a unit vector in a Euclidean space. Another relation is to Markov-modulated Poisson processes [117], [273], [276]. These are Poisson processes whose rate is controlled by a nonobservable continuous-time Markov chain. A Markov-modulated Poisson process may be viewed as a Markov renewal process and as an HMP. In both cases, a discrete-time Markov chain is defined by sampling the continuous-time chain at the Poisson event epochs, and the observation sequence is given by the interevent time durations.

One of the earliest applications of HMPs was to automatic character recognition. Raviv [265] studied the problem in 1967 at the IBM T. J. Watson Research Center. The characters of the language were represented by states of the Markov chain and the measurements constituted the observation process. Recognition in the minimum character error rate sense was performed. For that purpose, Raviv developed a new recursion for the conditional probability of a state given the observations.

In the mid-1970s, another major application of HMPs was taking place at the IBM T. J. Watson Research Center. Jelinek [172], Baker [21], Jelinek, Bahl, and Mercer [171], Bahl and Jelinek [18], along with their coworkers, developed a phonetic speech recognition system that relies on hidden Markov modeling of speech signals. The model for each word in the vocabulary was composed of individual phonetic models which were designed using the Baum algorithm. Linguistic decoding of an acoustic utterance was performed using the Viterbi algorithm [308], [124], [285] or the Stack graph-search algorithm of Jelinek [170]. In the early 1980s, applications of HMPs to automatic speech recognition were further studied primarily by Ferguson and his colleagues at the Institute for Defense Analysis [115], [256], and by Rabiner and his group at AT&T Bell Laboratories [262]. These studies popularized the theory of HMPs which have since become widespread in many applications. In Ferguson [115], probabilistic functions of Markov chains were probably first referred to as *hidden Markov models*.

In recent years, HMPs have been widely studied by statisticians and information theorists. Significant progress has been made in the theory of HMPs where the work of Baum and Petrie on finite-state finite-alphabet HMPs was expanded to HMPs with finite as well as continuous-state spaces and a general alphabet. In particular, new ergodic theorems for relative entropy densities of HMPs were developed by Leroux [214], Finesso [116], Le Gland and Mevel [210], and Douc and Matias [90]. Consistency and asymptotic normality of the ML estimator of the parameter of an HMP was proved by Leroux [214], Bickel, Ritov, and Rydén [36], Le Gland and Mevel [210], Jensen and Petersen [174], and Douc and Matias [90]. The ergodic theorems and asymptotic optimality of the ML parameter estimator were also proved for switching autoregressive processes with Markov regime by Francq and Roussignol [127], Krishnamurthy and Rydén [198], and Douc, Moulines, and Rydén [91]. Similar results were developed for a Markov-modulated Poisson process by Rydén [273], [276]. Exponential forgetting and geometric ergodicity in HMPs were studied by Le Gland and Mevel [210] and Douc and Matias [90]. A complete solution to identifiability of deterministic functions of nonstationary Markov chains was given by Ito, Amari, and Kobayashi [167]. Conditions for identifiability of a general HMP were developed by Leroux [214] and Rydén [274], [277]. Conditions for identifiability of a Markov modulated Poisson process were given by Rydén [278]. New stable recursions for prediction, filtering, and fixed-interval smoothing of the state sequence from an observation sequence were developed by Lindgren [219] and Askar and Derin [15]. These recursions provide conditional mean filters and smoothers for Markov chains observed through channels that are not necessarily Gaussian [203].

In addition to expanding the work of Baum and Petrie, other approaches to HMPs were developed in recent years. A comprehensive dynamical system approach to general HMPs was developed by Elliott, Aggoun, and Moore [99]. In particular, finite-dimensional recursions for conditional mean estimators of statistics of a general HMP were developed, and used in ML estimation of the parameter of the process. HMPs with discrete- as well as continuous-time state and observation processes, that have finite or continuous alphabet, were studied in [99]. Information-theoretic approaches for strongly consistent order estimation of a finite-alphabet HMP were developed by Finesso [116], Kieffer [187], and Liu and Narayan [223]. An order estimator for a general HMP that does not underestimate the true order was developed by Rydén [277]. A consistent asymptotically normal *recursive* estimator for the parameter of a general HMP was developed by Rydén [279]. A Gibbs sampling Bayesian approach for estimating the parameter of a general HMP was developed by Robert, Celeux, and Diebold [269].

In communications and information theory, several aspects of HMPs were studied in recent years. Minimum symbol error-rate decoding of convolutional and linear codes using the forward–backward recursions of Chang and Hancock [56] was proposed by Bahl, Cocke, Jelinek, and Raviv [17]. The algorithm has since been referred to as the BCJR algorithm, and a stabilized version of the recursions is commonly used in decoding turbo codes [32], [33]. *Turbo codes* use several concatenated convolutional codes and a feedback mechanism that allow iterative reduction of the bit error rate. They almost achieve the Shannon capacity in communication over memoryless Gaussian channels. Properties of *composite sources,* which are generalizations of HMPs, were studied by Fontana [119], and Fontana, Gray, and Kieffer [120]. The Lempel–Ziv universal data compression algorithm introduced in 1978 [326] is applicable to universal coding of finite-alphabet HMPs. This algorithm asymptotically outperforms any finite-state coding scheme in compressing sequences from any source, not necessarily an HMP. Large-deviations properties of the Lempel–Ziv algorithm for HMPs were developed by Merhav [236]. Significant progress in universal classification of Markov chains of any order using empirically observed sequences was made by Ziv [328], Gutman [155], and Zeitouni, Ziv, and Merhav [325]. Universal classification of HMPs using empirically observed training sequences was developed by Merhav [235], Merhav and Ephraim [238], and Kieffer [187]. A universal decoding algorithm for finite-state channels was developed by Ziv [327], and Lapidoth and Ziv [204]. An algorithm for decoding unknown intersymbol interference channels using the Baum algorithm was developed by Kaleh and Vallet [179].

Along with the advances in the theory of HMPs, numerous new applications of HMPs have emerged in recent years in areas such as neurophysiology, biology, economics, control, spectral estimation, radar, sonar and image signal processing, fault detection, computer vision, robotics, and metrology.

III. Preliminaries

In this section, we provide some preliminary background material. We also describe the notation that we use throughout the manuscript. Some additional notation will be introduced in Section IV-A where the specifics of the HMP are discussed.

A. General Definitions

All random variables in a given discussion are defined on a common probability space $(\Omega, \mathcal{F}, P)$. We use capital letters to denote random variables, lower case letters to denote realizations of random variables, and script letters to denote sets within which the random variables take values. For example, a random variable X takes values $\{x\}$ in $\mathcal{X}$. We write $P(F)$ to denote the probability of an event $F \in \mathcal{F}$. We also write $P(X = x)$ to denote the probability of the event $\{\omega \in \Omega\colon X(\omega) = x\}$.

A random variable X defined on the underlying probability space induces a probability space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, P_X)$. The random variable X takes values in the sample space $\mathcal{X}$. The σ-field $\mathcal{B}_{\mathcal{X}}$ denotes the Borel σ-field of open subsets of $\mathcal{X}$ with respect to a given metric. The probability measure P_X denotes the *distribution* of X. Usually $\mathcal{X}$ is the real line $\mathcal{R}$ or a subset of the real line. The probability space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, P_X)$ is referred to as the *associated probability space* of X [152, p. 11]. We shall usually work with this probability space rather than with the underlying probability space. The sample space $\mathcal{X}$ may also be referred to as the *alphabet* of X and members of $\mathcal{X}$ may be referred to as *letters* of the alphabet. We assume that all distributions are absolutely continuous with respect to some σ-finite measure, say μ, and hence possess *densities* or Radon–Nikodym derivatives [38, Theorem 32.2]. We denote absolute continuity of P_X with respect to μ by $P_X \ll \mu$. We denote the density of P_X with respect to μ by $p(x)$. We shall not stress the role of X in the notation of the density and use $p(x)$ instead of $p_X(x)$. When the dominating measure μ is the Lebesgue measure we may refer to the density as the probability density function (pdf). When the dominating measure μ is the counting measure we may use the term probability mass function (pmf) instead of density. These two dominating measures are of particular interest in applications of HMPs.

A discrete-time random process, say $\boldsymbol{X}$, is denoted by $\{X_t, t \in \mathcal{T}\}$ where $\mathcal{T}$ is the *index set* or a subset of all integers. For one-sided random processes, $\mathcal{T}$ is usually the set of all positive integers. In some discussions, $\mathcal{T}$ is more naturally chosen to be the set of all nonnegative integers. For two-sided random processes, $\mathcal{T}$ is the set of all integers. When the index set is clear from the context, we use the simpler notation of $\{X_t\}$. Assume that $x_t \in \mathcal{X}$ for all $t \in \mathcal{T}$. The random process is defined on the underlying probability space $(\Omega, \mathcal{F}, P)$ and has an associated measurable product space $(\mathcal{X}^{\mathcal{T}}, \mathcal{B}_{\mathcal{X}}^{\mathcal{T}})$. We are particularly interested in a random process defined by a distribution on $(\mathcal{X}^{\mathcal{T}}, \mathcal{B}_{\mathcal{X}}^{\mathcal{T}})$ which is a member of a given parametric family. Let $\phi \in \Phi$ denote the *parameter* of the process distribution where Φ is the *parameter set*. Usually $\Phi \subseteq \mathcal{R}^{d_0}$ where $\mathcal{R}^{d_0}$ is a d_0-dimensional Euclidean space. Let P_ϕ denote the parametric distribution of the process. The associated *sequence probability space* of the random process is $(\mathcal{X}^{\mathcal{T}}, \mathcal{B}_{\mathcal{X}}^{\mathcal{T}}, P_\phi)$. We denote by $\phi^0 \in \Phi$ the *true* parameter used to generate a given realization $\{x_t, t \in \mathcal{T}\}$ of the process.

A sequence of random variables of the process, $\{X_l, \ldots, X_m\}$, $m > l$, is denoted by X_l^m. A realization of X_l^m is denoted by x_l^m. Most commonly, we will consider a sequence of n random variables, X_1^n, which, for simplicity, we denote by X^n. Let $P_\phi^{(n)}$ denote the n-dimensional distribution of X^n induced by P_ϕ. For each n, the distribution $P_\phi^{(n)}$ is assumed absolutely continuous with respect to some σ-finite measure μ^n and its density with respect to that measure is denoted by $p(x^n; \phi)$. The explicit dependency of this density on ϕ may be suppressed when notation may be simplified. The expected value of a measurable function $g(X^n)$ with respect to the probability measure $P_{\phi^0}^{(n)}$ is denoted by $E_{\phi^0}\{g(X^n)\}$. Of particular interest is the expected value of $\log p(X^n; \phi)$ with respect to $P_{\phi^0}^{(n)}$ given by

$$E_{\phi^0}\{\log p(X^n; \phi)\} = \int \log p(x^n; \phi) P_{\phi^0}(dx^n).$$

The usual notation for conditional probabilities and densities is adopted here. For example, the density of X_t given X^{t-1} is denoted by $p(x_t|x^{t-1})$.

In some sections of the paper we report results that are applicable to *standard* measurable spaces $(\mathcal{X}^n, \mathcal{B}_{\mathcal{X}}^n)$. The definition and properties of standard spaces can be found in [151, Ch. 2], [152, p. 12]. Standard spaces include discrete spaces, the real line, Euclidean vector spaces, Polish spaces which are complete separable metric spaces, among other examples. Standard spaces form a general class of measurable spaces for which the Kolmogorov extension theorem holds, regular conditional probability measure exist, and the ergodic decomposition theorem holds [152, p. 12].

We shall also make the following conventions. We say that a stochastic matrix A satisfies $A > \delta$ if all of its entries are larger than δ. Let A and B be two stochastic matrices of possibly different order. Suppose that $\phi = (A, B)$. We say that $\phi \in \Phi_\delta$ if both $A > \delta$ and $B > \delta$. The transpose of a vector, say z, is denoted by z'. The gradient and Hessian of a function $g(\phi)$ with respect to ϕ are denoted by $D_\phi g(\phi)$ and $D_\phi^2 g(\phi)$, respectively. All logarithms in a given discussion are taken to the same arbitrarily chosen base. The most common choices are the natural base e and the base 2.

B. Entropy

Consider a random process $\{X_t, t \geq 1\}$ with parametric distribution P_{ϕ^0}, $\phi^0 \in \Phi$. Let $p(x^n; \phi^0)$ be the induced n-dimensional density of the process with respect to μ^n where μ is some σ-finite measure. The sample entropy is defined for finite-alphabet processes. Suppose μ is the counting measure. Then, the *sample entropy* of $\{X_t\}$ is defined as $-n^{-1}\log p(X^n; \phi^0)$ [152, p. 58]. The relative entropy density is defined for processes with finite as well as continuous alphabet. Suppose μ is any σ-finite measure which could possibly be the Lebesgue measure. The *relative entropy density* of $\{X_t\}$ is defined as $\log p(X^n; \phi^0)$ [152, p. 150]. We shall use this term for $\log p(X^n; \phi)$ as well, where $\phi \in \Phi$ may be different from the true parameter ϕ^0. These quantities have well-defined limits for HMPs when $n \to \infty$. The limits and conditions for their existence are given in Section IV-D.

C. Martingale Difference Sequence

Let X be a random variable on the probability space $(\Omega, \mathcal{F}, P)$, and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. The conditional mean $E\{X|\mathcal{G}\}$ exists if $E\{|X|\} < \infty$ [154, p. 348]. Let $\boldsymbol{\mathcal{F}} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots\}$ denote a sequence of sub-σ-fields of $\mathcal{F}$. The sequence $\boldsymbol{\mathcal{F}}$ is called a *filtration* if $\mathcal{F}_t \subseteq F_{t+1}$ for all t. Let $\boldsymbol{X} = \{X_t, t \geq 1\}$ denote a random process on the probability space. The process is said to be *adapted* to the filtration $\boldsymbol{\mathcal{F}}$ if X_t is $\mathcal{F}_t$-measurable for all t [154, p. 473]. For example, if $\mathcal{F}_t = \sigma\{X^t\}$ denotes the smallest σ-field generated by X^t then $\boldsymbol{\mathcal{F}}$ is a filtration and $\boldsymbol{X}$ is adapted to $\boldsymbol{\mathcal{F}}$. Suppose $\boldsymbol{\mathcal{F}}$ is a filtration and $\boldsymbol{X}$ is adapted to $\boldsymbol{\mathcal{F}}$. The pair $(\boldsymbol{X}, \boldsymbol{\mathcal{F}}) = \{(X_t, \mathcal{F}_t), t \geq 1\}$ is called a *martingale* if for all $t \geq 1$, $E\{|X_t|\} < \infty$, and $E\{X_{t+1}|\mathcal{F}_t\} = X_t$ [154, p. 474]. Suppose $(\boldsymbol{X}, \boldsymbol{\mathcal{F}})$ is a martingale. The sequence $\boldsymbol{D} = \{D_t, t > 1\}$, where $D_t = X_t - X_{t-1}$, is a called a *martingale difference sequence*. In particular, D_t is $\mathcal{F}_t$-measurable, $E\{|D_t|\} < \infty$, and $E\{D_{t+1}|\mathcal{F}_t\} = 0$ for all t [154, p. 476]. The class of zero-mean independent processes is a subsest of the class of martingale difference sequences, and the class of martingale difference sequences is a subset of the class of zero-mean noncorrelated processes when second-order moments exist [288]. Under these conditions, a martingale difference sequence comprises noncorrelated random variables which may also be statistically independent. A martingale difference sequence enjoys a central limit theorem [38, Theorem 35.12].

D. Ergodicity and Asymptotically Mean Stationarity

Consider a random process with associated sequence probability space $(\mathcal{X}^{\mathcal{T}}, \mathcal{B}_{\mathcal{X}}^{\mathcal{T}}, Q)$. Assume that this is a one-sided process with index set $\mathcal{T} = \{1, 2, \ldots\}$. Let $x^{\mathcal{T}} = (x_1, x_2, \ldots)$ denote a member of $\mathcal{X}^{\mathcal{T}}$. Define the *left-shift* transformation T: $\mathcal{X}^{\mathcal{T}} \to \mathcal{X}^{\mathcal{T}}$ by $T((x_1, x_2, \ldots)) = (x_2, x_3, \ldots)$. The measure Q is called *stationary* if $Q(G) = Q(T^{-1}G)$ for all $G \in \mathcal{B}_{\mathcal{X}}^{\mathcal{T}}$ where $T^{-1}G = \{x^{\mathcal{T}}: T(x^{\mathcal{T}}) \in G\}$. Stationary measures correspond to stationary random processes [154, p. 398]. An event

$G \in \mathcal{B}_{\mathcal{X}}^{\mathcal{T}}$ is called *invariant* if $G = T^{-1}G$ or when $x^{\mathcal{T}} \in G$ if and only if $T(x^{\mathcal{T}}) \in G$. The stationary measure Q is called *ergodic* if each invariant event has probability either zero or one, i.e., $Q(G) = 0$ or $Q(G) = 1$ for all invariant events G [154, p. 398]. Define $T^n(x^{\mathcal{T}}) = (x_{n+1}, x_{n+2}, \ldots)$. The random process is called *asymptotically mean stationary* (AMS) with respect to the left-shift T if the limit of the Cesáro mean

$$\overline{Q}(G) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} Q\left(T^{-j}G\right) \tag{3.1}$$

exists for all $G \in \mathcal{B}_{\mathcal{X}}^{\mathcal{T}}$ [152, p. 16]. The limit $\overline{Q}$ is a stationary probability measure on $(\mathcal{X}^{\mathcal{T}}, \mathcal{B}_{\mathcal{X}}^{\mathcal{T}})$. It is called the *stationary mean* of Q [152, p. 16]. The stationary mean $\overline{Q}$ *asymptotically dominates* Q in the sense that $\overline{Q}(G) = 0$ implies $\lim_{n\to\infty} Q(T^{-n}G) = 0$ [151, Corollary 6.3.2]. Conversely, if Q is asymptotically dominated by a stationary measure then Q is AMS [148, Theorem 2]. These properties demonstrate intuitive aspects of AMS processes gained by considering events determinable by samples of the process in the distant future. Asymptotic mean stationarity is necessary and sufficient for an ergodic theorem to hold [151, Corollary 7.2.2].

Note that the left-shift transformation for one-sided processes is not invertible. Some of the results discussed in this paper were derived for two-sided processes. For that case, an invertible (one-to-one) shift transformation can be defined.

E. Mixing

A process $\boldsymbol{X} = \{X_t, t \geq 1\}$ with distribution P_X is said to be *α-mixing* if for every set $F \in \sigma(X_1, \ldots, X_k)$ and set $G \in \sigma(X_{k+n}, X_{k+n+1}, \ldots)$, $n \geq 1$, $k \geq 1$

$$|P_X(F \cap G) - P_X(F)P_X(G)| \leq \alpha(n) \tag{3.2}$$

where $\alpha(n)$ is independent of F and G and $\lim_{n\to\infty} \alpha(n) = 0$ [38, p. 363]. Thus, X_k and X_{k+n} are approximately independent for large n. The process is said to be *φ-mixing* if

$$\sup_{G, F} |P_X(G|F) - P_X(G)| \leq \varphi(n) \tag{3.3}$$

where $\varphi(n)$ is independent of F and G and $\lim_{n\to\infty} \varphi(n) = 0$ [37, p. 166]. This is a nonsymmetric measure of approximate independence.

F. Channels

A channel is defined as follows [152, Sec. 9.2]. Consider an input probability space $(\mathcal{X}^{\mathcal{T}}, \mathcal{B}_{\mathcal{X}}^{\mathcal{T}}, \eta)$ and an output measurable space $(\mathcal{Y}^{\mathcal{T}}, \mathcal{B}_{\mathcal{Y}}^{\mathcal{T}})$. Assume that the two measurable spaces are standard. A *channel* is a family of probability measures $\{\nu(\cdot|\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}^{\mathcal{T}}\}$ on $(\mathcal{Y}^{\mathcal{T}}, \mathcal{B}_{\mathcal{Y}}^{\mathcal{T}})$ such that for every output event $F \in \mathcal{B}_{\mathcal{Y}}^{\mathcal{T}}$, $\nu(F|\boldsymbol{x})$ is a measurable function of $\boldsymbol{x}$. For every rectangle $G \times F \in \mathcal{B}_{\mathcal{X}}^{\mathcal{T}} \times \mathcal{B}_{\mathcal{Y}}^{\mathcal{T}}$, the set function

$$P(G \times F) = \int_G \nu(F|\boldsymbol{x})\, d\eta(\boldsymbol{x}) \tag{3.4}$$

is well defined, and it extends to a probability measure on the joint input/output space which is sometimes called the *hookup* of the source η and channel ν. Thus, a channel is simply a regular conditional probability [152, p. 5].

Let $T_{\mathcal{X}}$ and $T_{\mathcal{Y}}$ be the shift transformations on the input sequence space $\mathcal{X}$ and output sequence space $\mathcal{Y}$, respectively. A channel is said to be *stationary* with respect to $T_{\mathcal{X}}$ and $T_{\mathcal{Y}}$, or simply stationary if the shifts are clear from the context, if [152, p. 184]

$$\nu\left(T_{\mathcal{Y}}^{-1}F|\boldsymbol{x}\right) = \nu(F|T_{\mathcal{X}}\,\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathcal{X}^{\mathcal{T}},\ F \in \mathcal{B}_{\mathcal{Y}}^{\mathcal{T}}. \tag{3.5}$$

Intuitively, a right-shift of an output event yields the same probability as a left-shift of an input event. Two shifts are required since in general $T_{\mathcal{X}}^{-1}\boldsymbol{x}$ and $T_{\mathcal{Y}}F$ may not exist. If the shifts are invertible, as for two-sided processes, then the definition is equivalent to

$$\nu(T_{\mathcal{Y}}F|T_{\mathcal{X}}\boldsymbol{x}) = \nu\left(T_{\mathcal{Y}}^{-1}F|T_{\mathcal{X}}^{-1}\,\boldsymbol{x}\right) = \nu(F|\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathcal{X}^{\mathcal{T}},\ F \in \mathcal{B}_{\mathcal{Y}}^{\mathcal{T}}. \tag{3.6}$$

Thus, shifting the input sequence and output sequence in the same direction does not change the probability. In that case, a single shift may be used for both input and output sequences.

A channel is said to be *output strongly mixing*, or asymptotically output memoryless, if for all output rectangles F and G and all input sequences $\boldsymbol{x}$ [152, p. 196]

$$\lim_{n\to\infty} \left|\nu\left(T^{-n}F \cap G\,|\,\boldsymbol{x}\right) - \nu\left(T^{-n}F\,|\,\boldsymbol{x}\right)\nu(G\,|\,\boldsymbol{x})\right| = 0. \tag{3.7}$$

More generally, the channel is said to be *output weakly mixing* if

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1}\left|\nu\left(T^{-i}F \cap G\,|\,\boldsymbol{x}\right) - \nu\left(T^{-i}F\,|\,\boldsymbol{x}\right)\nu(G\,|\,\boldsymbol{x})\right| = 0. \tag{3.8}$$

Of particular interest for our discussion are memoryless invariant channels. Suppose that $\upsilon(\cdot|x)$ is a probability measure on $\mathcal{B}_{\mathcal{Y}}$ for all $x \in \mathcal{X}$ and that $\upsilon(F|x)$ is a measurable function of x for fixed F. Let $\{F_i\}$ denote a sequence of output events. The channel ν is said to be *memoryless* if

$$\nu\left(\prod_{i\in\mathcal{I}} F_i|\boldsymbol{x}\right) = \prod_{i\in\mathcal{I}} \upsilon(F_i|x_i) \tag{3.9}$$

for any finite index set $\mathcal{I} \subset \mathcal{T}$ [152, p. 193]. The channel ν is said to be *invariant* if $\upsilon(F_i|x_i)$ is independent of i. When densities exist, the channel is defined by its transition density or by the n-dimensional conditional density $p(y^n|x^n)$ for all finite n. Memoryless invariant channels satisfy

$$p(y^n|x^n) = \prod_{t=1}^{n} p(y_t|x_t)$$

and $p(y_t|x_t)$ is time-invariant, i.e., for any $F \in \mathcal{B}_{\mathcal{Y}}$ and $x \in \mathcal{X}$, the probability of $Y_t \in F$ given $X_t = x$ is the same for all t.

IV. Statistical Properties

In this section, we define HMPs and discuss their relations to mixture processes, switching autoregressive processes, dynamical systems, Markov-modulated Poisson processes, com-

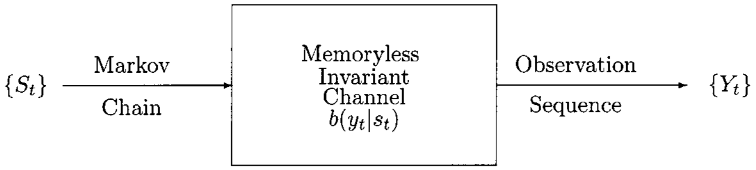

Fig. 1. An HMP.

posite sources, deterministic functions of Markov chains, and unifilar sources. We state conditions for HMPs to be stationary, ergodic, and mixing processes. We provide ergodic theorems for almost-sure convergence of the sample entropy and relative entropy densities of stationary-ergodic general HMPs. Similar ergodic theorems for switching autoregressive processes with finite and continuous state spaces are also reviewed.

A. Definitions and Structure

Let $\{S_1, S_2, \ldots\}$ denote a discrete-time Markov chain that takes values in a finite set $\mathcal{S}$ called the *state space*. Let M denote the number of states. We assume without loss of generality that $\mathcal{S} = \{1, 2, \ldots, M\}$. Let $s_t \in \mathcal{S}$ denote a value that S_t can take. Let $\pi_j = P(S_1 = j)$ denote the probability that the initial state is j. Let $\pi = \{\pi_j\}$ be a $1 \times M$ vector representing the *initial distribution*. The Markov chain is always assumed homogeneous unless stated otherwise. Let $a_{ij} = P(S_t = j | S_{t-1} = i)$ denote the *transition probability*. Let $A = \{a_{ij}\}$ denote the $M \times M$ *transition matrix*. Consider a discrete-time channel with input $\{S_1, S_2, \ldots\}$ and output $\{Y_1, Y_2, \ldots\}$. For each t, Y_t takes values in an *observation space* $\mathcal{Y}$. The nature of $\mathcal{Y}$ will be discussed shortly. Let $y_t \in \mathcal{Y}$ denote a value that Y_t can take. Assume that the channel is memoryless and invariant. For a given s_t, let $b(y_t|s_t)$, $y_t \in \mathcal{Y}$, denote a transition density of the channel with respect to some σ-finite measure μ. Of particular interest are the Lebesgue and counting measures. The counting measure is denoted by κ. The channel is characterized by a set of M transition densities $\{b(\cdot|s_t),\, s_t = 1, \ldots, M\}$. We shall refer to $b(\cdot|s_t)$ as an *observation conditional density* [210].

In information theory, an HMP is viewed as a discrete-time finite-state homogeneous Markov chain observed through a discrete-time memoryless invariant channel as described in Fig. 1. In the statistical literature, see, e.g., [36], an HMP is viewed as a discrete-time bivariate random process $\{(S_t, Y_t)\}$ with Markov *regime* $\{S_t\}$ and conditionally independent random variables $\{Y_t\}$. The distribution of Y_t is time-invariant and it depends on $\{S_t\}$ only through S_t.

The n-dimensional density of (Y^n, S^n) with respect to $\mu^n \times \kappa^n$ can be written as

$$p(y^n, s^n) = p(y_1, s_1) \prod_{t=2}^{n} p(y_t, s_t | s_{t-1}) \tag{4.1}$$

where

$$\begin{aligned} p(y_1, s_1) &= \pi_{s_1} b(y_1|s_1) \\ p(y_t, s_t|s_{t-1}) &= a_{s_{t-1}s_t} b(y_t|s_t), \qquad t = 2, 3, \ldots. \end{aligned} \tag{4.2}$$

The convention $a_{s_0 s_1} = \pi_{s_1}$ for all $s_1 \in \mathcal{S}$ is often convenient. The n-dimensional density of Y^n with respect to μ^n is given by

$$p(y^n) = \sum_{s^n} \prod_{t=1}^{n} a_{s_{t-1}s_t} b(y_t|s_t). \tag{4.3}$$

This function is often referred to as the *likelihood function* of the HMP. Note that the saummation in (4.3) is over M^n product terms.

The likelihood function may also be expressed in an alternative useful form in terms of $b(y_t|s_t)$ and $p(s_t|y^{t-1})$. It is easy to check, see, e.g., Ott [248], Lindgren [219], and Devijver [81], that

$$\begin{aligned} p(y^n) &= p(y_1) \prod_{t=2}^{n} p(y_t|y^{t-1}) \\ p(y_1) &= \sum_{s_1=1}^{M} \pi_{s_1} b(y_1|s_1) \\ p(y_t|y^{t-1}) &= \sum_{s_t=1}^{M} p(s_t|y^{t-1}) b(y_t|s_t). \end{aligned} \tag{4.4}$$

We refer to $p(s_t|y^{t-1})$ as the *predictive density* of S_t given y^{t-1} [90]. Thus, properties of the likelihood function of the HMP are determined by the predictive density sequence and by the observation conditional densities. These properties will be discussed in Section IV-C3. A computationally efficient recursion for calculating $p(s_t|y^{t-1})$ is provided in (4.30).

It follows from (4.4) that each observation of the HMP has a mixture density

$$p(y_t) = \sum_{s_t=1}^{M} p(s_t) b(y_t|s_t).$$

If the Markov chain is stationary, then $P(S_t = j) = \pi_j$ for all t, and the observations $\{Y_t\}$ are identically distributed with mixture density given by

$$p(y_t) = \sum_{j=1}^{M} \pi_j b(y_t|S_t = j). \tag{4.5}$$

Conditions for stationarity of the Markov chain are given in Section IV-C. The observations $\{Y_t\}$ are generally dependent but they may also be independent. For example, let $\{X_t\}$ denote a sequence of independent and identically distributed (i.i.d.) random variables and define a Markov chain $\{S_t\}$ by $S_t = (X_{t-1}, X_t)$. Let $Y_t = g(S_t) = X_t$ for some deterministic function $g(\cdot, x) = x$. The sequence $\{Y_t\}$ is an HMP with i.i.d. observations. A stationary HMP is thus a sequence of possibly dependent identically distributed random variables with a marginal mixture density. Each mixture density is overdispersed relative to any given single density $b(\cdot|s_t)$. Leroux [212] referred to HMPs as *mixture processes with Markov dependence*.

It also follows from (4.4) that if the density $b(y_t|s_t)$ has zero mean for all $s_t \in \mathcal{S}$, then almost surely

$$E\{Y_t|Y^{t-1}\} = 0, \qquad t > 1. \tag{4.6}$$

Under this condition, $\{Y_t\}$ is a martingale difference sequence, as pointed out by Francq and Roussignol [126]. As such, an HMP is a sequence of noncorrelated random variables that may also be statistically independent. This implies that the observations $\{Y_\tau, \tau < t\}$ of an HMP are not useful in predicting Y_t in the minimum mean square error (MMSE) sense [288].

If the regime $\{S_t\}$ of an HMP is i.i.d. instead of Markov, the observations $\{Y_t\}$ are necessarily i.i.d. From (4.3) we have

$$\begin{aligned} p(y^n) &= \sum_{s^n} \prod_{t=1}^{n} p(s_t) b(y_t|s_t) \\ &= \prod_{t=1}^{n} \sum_{j=1}^{M} P(S_t = j) b(y_t|S_t = j). \end{aligned} \tag{4.7}$$

An HMP with i.i.d. regime is a *mixture process*. Mixture processes have been extensively studied and a wealth of results is available, see, e.g., [109], [266], [301], [232]. The close relation between HMPs and mixture processes is often exploited in proving properties of HMPs using similar properties of mixture processes.

When the observation space $\mathcal{Y}$ is finite, the HMP is referred to as a *finite-alphabet* HMP. When $\mathcal{Y}$ is not necessarily finite, the HMP is referred to as a *general* HMP [36]. For a finite-alphabet HMP, we assume without loss of generality that $\mathcal{Y} = \{1, 2, \dots, L\}$. Let

$$b_{jl} = P(Y_t = l|S_t = j)$$

denote the time-invariant state-to-observation transition probability. Let $B = \{b_{jl}\}$ denote the $M \times L$ state-to-observation transition matrix. The parameter of the channel is denoted by $\theta = B$. For a general HMP, $\mathcal{Y}$ is usually a subset of a Euclidean space $\mathcal{R}^k$ for some k. Other higher dimensional spaces are also possible. The parameter of the observation conditional density for state j is denoted by $\theta_j \in \Theta \subseteq \mathcal{R}^{d_1}$ for some d_1. The parameter of the channel is denoted by $\theta = \{\theta_j\}$. We shall sometimes emphasize the dependency of the observation conditional density on its parameter. We may write $b(y_t|s_t; \theta)$ or use the more customary notation of $b(y_t; \theta_{s_t})$.

The parameter of the HMP is given by (π, A, θ). For a stationary Markov chain with a unique stationary distribution $\pi = \pi A$ the parameter of the HMP is simply (A, θ). Conditions for uniqueness of a stationary distribution are given in Section IV-C1. In some applications, the triplet (π, A, θ) depends on a parameter ϕ in some parameter set Φ and we have the parametrization $(\pi(\phi), A(\phi), \theta(\phi))$. The parameter $\phi = (\pi, A, \theta)$ is a particular case obtained using coordinate projections, i.e., $\pi(\phi) = \pi$, $A(\phi) = A$, and $\theta(\phi) = \theta$. This is the most common parametrization of the HMP which is referred to as the *usual parametrization*. Throughout this paper, ϕ is referred to as the *parameter* of the HMP where in general it need not represent the usual parametrization.

We shall sometimes emphasize the dependency of the n-dimensional density of the HMP on its parameter by rewriting (4.3) as

$$p(y^n; \phi) = \sum_{s^n} \prod_{t=1}^{n} a_{s_{t-1}s_t}(\phi) b(y_t; \theta_{s_t}(\phi)). \tag{4.8}$$

In some discussions, such as in ML parameter estimation, we must distinguish between the true parameter that was used to produce a given sequence of observations, say y^n, and any other value ϕ of the parameter of the HMP. We denote the true parameter by ϕ^0. For the usual parametrization, $\phi^0 = (\pi^0, A^0, \theta^0)$. A stationary HMP is said to be *identifiable* if for each $\phi \in \Phi$ such that $\phi \neq \phi^0$, $p(y^n; \phi) \neq p(y^n; \phi^0)$ a.e. for some $n > 0$ [274]. Note that states may always be permuted without affecting the distribution of the HMP. This ambiguity can be removed by ordering the states, for example, according to their $\{\theta_j\}$.

Two parameters $\phi^{(1)}$ and $\phi^{(2)}$ in Φ are said to be *equivalent* if they induce the same stationary law for $\{Y_t\}$. We denote this relation by $\phi^{(1)} \sim \phi^{(2)}$. The parameter set Φ can be partitioned into the *equivalence classes* of $\sim$. The equivalence class of a parameter ϕ of an identifiable HMP comprises all points in Φ obtained by permutations of the states of the HMP [214, Lemma 2].

In some applications such as modeling of speech signals [173], and representing Markov modulated Poisson processes as HMPs [273], the assumption (4.2) is replaced by

$$p(y_t, s_t|s_{t-1}) = p(s_t|s_{t-1}) p(y_t|s_t, s_{t-1}), \qquad t = 2, 3, \dots. \tag{4.9}$$

Since pairs of states in (4.9) may be renamed as new states in (4.2), the two assumptions are equivalent [36], [173]. Finite-alphabet HMPs that obey (4.9) were referred to as *finite-state sources* in [236], [325], [330].

There are many extensions of the HMP as defined in this section. Some of them will be discussed in the next subsection. Throughout this paper, we refer to the discrete-time finite-state process with finite or general alphabet defined by (4.1) and (4.2) as an HMP or even more specifically as a *standard HMP*. This is not to be confused with an HMP that has standard alphabet. Other forms of HMPs such as HMPs with a countably infinite state space, a continuous state space, or continuous-time HMPs will be specifically noted.

B. Examples

HMPs appear in many forms. In this subsection we provide some examples to demonstrate the scope of this rich family of processes.

1) Gaussian Mixture Processes With Markov Dependence: HMPs with multivariate Gaussian observation conditional densities are commonly used in automatic speech recognition applications. Gaussian densities are suitable when modeling is applied to representations of the signal for which a central limit theorem holds [47]. This indeed is the case in automatic speech recognition applications where modeling is applied to vectors of spectral or cepstral components of the signal [19], [262], [106]. Let k denote the dimension of each vector. Parametrization of each $k \times k$ covariance matrix as a matrix of an autoregressive process of order $r \ll k$ [149] was studied in [254], [255], [175]. HMPs with zero-mean Gaussian observation conditional densities also appear in the form of $Y_t = S_t W_t$, where $\{S_t\}$ is a Markov chain that takes values $\{\sigma_1, \ldots, \sigma_M\}$, $\{W_t\}$ is a sequence of i.i.d. standard Gaussian random variables, and $\{S_t\}$ and $\{W_t\}$ are statistically independent. This model was thoroughly studied by Francq and Roussignol [126]. Another popular form of HMPs with Gaussian observation conditional densities is given by $Y_t = S_t + W_t$, where $\{S_t\}$ is a Markov chain that takes values $\{\varsigma_1, \ldots, \varsigma_M\}$, $\{W_t\}$ is a sequence of zero mean i.i.d. Gaussian random variables with variance σ_w^2, and $\{S_t\}$ and $\{W_t\}$ are statistically independent [59], [194], [197].

2) Poisson Mixture Processes With Markov Dependence: HMPs with Poisson observation conditional densities are used for modeling counting processes. Here Y_t given $S_t = j$ is a Poisson random variable with rate λ_j. Such HMPs are often encountered in biomedical applications, for example, in monitoring epileptic seizure counts [6], [207].

3) Switching Processes With Markov Regime: A switching process with Markov regime is a random process whose dynamics at any given time depend on the state of a Markov chain at that time. Examples include *switching regression* [219] and *switching autoregressive* processes [156, Ch. 22], [164]. In this subsection, we focus on switching autoregressive processes only. Let $\{Y_t\}$ denote the process and let $\{S_t\}$ denote its Markov regime of M states. Consider first a switching autoregressive process that is linear in its parameter when the state sequence $\{s_t\}$ is given. Assume that all states have the same autoregressive order r. Let $\theta_j = \{\sigma_j, c_j(i), i = 1, \ldots, r\}$ denote the autoregressive parameter for state j, where σ_j denotes the gain and $\{c_j(i), i = 1, \ldots, r\}$ are the autoregressive coefficients. Let $\{W_j(t)\}$ denote the i.i.d. sequence of innovations when the Markov chain is in state j. It is assumed that $\{\{S_t\}, \{W_j(t)\}, j = 1, \ldots, M\}$ are mutually statistically independent. Assuming that the Markov chain is in state s_t at time t, then the process can be described by the difference equation

$$Y_t = -\sum_{i=1}^{r} c_{s_t}(i) Y_{t-i} + \sigma_{s_t} W_{s_t}(t), \qquad t = 1, 2, \ldots. \tag{4.10}$$

The conditional n-dimensional density of Y^n is given by

$$p(y^n | y_{-r+1}^0; \phi) = \sum_{s^n} p(y^n, s^n | y_{-r+1}^0; \phi) \tag{4.11}$$

$$p(y^n, s^n | y_{-r+1}^0; \phi) = \prod_{t=1}^{n} a_{s_{t-1}s_t} b(y_t | y_{t-r}^{t-1}; \theta_{s_t}) \tag{4.12}$$

where y_{-r+1}^0 is a realization of a vector of initial conditions which is assumed independent of $\{S_t\}$, the density $b(y_t|y_{t-r}^{t-1}; \theta_{s_t})$ is determined by the distribution of $W_{s_t}(t)$, and ϕ is the parameter of the process. Let $Z_t = (Y_t, Y_{t-1}, \ldots, Y_{t-r+1})'$, $V_{s_t} = (W_{s_t}(t), 0, \ldots, 0)'$, and C_{s_t} denote the $r \times r$ companion matrix of the autoregression associated with state $S_t = s_t$. The process $\{Y_t\}$ has the following state-space representation, see, e.g., [258, p. 797], [240]:

$$Z_t = C_{s_t} Z_{t-1} + \sigma_{s_t} V_{s_t}. \tag{4.13}$$

The switching autoregressive process (4.10) is not guaranteed to be second-order stationary even if each individual autoregressive process is stable. Conversely, the switching autoregressive processes (4.10) may be second-order stationary even if some individual autoregressive processes are not [164]. A sufficient condition for the switching autoregressive process (4.10) to be second-order stationary was given by Holst, Lindgren, Holst, and Thuvesholmen [164]. Assume that for each $i = 1, \ldots, M$ the innovation process $\{W_i(t)\}$ has zero mean and unit variance. Let a_{ji} denote the transition probability from state j to state i. For each $i, j = 1, \ldots, M$, define the $r^2 \times r^2$ matrix $F_{ij} = (C_i \otimes C_i) a_{ji}$ where $\otimes$ denotes the Kronecker product. Let $F = \{F_{ij}\}$ denote the resulting $r^2 M \times r^2 M$ matrix and denote by $\rho(F)$ its spectral radius. The switching autoregressive process (4.10) is second-order stationary if $\rho(F) < 1$.

A more general form of the switching autoregressive process with Markov regime (4.13) was studied by Francq and Roussignol [127]. The process is defined by

$$Z_t = g(Z_{t-1}, S_t; \phi) + h(V_t, S_t; \phi), \qquad t = 1, 2, \ldots \tag{4.14}$$

where $\{Z_t\}$ is a sequence of r-dimensional random vectors, $\{S_t\}$ is a finite-state Markov chain, $\{V_t\}$ is a sequence of i.i.d. k-dimensional random vectors independent of $\{S_t\}$, and $g(\cdot, \cdot, \cdot)$ and $h(\cdot, \cdot, \cdot)$ are measurable functions from $\mathcal{R}^r \times \mathcal{S} \times \Phi$ to $\mathcal{R}^r$ and from $\mathcal{R}^k \times \mathcal{S} \times \Phi$ to $\mathcal{R}^r$, respectively. In general, the driving i.i.d. noise $\{V_t\}$ need not be Gaussian. The standard HMP (4.3) is a special case of (4.14) which corresponds to $g(\cdot, \cdot, \cdot) \equiv 0$. Krishnamurthy and Rydén [198] studied an even more general class of switching autoregressive process characterized by

$$Y_t = g\left(Y_{t-r}^{t-1}, S_t, W_t; \phi\right), \qquad t = 1, 2, \ldots \tag{4.15}$$

where $\{Y_t\}$ is a scalar process, g is an arbitrary measurable function, and $\{W_t\}$ is a scalar i.i.d. process. The conditional n-dimensional densities of (4.14) and (4.15) may be written similarly to (4.11) and (4.12). The scalar case was chosen in [198] for notational convenience only. Douc, Moulines, and Rydén [91] studied general forms of switching autoregressive processes, of which the above functional forms are special cases, when the Markov chain $\{S_t\}$ takes values in a separable compact state space that is not necessarily finite. For example, the state space $\mathcal{S}$ may be a compact set in a Euclidean space. Sufficient conditions for the existence of a stationary ergodic solution for the difference equation (4.14) will be detailed in Section IV-C4. Ergodic theorems for switching autoregressive processes will be presented in Section IV-D. Theorems for asymptotic optimality of their ML parameter estimators will be given in Section VI-B.

The switching autoregressive process (4.13) is a special case of the r-dimensional vector process $\{Z_t\}$ defined by the difference equation

$$Z_t = G_t Z_{t-1} + H_t \tag{4.16}$$

where $\{G_t\}$ and $\{H_t\}$ are sequences of random matrices. When $\{G_t\}$ and $\{H_t\}$ are statistically independent sequences of i.i.d. random matrices, the Markov process $\{Z_t\}$ is referred to as *random coefficient autoregressive* (RCA) process. Conditions for stationarity and second-order stationarity as well as algorithms for parameter estimation of RCA processes were given by Nicholls and Quinn [245]. Conditions for geometric ergodicity and existence of moments of RCA processes were provided by Feigin and Tweedie [114]. The important concept of geometric ergodicity will be defined in (4.28) for finite-state Markov chains. For more general cases see Meyn and Tweedie [240]. Conditions for existence of moments of a scalar process $\{Z_t\}$ satisfying (4.16), when $\{(G_t, H_t)\}$ is a stationary sequence of random variables, were given by Karlsen [181]. A sufficient condition for existence of a unique stationarity solution $\{Z_t\}$ of (4.16), when $\{(G_t, H_t)\}$ is a stationary ergodic sequence of random matrices, was given by Brandt [46] and by Bougerol and Picard [45]. The condition was shown to be necessary in [45].

Note that the switching autoregressive process (4.13) differs from the HMP with autoregressive observation conditional densities of the example in Section IV-B1. In that example, observations are conditionally independent given the state sequence. Applications of switching autoregressive processes of the form (4.10) in econometrics were studied by Hamilton [156]. See also Krolzig [202] and the references therein. First-order switching autoregressive processes of the form (4.10) were used in automatic speech recognition applications by Wellekens [312] and in speech enhancement applications by Ephraim [103].

4) Communication Channels Driven by HMPs: Consider a communication channel with input $\{X_1, X_2, \ldots\}$, output $\{Y_1, Y_2, \ldots\}$, and a state sequence $\{C_0, C_1, \ldots\}$. Assume that for each t, X_t takes values in an input space $\mathcal{X}$, Y_t takes values in an output space $\mathcal{Y}$, and C_t takes values in a state space $\mathcal{C}$. The channel is called a *finite-state channel* (FSC) if the following conditions are met [133, Sec. 4.6]. i) $\mathcal{C}$ is finite. ii) The state sequence $\{C_t\}$ is a Markov chain given $\{X_t\}$, and the distribution of C_t depends on $\{X_t\}$ only through X_t. iii) The observations $\{Y_t\}$ are conditionally independent given $\{(X_t, C_t)\}$, and the distribution of Y_t depends on $\{(X_t, C_t)\}$ only through (X_t, C_t, C_{t-1}). An FSC is characterized by the time-invariant transition density $p(y_t, c_t|x_t, c_{t-1})$ and by the initial state c_0. The conditional n-dimensional transition density of the channel is given by

$$p(y^n|x^n, c_0) = \sum_{c^n} p(y^n, c^n|x^n, c_0) \qquad (4.17)$$

where

$$p(y^n, c^n|x^n, c_0) = \prod_{t=1}^{n} p(y_t, c_t|c_{t-1}, x_t). \qquad (4.18)$$

Equation (4.18) is an example of a Markov channel [186], [150]. FSCs play an important role in information theory, see [133], [205]. Properties and universal decoding of FSCs will be discussed in Section XIII. It is easy to check that if $\{(X_t, S_t)\}$ is an HMP with state space $\mathcal{S}$, then $\{Y_t, (C_t, S_t)\}$ is an HMP with an augmented state space $\mathcal{C} \times \mathcal{S}$. A special case of this example is an HMP observed through a memoryless invariant channel [17]. Note also that an FSC with a degenerate input sequence of $x^n = (x, x, \ldots, x)$, $x \in \mathcal{X}$, is an HMP.

5) The Gilbert–Elliott Channel: The Gilbert–Elliott channel is a special FSC [137], [97], [14], [243], [204]. For this example, $\mathcal{C}$, $\mathcal{X}$, and $\mathcal{Y}$ are binary. In addition

$$p(y_t|c_t, c_{t-1}, x_t) = p(y_t|c_t, x_t)$$

and

$$p(c_t|c_{t-1}, x_t) = p(c_t|c_{t-1}).$$

The channel introduces an additive two-state hidden Markov noise process $\{Z_t\}$ that is statistically independent of the input process $\{X_t\}$. For each t, $Y_t = X_t \oplus Z_t$ where $\oplus$ denotes modulo-two addition. The two states of the channel represent low and high error conditions. The channel is particularly suitable for modeling communications under fading conditions characterized by irregular patterns of burst errors. Properties of the Gilbert–Elliott channel depend on its memory length characterized by the parameter $v = 1 - (a_{12} + a_{21})$ which satisfies $|v| \leq 1$ [243]. When $v = 0$, the Markov regime $\{C_t\}$ becomes an i.i.d. regime and the channel is memoryless. When $v = 1$, the Markov chain is reducible and the state of the channel is determined by its initial distribution. This is a degenerate channel whose underlying state can be inferred from the observed sequence $\{Y_t\}$. When $v = -1$, the chain is periodic and the states constantly alternate. Additional properties of this channel are given in Section XIII.

6) Dynamical Systems: HMPs have dynamical system representations in the sense of control theory. A dynamical system representation for discrete-time point processes was first given by Segall [289]. These processes are briefly described in Section IV-B7. A dynamical system representation of an HMP was developed by Hamilton [156, Ch. 22]. Elliott, Aggoun, and Moore [99] applied this representation to a range of general HMPs. In this example, we demonstrate the approach for a finite-alphabet HMP with M states and L letters. We will revisit this representation in Section IX which is dedicated to the dynamical system approach to HMPs. Our presentation follows [99, Sec.2.2].

Let e_m denote a unit vector representing the mth state of the HMP in an M-dimensional Euclidean space $\mathcal{R}^M$. The mth component of e_m is one while all other components are zero. The state space of the HMP is given by $\{e_1, \ldots, e_M\}$. Similarly, let f_l denote a unit vector in $\mathcal{R}^L$ representing the lth letter from the alphabet of the HMP. The observation space of the HMP is given by $\{f_1, \ldots, f_L\}$. Let $a_{ij} = P(S_{t+1} = e_j|S_t = e_i)$ denote the state transition probability and let $A = \{a_{ij}\}$. Let $b_{jl} = P(Y_{t+1} = f_l|S_t = e_j)$ denote the state-to-observation transition probability and let $B = \{b_{jl}\}$. The unit delay between the state and output variables indicates a noninstantaneous response of the system to S_t. Let $\mathcal{F}_t = \sigma(S_0^t)$ denote the smallest σ-field generated by the random variables S_0^t. Let $\mathcal{G}_t = \sigma(S_0^t, Y^t)$ denote the smallest σ-field generated by the random variables $\{S_0^t, Y^t\}$. Note that $E\{S_{t+1}|\mathcal{F}_t\} = A'S_t$ and $E\{Y_{t+1}|\mathcal{G}_t\} = B'S_t$. Define $V_{t+1} = S_{t+1} - A'S_t$ and note that $E\{V_{t+1}|\mathcal{F}_t\} = 0$. Similarly, define $W_{t+1} = Y_{t+1} - B'S_t$ and note that $E\{W_{t+1}|\mathcal{G}_t\} = 0$. The HMP can now be written as a dynamical system that has the same probability law. This system is given by

$$\begin{aligned} S_{t+1} &= A' S_t + V_{t+1} \\ Y_{t+1} &= B' S_t + W_{t+1}, \qquad t = 0, 1, \ldots. \end{aligned} \tag{4.19}$$

The martingale difference processes $\{V_t\}$ and $\{W_t\}$ may be statistically dependent as in [289]. They are statistically independent for the HMP defined in Section IV-A.

7) Markov-Modulated Poisson Processes (MMPPs): Consider a Poisson process whose rate is controlled by a nonobservable continuous-time finite-state homogeneous Markov chain. Such process is called a *Markov-modulated Poisson process*. Markov-modulated Poisson processes have many applications in medicine [295, Ch. 7], computer networks [158], [159], and queueing theory [117]. A survey of this class of processes can be found in Fischer and Meier-Hellstern [117]. Some properties of Markov-modulated Poisson processes and their relation to HMPs are discussed in this subsection. Our presentation follows Rydén [273]. Additional results will be given in Sections IV-D, VI-A–VI-C, and X.

Let $\{S(t), t \geq 0\}$ be the continuous-time Markov chain with state space $\mathcal{S} = \{1, 2, \ldots, M\}$. Let

$$p_{ij}^{(t)} = P(S(\tau + t) = j | S(\tau) = i), \qquad t \geq 0$$

denote the transition probability from state i to state j in t seconds. Assume that for any pair i, j of states, $p_{ij}^{(t)} > 0$ for some $t > 0$. This implies that $p_{ij}^{(t)} > 0$ for all $t > 0$ [154, p. 260]. A Markov chain with this property is called *irreducible*. Let $P_t = \{p_{ij}^{(t)}\}$ and assume that the entries of P_t are continuous functions of t. This assumption is equivalent to $P_t \rightarrow I$ as $t \downarrow 0$ where I denotes the identity matrix [154, p. 257]. The transition probability $p_{ij}^{(t)}$ is approximately linear in t for sufficiently small t. There exist constants $\{g_{ij}\}, i, j \in \mathcal{S}$, such that

$$p_{ij}^{(t)} \simeq \begin{cases} 1 + g_{ii} t, & i = j \\ g_{ij} t, & i \neq j \end{cases} \tag{4.20}$$

where $g_{ij} \geq 0$ for $i \neq j$, and $g_{ii} \leq 0$ and $\sum_j g_{ij} = 0$ for all i. The matrix $G = \{g_{ij}\}$ is called the *generator* of the chain [154, p. 256]. The matrix P_t satisfies Kolmogorov's forward and backward equations, $\partial P_t / \partial t = P_t G$ and $\partial P_t / \partial t = G P_t$, respectively, where $P_0 = I$. These equations often have a unique solution given by $P_t = \exp(tG)$ [154, p. 259]. Next, let $\{N(t), t \geq 0\}$ denote the Markov-modulated Poisson process. Let λ_i denote the rate of the process when the chain is in state i. Assume that at least one $\lambda_i > 0$. Let Λ denote a diagonal matrix of rates $\{\lambda_i\}$. Let $\phi = (G, \Lambda)$ denote the parameter of the Markov-modulated Poisson process satisfying the above conditions.

The process $\{N(t)\}$ may be regarded as a *Markov renewal process*. To see this, let S_m denote the state of the continuous-time chain at the time of the mth Poisson event. Introduce an initial state S_0 with distribution π. Define Y_1 as the time until the first event, and let $Y_m, m \geq 2$, denote the time between event $(m-1)$ and event m. It follows that

$$\begin{aligned} P(Y_m \leq y, S_m = j | S_{m-1} = i, y^{m-1}, s_0^{m-2}) \\ = P(Y_m \leq y, S_m = j | S_{m-1} = i). \end{aligned} \tag{4.21}$$

Note that there is a one-to-one correspondence between $\{Y_m, m \geq 1\}$ and $\{N(t), t \geq 0\}$. Also, $\{S_m\}$ is a discrete-time Markov chain, and $\{Y_m\}$ is a sequence of conditionally independent random variables given $\{S_m\}$. The distribution of Y_m depends on $\{S_m\}$ only through S_{m-1} and S_m. This suggests that the Markov-modulated Poisson process may also be viewed as an HMP with Markov chain $\{S_m\}$ and observations $\{Y_m\}$. The density of this HMP is given by (4.1) and (4.9). The formulations of the Markov-modulated Poisson process as a Markov renewal process and as an HMP are similar but there is a subtle conceptual difference. For a Markov renewal process, the discrete-time Markov chain and the observations evolve sequentially in time, i.e., S_0 is first chosen according to the initial distribution π, then S_1 and Y_1 are chosen according to (4.21), and so on. For an HMP, the entire Markov chain first evolves and only then the observations follow [273].

Let $F_{ij}(y) = P(Y_m \leq y, S_m = j | S_{m-1} = i)$ and $F(y) = \{F_{ij}(y)\}$. The *transition density matrix* which corresponds to $F(y)$ is given by [130], [117]

$$f(y; \phi) = \frac{\partial F(y)}{\partial y} = \exp\{(G - \Lambda) y\} \Lambda. \tag{4.22}$$

The transition matrix of $\{S_m\}$ is given by $A = F(\infty)$. Integrating (4.22) with respect to y over $[0, \infty)$ gives

$$A = (\Lambda - G)^{-1} \Lambda. \tag{4.23}$$

The likelihood function of an observation sequence y^n is given by

$$p(y^n; \phi) = \pi(\phi) \prod_{l=1}^{n} f(y_l; \phi) \mathbf{1} \tag{4.24}$$

where $\pi(\phi)$ is a row vector of $\{\pi_i(\phi)\}$ and $\mathbf{1}$ denotes a column vector of M 1's. Conditions for stationarity and ergodicity of Markov modulated Poisson processes will be given in Section IV-C2.

When the Markov-modulated Poisson process has only two states it is called a *switched Poisson process*. If the rate of one of the states is zero, the process is referred to as an *interrupted Poisson process*. These processes were studied in [130], [233], and [273], where more explicit results could be derived. In particular, Freed and Shepp [130] considered interrupted Poisson processes, and derived a simple formula for the asymptotic likelihood ratio for estimating the state at any instant from a stream of past events. Bounds on the likelihood ratio were given for a switched Poisson process.

Related to Markov-modulated Poisson processes are *discrete-time point processes*. A discrete-time point process is a binary process $\{Y_t\}$ with rate $\{\lambda_t\}$ determined by another random process, such as a Markov chain, and possibly by past observations. $Y_t = 1$ signifies the occurrence of an event at time t, e.g., emission of an electron, while $Y_t = 0$ indicates that no such occurrence has taken place at that time. A recursion for estimating λ_t was developed by Segall [289].

8) Composite Sources: A composite source comprises a collection of discrete-time stationary ergodic subsources and a random switch. At any given time, a single observation or multiple observations are drawn from a subsource selected by the switch. Composite sources become HMPs when the switch is controlled by a first-order discrete-time homogeneous Markov chain, the number of subsources is finite, and the subsources are statistically independent i.i.d. random processes. Composite sources with i.i.d. switch processes and finite number of statistically independent i.i.d. subsources were first introduced by Berger [31]. When the switch position is randomly chosen at time minus infinity, and the switch remains in that position forever, a stationary ergodic process from one of the subsources is observed. The identity or the index of the subsource is not known. The frozen switch position composite source is a mixture process. The ergodic decomposition theorem shows that discrete-time standard alphabet stationary nonergodic processes are composite sources with a switch soldered to its randomly chosen initial position [151, Ch. 7.4]. The special case of discrete-alphabet sources was developed by Gray and Davisson [147].

Composite sources have been found useful in applications such as coding and enhancement of speech signals [93], [9], [104]. A composite source with about 50 stationary subsources, and a switch that may change position every 10–400 ms, can adequately represent the modes of speech signals and their durations [9], [10]. Most of the information in a speech waveform lies in the sequence of modes. The set of modes is essentially independent of the speaker while the switch process is characteristic of the speaker [119]. A collection of universal modes may therefore be used to describe all speech signals as it is done in vector quantization [135]. Composite sources with a switch soldered to its randomly chosen initial position are natural models in universal source coding [75], [147]. The composite source represents a family of possible sources for which a coder is designed. The coder is universal in the sense that it must perform well for all subsources while the identity of the subsource selected by nature is not known. Existence of universal codes for composite sources was proved in [75], [118], [121].

A summary of properties of two-sided composite sources with finite number of subsources was given by Fontana [119]. A composite source is said to be *decomposable* if the switch process is statistically independent of the collection of subsources, i.e., the switch only chooses a subsource but does not otherwise affect its output. Any decomposable composite source has a regular conditional probability $P(G|\boldsymbol{s})$ where G is a set in the σ-field of the observation sequence space and $\boldsymbol{s}$ denotes a switch sequence. The existence of $P(G|\boldsymbol{s})$ is guaranteed for any alphabet of the subsources. If the subsources are jointly stationary then $P(G|\boldsymbol{s})$ is stationary in the sense that $P(G|\boldsymbol{s}) = P(TG|T\boldsymbol{s})$ where T denotes the shift transformation on any two-sided infinite product space. Stationary, mixing, and ergodic properties of a composite source are inherited from the switch process much like what we shall see in Section IV-C for HMPs.

The entropy rate of a sequence of finite-alphabet stationary decomposable composite sources with statistically independent subsources and slowly varying switch processes was studied in [119, Theorem 12]. It is given by a weighted sum of the entropy rates of the individual subsources where the weights are the asymptotic probabilities of the switch process. Limit theorems for the distortion-rate function of a sequence of composite sources with vanishingly slow switch processes were also developed in [119]. Rate-distortion functions for composite sources with an i.i.d. switch process and under varying degrees of knowledge of the switch process at the encoder and decoder were determined by Berger [31]. A correct version of [31, Theorem 6.1.1] was given by Wyner and Ziv [318] where the rate-distortion function of sources with side information at the decoder was developed.

9) The Telegraph Signal: The telegraph signal is an example of a continuous-time binary Markov process. The state space $\mathcal{S} = \{+1, -1\}$ and the generator of the chain is given by $g_{ij} = -g_{ii} = g$ for $i, j = 1, 2$ [315]. When this signal is observed in white noise, it becomes a continuous-time HMP. Finite-dimensional causal MMSE estimation of an M-state continuous-time Markov chain observed in white noise was first developed by Wonham [315]. Noncausal estimation of the states was studied by Yao [323].

C. Stationarity and Ergodicity

Statistical properties of an HMP such as stationarity, ergodicity, mixing, and asymptotic stationarity, are inherited from similar properties of the underlying Markov chain. In the first and second parts of this subsection, we review these concepts for Markov chains and HMPs, respectively. Our presentation in Section IV-C1 follows Grimmett and Stirzaker [154] and Billingsley [38]. In the third part, we discuss exponential forgetting and geometric ergodicity in HMPs. In the fourth part, we provide conditions for stationarity and ergodicity of a switching autoregressive process of the form (4.14). We conclude this section with a local limit theorem for HMPs.

1) The Markov Chain: Consider a discrete-time homogeneous Markov chain $\{S_t, t = 1, 2, \ldots\}$ with finite or countably infinite state space $\mathcal{S}$. Let $\pi_i = P(S_1 = i)$ denote the probability that the chain starts from some state $i \in \mathcal{S}$. Let π denote a row vector with entries $\{\pi_i\}$. This vector represents the initial distribution of the chain. Let $a_{ij} = P(S_{t+1} = j \,|\, S_t = i)$ denote the transition probability for states $i, j \in \mathcal{S}$. Let $A = \{a_{ij}\}$ denote the transition matrix of the chain. Let $\pi_i^{(n)} = P(S_n = i)$ denote the probability of the chain to be in state $i \in \mathcal{S}$ at time $t = n$. Let $\pi^{(n)}$ denote a row vector with entries $\{\pi_i^{(n)}\}$. Let $p_{ij}^{(n)} = P(S_{t+n} = j \,|\, S_t = i)$ denote the n-step transition probability for states $i, j \in \mathcal{S}$. Let $A^{(n)}$ denote a matrix with entries $\{p_{ij}^{(n)}\}$. We have that $\pi^{(1)} = \pi$ and $A^{(1)} = A$. The Chapman–Kolmogorov theorem establishes that $A^{(n)} = A^n$, the nth power of A. Furthermore, $\pi^{(n+1)} = \pi A^n$ [154, p. 215].

To establish conditions for stationarity and ergodicity of Markov chains we need to characterize states and subsets of states within the chain. A state $i \in \mathcal{S}$ is said to be *recurrent*, or *persistent*, if $P(S_{t+1} = i$ for some $t \geq 1 \,|\, X_1 = i) = 1$.

If this probability is smaller than one then the state is called *transient*. Intuitively, the probability that a recurrent state i will be revisited infinitely many times is one. The probability that a transient state i will be revisited infinitely many times is zero. More formally, define the event $F_n = \{S_n = i\}$. The event that infinitely many of the F_n occur, written as F_n *infinitely often* or $\{F_n \text{ i.o.}\}$, satisfies

$$\{F_n \text{ i.o. }\} = \limsup_{n\to\infty} F_n = \bigcap_n \bigcup_{m=n}^{\infty} F_m.$$

It is shown in Billingsley [38, Theorem 8.2] that persistence of a state i is equivalent to $P(S_t = i \text{ i.o.}) = 1$ and to $\sum_n p_{ii}^{(n)} = \infty$. Transience is equivalent to $P(S_t = i \text{ i.o.}) = 0$ and to $\sum_n p_{ii}^{(n)} < \infty$.

Suppose that a chain starts in state i. Let $f_{ij}^{(n)}$ denote the probability that the first visit of the chain to state j occurs after n steps. This probability is given by

$$f_{ij}^{(n)} = P(S_2 \neq j, \ldots, S_n \neq j, S_{n+1} = j | S_1 = i). \quad (4.25)$$

Let T_i denote the time of the first visit to state i, i.e., $T_i = \min\{t > 1: S_t = i\}$. If the visit never occurs then $T_i = \infty$. The probability $P(T_i = \infty \mid S_1 = i) > 0$ if and only if i is transient, and in this case $E\{T_i \mid S_1 = i\} = \infty$. The *mean recurrence time* τ_i of a state i is defined as

$$\tau_i = E\{T_i \mid S_1 = i\} = \begin{cases} \sum_{n=1}^{\infty} n f_{ii}^{(n)}, & \text{if } i \text{ is recurrent} \\ \infty, & \text{if } i \text{ is transient.} \end{cases} \quad (4.26)$$

Note that the mean recurrence time may be infinite even if i is recurrent. A recurrent state i is said to be *positive recurrent*, or *nonnull recurrent*, if $\tau_i < \infty$. Otherwise, the state is called *null recurrent*.

The *period* of a state i is defined as $c(i) = \gcd\{n: p_{ii}^{(n)} > 0\}$, or as the greatest common divisor of the epochs at which returns to i are possible. A state $i \in \mathcal{S}$ is called *periodic* if $c(i) > 1$ and *aperiodic* if $c(i) = 1$. A state is called *ergodic* if it is positive recurrent and aperiodic.

A set $\mathcal{C}$ of states is called *irreducible* if for every pair of states i and j in $\mathcal{C}$, $p_{ij}^{(n)} > 0$ for some n. Thus, $\mathcal{C}$ is irreducible if there is a positive probability of ever visiting a state in the set having started from another state in the set. A set $\mathcal{C}$ of states is called *closed* if $a_{ij} = 0$ for all $i \in \mathcal{C}$ and $j \notin \mathcal{C}$. Thus, the probability of leaving the set is zero. A state is called *absorbing* if the chain never leaves that state. The decomposition theorem for Markov chains establishes that the state space of a Markov chain can be uniquely partitioned as $\mathcal{S} = \mathcal{T}_0 \cup \mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots$ where $\mathcal{T}_0$ is the set of transient states, and $\{\mathcal{C}_m\}$ are irreducible closed sets of recurrent states [154, Theorem 6.3.4]. If $s_1 \in \mathcal{C}_m$ for some m, then the chain never leaves $\mathcal{C}_m$ and that set may be taken to be the whole state space. On the other hand, if $s_1 \in \mathcal{T}_0$, the chain will either stay in $\mathcal{T}_0$ forever or move eventually to one of the $\{\mathcal{C}_m\}$ where it subsequently resides. Thus, if the Markov chain is irreducible, then all states are either transient or recurrent [38, Theorem 8.3]. In an irreducible chain, all states are either positive recurrent or null recurrent. Also, all states are either aperiodic or periodic with the same period [154, Lemma 6.3.2]. When $\mathcal{S}$ is finite, the chain cannot stay in $\mathcal{T}_0$ forever, and there exists at least one recurrent state. Furthermore, all recurrent states are positive recurrent [154, Lemma 6.3.5]. Thus, all states of an irreducible finite-state Markov chain are positive recurrent.

A homogeneous Markov chain is a stationary process if and only if $\pi^{(n)} = \pi^{(1)}$ for all n. Since $\pi^{(n+1)} = \pi^{(1)} A^n$, the process is stationary if and only if the initial distribution $\pi = \pi^{(1)}$ satisfies $\pi = \pi A$. This equation may not have a solution, and when it has one, it may not be unique. Any distribution π that satisfies $\pi = \pi A$ is called a *stationary distribution*. The following summarizes conditions for existence and uniqueness of a stationary distribution [162, Corollary 7, p. 68]. Let $\mathcal{H}_P$ denote the set of positive recurrent states of a Markov chain. If $\mathcal{H}_P$ is empty, the chain has no stationary distributions. If $\mathcal{H}_P$ is a nonempty irreducible set, the chain has a unique stationary distribution π given by $\pi_i = 1/\tau_i$ for $i \in \mathcal{H}_P$ and by $\pi_i = 0$ otherwise. If $\mathcal{H}_P$ is nonempty but not irreducible, the chain has an infinite number of distinct stationary distributions. For example, suppose that $\mathcal{S} = \mathcal{H}_P = \mathcal{C}_1 \cup \mathcal{C}_2$. Any convex combination of the unique stationary distributions of $\mathcal{C}_1$ and of $\mathcal{C}_2$ is a stationary distribution for $\mathcal{S}$. For a finite-state Markov chain, $\mathcal{H}_P$ is a nonempty set, and the chain has a unique stationary distribution if and only if $\mathcal{H}_P$ is irreducible. If the finite-state Markov chain itself is irreducible then it has a unique positive stationary distribution.

Consider next the asymptotic behavior of $p_{ij}^{(n)}$ [154, Theorem 6.4.17]. If the Markov chain is irreducible and aperiodic, then $\lim_{n\to\infty} p_{ij}^{(n)} = 1/\tau_j$ for all i and j. If the chain is transient or null recurrent, $p_{ij}^{(n)} \to 0$ for all i and j since $\tau_j = \infty$. If the Markov chain is irreducible, aperiodic, and positive recurrent, convergence is to the unique stationary distribution, say π, for all states i and j in $\mathcal{S}$

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j = 1/\tau_j. \quad (4.27)$$

For a *finite-state* irreducible aperiodic Markov chain, (4.27) holds, convergence is at an exponential rate

$$\left| p_{ij}^{(n)} - \pi_j \right| \leq D\rho^n \quad (4.28)$$

where $D \geq 0$ and $0 \leq \rho < 1$, and the chain is an ergodic process, see Billingsley [38, Theorem 8.9 and Lemma 2, p. 315]. An ergodic Markov chain satisfying (4.28) is called *geometrically ergodic* [114]. This concept usually applies to a much more general situation of a Markov chain with a continuous state space, see Meyn and Tweedie [240]. Note that the aperiodic condition for ergodicity of the chain is sufficient but not necessary. For example, consider a Markov chain with $a_{ii} = 0$ and $a_{ij} = 1$, $i, j = 1, 2$. This periodic chain has a unique stationary distribution, (4.27) does not hold for this chain, but the chain is an ergodic process.

The transition matrix A of a Markov chain is called *primitive* if there exists some positive integer d such that the d-step transition matrix $A^d = \{p_{ij}^{(d)}\}$ has positive entries, i.e., $A^d > 0$. The smallest such integer d is called the *index of primitivity* of A. The transition matrix of an irreducible aperiodic finite-state

Markov chain is primitive [38, Lemma 2, p. 125]. A finite-state chain with primitive transition matrix has a unique positive stationary distribution and the chain is geometrically ergodic [38, Theorem 8.9]. A corollary of these results is that the chain has a unique positive stationary distribution and is geometrically ergodic when $A > 0$.

An M-state Markov chain with $A > 0$ is *α-mixing* with $\alpha(n) = M\rho^n$ for $0 \leq \rho < 1$. Moreover, any deterministic function of the chain is α-mixing with the same coefficients [38, p. 363]. Since mixing implies ergodicity [38, p. 325], a stationary finite-state Markov chain with positive transition probabilities is stationary ergodic as we have seen before under weaker conditions.

2) The HMP: HMPs are Markov chains observed through channels. Statistical properties of sources observed through channels were developed by Adler [1] for two-sided processes and by Gray [152] for one-sided processes. In particular, when a stationary source is connected to a stationary channel then the source–channel hookup is stationary [152, Lemma 9.3.1]. When a stationary ergodic source is connected to a stationary output weakly mixing channel then the source–channel hookup is stationary and ergodic [152, Lemma 9.4.3]. The channel associated with an HMP is a memoryless invariant channel. As such, it is stationary and output strongly mixing. Hence, an HMP is stationary and ergodic if the Markov chain is stationary, irreducible, and aperiodic. A similar result was directly proved by Leroux [214, Lemma 1] without resorting to the information-theoretic model of the process.

When a stationary mixing source is observed through a stationary output strongly mixing channel, the source–channel hookup is stationary mixing [1]. Hence, an HMP is stationary mixing if the Markov chain is stationary and its transition probabilities are positive. Mixing properties of the two-sided HMP $Y_t = S_t W_t$ in Section IV-B1 were demonstrated by Francq and Roussignol [126].

An additional result showing that when an AMS source is observed through a stationary channel then the source-channel hookup is AMS was developed by Fontana, Gray, and Kieffer [120, Theorem 4], see also [152, Lemma 9.3.2]. Finite-state Markov chains and deterministic functions of such chains are AMS, see Kieffer and Rahe [186, Theorem 9]. Hence, HMPs are AMS.

Conditions for stationarity and ergodicity of a Markov-modulated Poisson process, defined in Section IV-B7, were given by Rydén [273, Lemma 1]. Consider a process with an irreducible continuous-time Markov chain, a generator G, and a diagonal matrix of rates Λ with at least one $\lambda_i > 0$. Let $\phi = (G, \Lambda)$ denote the parameter of the process. A vector ϖ is a stationary distribution of the chain if $\varpi_i \geq 0$, $\sum_i \varpi_i = 1$, and $\varpi = \varpi P_t$ for all $t \geq 0$. This equation is satisfied for all t if and only if $\varpi G = [0, \ldots, 0]$ [154, p. 261]. If a stationary distribution ϖ exists, then it is unique and $\lim_{t\to\infty} p_{ij}^{(t)} = \varpi_j$ for all i, j [154, p. 261]. Recall that $A = (\Lambda - G)^{-1}\Lambda$ is the transition matrix of the discrete-time Markov chain embedded at event epochs. Let $\mathcal{H} = \{i \in \mathcal{S}: \lambda_i > 0\}$ denote the subset of states with corresponding positive Poisson rate. It was shown that for each parameter ϕ, $\mathcal{H}$ is the only set of recurrent aperiodic states. The remaining states are transient. The discrete-time Markov chain has therefore a unique stationary distribution which is positive for all states in $\mathcal{H}$ and zero otherwise. The stationary distribution of A is given by [117, eq. (6)]

$$\pi = \frac{1}{\sum_{i=1}^{M} \varpi_i \lambda_i} \varpi \Lambda. \tag{4.29}$$

Stationarity and ergodicity of the Markov modulated Poisson process are inherited from the Markov chain.

3) Exponential Forgetting and Geometric Ergodicity: We have seen in (4.4) that the likelihood function $p(y^n)$ of an HMP is determined by the state predictive densities and the observation conditional densities. Recall that π is the $1 \times M$ vector representing the initial distribution of the Markov chain. Let ξ_t denote the state predictive density vector at time t. For $j = 1, \ldots, M$, the jth component of this vector is given by $\xi_1(j) = \pi_j$ for $t = 1$ and by $\xi_t(j) = P(S_t = j|y^{t-1})$ for $t = 2, \ldots, n$. Let η_t denote a column vector whose jth element is given by $\eta_t(j) = b(y_t|S_t = j)$. Recall that θ_j denotes the parameter of $\eta_t(j)$ and $\theta = \{\theta_j\}$. Let B_t denote a diagonal matrix whose (j, j) element is $\eta_t(j)$. The state predictive density vector satisfies the following recursion which will be discussed in more details in Section V-A:

$$\begin{aligned} \xi_1 &= \pi' \\ \xi_t &= \frac{A' B_{t-1} \xi_{t-1}}{\eta'_{t-1} \xi_{t-1}}, \qquad t = 2, \ldots, n. \end{aligned} \tag{4.30}$$

The log-likelihood function is given by

$$\log p(y^n; \phi) = \sum_{t=1}^{n} \log(\eta'_t \xi_t). \tag{4.31}$$

Assume the usual parametrization $\phi = (\pi, A, \theta) \in \Phi$ for the HMP in (4.30) and (4.31). Let $\phi^0 \in \Phi$ denote the true value of ϕ used to produce the observation sequence y^n. Assume that ϕ^0 is not known. For identification of ϕ^0, $\log p(y^n; \phi)$ is expected to take different values for different pairs of (A, θ). The effects of π on $\log p(y^n; \phi)$ is expected to be rapidly forgotten so that an arbitrary initial distribution can be used in the recursion (4.30) with no lasting effect. Conditions for identifiability of an HMP are given in Section VI-A.

Le Gland and Mevel [210, Theorem 2.2] proved exponential forgetting of the initial distribution for the prediction recursion (4.30) when ϕ^0 is not known. They referred to this situation as that of a *misspecified* HMP. They assumed that the transition matrix A and its true value A^0 are primitive, but no restrictions were imposed either on θ or on its true value θ^0. To emphasize the dependence of ξ on the observation sequence y_m^n and on the initial distribution $\xi_m = \zeta$, we rewrite it as $\xi(y_m^n, \zeta)$. Let $\tilde{\zeta}$ be another initial distribution. It was shown that

$$\begin{aligned} &\limsup_{n\to\infty} \frac{1}{n} \log \left\| \xi\left(Y_m^n, \zeta\right) - \xi\left(Y_m^n, \tilde{\zeta}\right) \right\| \\ &\qquad \leq \frac{1}{d} \log(1 - D) \qquad P_{\phi^0}\text{-a.s.} \end{aligned} \tag{4.32}$$

where $\|\cdot\|$ denotes the L^1 norm, d denotes the index of primitivity of A, and $D > 0$ is a constant depending on the observation conditional densities of the HMP. An implication of this property is that the log-likelihood function is Lipschitz continuous with respect to some parameter of the model, uniformly in time [210].

For a misspecified HMP, the predictive density vector sequence $\{\xi_t\}$ is not a Markov chain under P_{ϕ^0}, but the triplet (state, observation, wrong predictive density) is a Markov chain. Let $\{Z_t = (S_t, Y_t, \xi_t), t \geq 1\}$ denote that extended Markov chain. Le Gland and Mevel [210, Theorem 3.5, Corollary 3.6] proved geometric ergodicity of the extended Markov chain $\{Z_t\}$ and showed existence of a unique invariant distribution under the assumption that the true and unknown transition matrices are primitive. In particular, this theorem implies an ergodic theorem for the relative entropy density of the HMP [210]. This limit theorem is key to proving consistency of the ML parameter estimator. These subjects are discussed in Sections IV-D and VI-B, respectively.

Exponential forgetting, geometric ergodicity, and existence of a unique invariant distribution for an extended Markov chain defined similarly to $\{Z_t\}$ above, for an HMP with a separable compact state space that is not necessarily finite, were proved by Douc and Matias [90, Proposition 1, Corollaries 1, 2].

A recursion for the gradient of the predictive density vector with respect to a scalar parameter of the HMP can be obtained from (4.30). Exponential forgetting of the initial condition for this recursion were established by Le Gland and Mevel [210, Theorem 4.6]. This result implies that the score function of the HMP is Lipschitz continuous with respect to some parameter of the model, uniformly in time [210]. Let $\{G_t\}$ denote the gradient sequence. Geometric ergodicity of the extended Markov chain

$$\{\tilde{Z}_t = (S_t, Y_t, \xi_t, G_t), t \geq 1\}$$

and existence of a unique invariant distribution, were proved by Le Gland and Mevel [210, Theorem 5.4, Corollary 5.5] under some integrability assumptions. The implications of this result are that a central limit theorem for the score function and a law of large numbers for the Hessian matrix follow [210]. These limit theorems are key in proving asymptotic normality of the ML parameter estimator. This subject will be discussed in Section VI-B.

Exponential forgetting and geometric ergodicity for similarly defined extended Markov chains, involving the score function and Hessian of a misspecified HMP with a separable compact state space that is not necessarily finite, were proved by Douc and Matias [90, Appendix D].

Another form of exponential forgetting was demonstrated by Douc, Moulines, and Rydén [91] for switching autoregressive processes with a separable compact state space that is not necessarily finite. Let r denote the order of the autoregressive process and let $\phi^0 \in \Phi$ be the true parameter. They showed that for any $n \geq m$ and $\phi \in \Phi$, the state sequence $\{S_t, t \geq m\}$ given an observation sequence y_{m-r+1}^n, is an inhomogeneous Markov chain under the stationary measure P_ϕ [91, Lemma 1]. Exponential forgetting of the initial distribution for this inhomogeneous Markov chain was shown in [91, Corollary 1]. This property is key in proving consistency and asymptotic normality of the ML parameter estimator of ϕ^0.

4) Switching Autoregressive Processes: Conditions for stationarity and ergodicity of a switching autoregressive process of the form (4.14) were given by Francq and Roussignol [127, Theorem 1]. Recall that an HMP is a special case of this process. Let ϕ^0 denote the true parameter of the switching autoregressive process. It was assumed that i) the Markov chain $\{S_t\}$ is irreducible and aperiodic, ii) $E\{\|h(V_t, j; \phi^0)\|\} < \infty$ for all $j = 1, \ldots, M$ where $\|\cdot\|$ denotes the usual Euclidean norm, and iii) there exist constants $\alpha_1, \ldots \alpha_M$ such that for all $j = 1, \ldots, M$ and all $(x, y) \in \mathcal{R}^r \times \mathcal{R}^r$

$$\left\|g\left(x, j; \phi^0\right) - g\left(y, j; \phi^0\right)\right\| \leq \alpha_j \|x - y\| \tag{4.33}$$

and the matrix $Q = D_\alpha (A^0)'$, where $D_\alpha = \operatorname{diag}(\alpha_1, \ldots, \alpha_M)$ and A^0 is the true transition matrix of $\{S_t\}$, has spectral radius smaller than 1. Under these conditions, it was shown that the Markov chain $\{Z_t, S_t\}$ on $\mathcal{R}^r \times \mathcal{S}$ admits a unique stationary probability ν. The second marginal of ν is equal to the stationary probability of $\{S_t\}$. Moreover, a stationary Markov chain $\{Z_t, S_t\}$ satisfying (4.14) with ν as initial distribution is an aperiodic ergodic Harris process [114], [240].

5) A Local Limit Theorem: A local limit theorem for zero-mean stationary ergodic general HMPs with finite second-order moment that satisfy some mild conditions was proven by Maxwell and Woodroofe [231]. Let P denote the distribution of the HMP. For the partial sum of HMP observations, $\Sigma_n = Y_1 + \cdots + Y_n$, it was shown that

$$\lim_{n \to \infty} \sqrt{n} P(\alpha < \Sigma_n \leq \beta) = c(\beta - \alpha) \tag{4.34}$$

for some positive constant c and $-\infty < \alpha < \beta < \infty$, and

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \sqrt{k} P(S_k \in D, \alpha < \Sigma_n \leq \beta) = cP(D)(\beta - \alpha). \tag{4.35}$$

D. Entropy Ergodic Theorems

In this subsection, we review ergodic theorems for the sample entropy and relative entropy densities of an HMP. The fundamental ergodic theorem for the sample entropy of a stationary ergodic finite-alphabet process, not necessarily an HMP, is given by the Shannon–McMillan–Breiman theorem [68, Theorem 15.7.1]. Let $\{Y_t\}$ denote such a process and let P_Y denote its distribution. Let $p(y^n)$ denote the n-dimensional pmf induced by P_Y. The theorem states that

$$\lim_{n \to \infty} -\frac{1}{n} \log p(Y^n) = \overline{H}(P_Y) \qquad P_Y\text{-a.s.} \tag{4.36}$$

where

$$\begin{aligned} \overline{H}(P_Y) &= \lim_{n \to \infty} \frac{1}{n} E_{P_Y}\{-\log p(Y^n)\} \\ &= \lim_{n \to \infty} E_{P_Y}\left\{-\log p\left(Y_n | Y^{n-1}\right)\right\} \\ &= E_{P_Y}\left\{-\log p\left(Y_0 | Y_{-\infty}^{-1}\right)\right\} < \infty \end{aligned} \tag{4.37}$$

is the *entropy rate* of $\{Y_t\}$ [152, p. 24]. Another common notation for the entropy rate is $\overline{H}(Y)$.

Let $\{Y_t,\, t \geq 1\}$ denote a stationary ergodic HMP with distribution P_{ϕ^0} for some parameter $\phi^0 \in \Phi$. This one-sided process may be considered part of a two-sided stationary ergodic process with the index set of all integers. The n-dimensional density of the HMP with respect to μ^n is the density $p(y^n;\, \phi^0)$ given by (4.8). For a finite-alphabet HMP, μ is the counting measure κ. For a general HMP, μ is any σ-finite measure. For a finite-alphabet HMP we have from (4.36)

$$\lim_{n\to\infty} -\frac{1}{n} \log p(Y^n;\, \phi^0) = \overline{H}(P_{\phi^0}) \qquad P_{\phi^0}\text{-a.s.} \tag{4.38}$$

Leroux [214, Theorem 1] proved (4.38) for a stationary ergodic general HMP. He assumed an irreducible aperiodic Markov chain and observation conditional densities that satisfy

$$E_{\phi^0}\{|\log b(Y_1;\, \theta_j(\phi^0))|\} < \infty, \qquad \text{for } j = 1, \ldots, M.$$

This extension is in fact a special case of Barron's ergodic theorem [23] which we discuss shortly.

Let $P_\phi,\ \phi \in \Phi$, denote a distribution of the HMP and let $p(y^n;\, \phi)$ denote the induced n-dimensional density with respect to μ^n as given by (4.8). The parameters ϕ and ϕ^0 are not necessarily equivalent. We are now interested in ergodic theorem for $n^{-1} \log p(Y^n;\, \phi)$ when $\{Y_t\}$ is the stationary ergodic HMP with distribution P_{ϕ^0}. Baum and Petrie [25, Theorem 3.2] and Petrie [251, Theorem 2.1] developed the theorem for a finite-alphabet HMP. Petrie [251] relaxed the assumption that $\phi \in \Phi_\delta$ made in [25]. Leroux [214, Theorem 2] proved the theorem for a general HMP. The ergodic theorem states that

$$\lim_{n\to\infty} \frac{1}{n} \log p(Y^n;\, \phi) = \overline{H}(P_{\phi^0},\, P_\phi) \qquad P_{\phi^0}\text{-a.s.} \tag{4.39}$$

where

$$\overline{H}(P_{\phi^0},\, P_\phi) = \lim_{n\to\infty} \frac{1}{n} E_{\phi^0}\{\log p(Y^n;\, \phi)\} < \infty. \tag{4.40}$$

Define

$$\begin{aligned}\overline{D}(P_{\phi^0} \| P_\phi) &= \lim_{n\to\infty} \frac{1}{n} E_{\phi^0}\left\{\log \frac{p(Y^n;\, \phi^0)}{p(Y^n;\, \phi)}\right\} \\ &= \overline{H}(P_{\phi^0},\, P_{\phi^0}) - \overline{H}(P_{\phi^0},\, P_\phi)\end{aligned} \tag{4.41}$$

and note that $\overline{H}(P_{\phi^0},\, P_{\phi^0}) = -\overline{H}(P_{\phi^0})$. Baum and Petrie [25, Theorem 3.1], Petrie [251, Proposition 2.2, Theorem 2.5], and Leroux [214, Lemma 6] proved that

$$\overline{D}(P_{\phi^0} \| P_\phi) \geq 0 \quad \text{with equality iff } \phi \sim \phi^0. \tag{4.42}$$

This important property provides a criterion for distinguishing between the equivalence classes of ϕ and ϕ^0, and is key in proving consistency of the ML estimator of ϕ^0. For an identifiable HMP, the equivalence class of ϕ^0 comprises all points in Φ obtained by permutations of the states of the HMP. A similar statement holds for the equivalence class of ϕ.

Leroux showed that theorem (4.39) holds for any choice of positive initial distribution π and $\overline{H}(P_{\phi^0},\, P_\phi)$ is the same for any such choice. $\overline{H}(P_{\phi^0},\, P_\phi)$ may possibly be equal to $-\infty$. He proved the theorem using Kingman's [188] ergodic theorem for subadditive processes assuming an irreducible aperiodic Markov chain and observation conditional densities $\{b(y_t;\, \theta_j),\, \theta_j \in \Theta\}$ that satisfy

$$E_{\phi^0}\left\{\sup_{\|\tilde{\theta}_j - \theta_j\| < \delta} (\log b(Y_1;\, \tilde{\theta}_j))^+\right\} < \infty$$

for some $\delta > 0$ where $\|\cdot\|$ denotes the Euclidean distance and $x^+ = \max\{x,\, 0\}$. Theorems (4.39) and (4.42) hold for any ϕ in the one-point compactified parameter space Φ_c. Compactification extends the parameter set Φ into a compact set Φ_c. For the usual parametrization, Φ_c is obtained from compactification Θ_c of the parameter space Θ. The latter is done by attaching to Θ a point denoted ∞, and extending $b(y_t;\, \theta_j)$ to Θ_c by defining $b(y_t;\, \infty) = 0$. For example, if $b(y_t;\, \theta_j)$ is the Poisson density with mean θ_j then $\Theta_c = [0,\, \infty]$. A regularity condition assumed in [214] ensures continuity of $b(y_t;\, \cdot)$ over Θ_c. For any other parametrization ϕ of the HMP, $\theta_j(\phi) \in \Theta_c$ for $j = 1, \ldots, M$. In proving (4.42), the assumption quoted after (4.38) was also made.

If we assume that $\mu^n \ll P_\phi^{(n)}$ in addition to our earlier assumption that $P_\phi^{(n)} \ll \mu^n$, then the two measures are *equivalent*, and

$$\frac{dP_{\phi^0}^{(n)}}{dP_\phi^{(n)}} = \frac{p(Y^n;\, \phi^0)}{p(Y^n;\, \phi)}. \tag{4.43}$$

For this case, (4.38) and (4.39) imply an ergodic theorem for the relative entropy density of one general HMP with respect to another

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{dP_{\phi^0}^{(n)}}{dP_\phi^{(n)}} = \overline{D}(P_{\phi^0} \| P_\phi) \qquad P_{\phi^0}\text{-a.s.} \tag{4.44}$$

In addition, we may now call $\overline{D}(P_{\phi^0} \| P_\phi)$ the *relative entropy rate* [152, p. 150].

Similar ergodic theorems for relative entropy densities of several extensions of standard HMPs were recently proved under suitable conditions. Francq and Roussignol [127] studied stationary ergodic switching autoregressive processes with finite-state Markov regime given by (4.14). They proved an ergodic theorem similar to (4.39) for the normalized conditional log-likelihood $n^{-1} \log p(Z^n|z_0;\, \phi)$ [127, eq. (11)]. They expressed the conditional density as a product of random matrices and applied Furstenberg and Kesten [132] ergodic theorem. The sequence converges almost surely to the upper Lyapunov exponent of the sequence of random matrices. They also proved (4.42) [127, Theorem 2]. Conditions for a switching autoregressive process of the form (4.14) to be stationary ergodic were given in Section IV-C4. For the matrix product form of the likelihood function of a standard HMP see (5.12).

Krishnamurthy and Rydén [198, Lemma 1] studied stationary ergodic switching autoregressive processes with finite-state Markov regime described by (4.15) and arrived at a similar ergodic theorem for the normalized conditional log likelihood. They used Kingman's ergodic theorem following Leroux [214]. They also showed in [198, Lemma 4] that $\overline{D}(P_{\phi^0} \| P_\phi) \geq 0$ but the implications of $\overline{D}(P_{\phi^0} \| P_\phi) = 0$ are not as explicit as for the process studied in [127, Theorem 2].

Le Gland and Mevel [210] proved an ergodic theorem similar to (4.39) for a finite-state general HMP using geometric ergodicity of an extended Markov chain as described in Section IV-C3. Douc and Matias [90] extended this approach to a general HMP with a separable compact state space that is not necessarily finite. They developed an ergodic theorem similar to (4.39) for an HMP with arbitrary initial state density not necessarily a stationary density [90, Proposition 4]. They also proved (4.42) [90, Theorem 1]. It was noted in [90] that Leroux's approach does not immediately apply to HMPs with a continuous state space.

Douc, Moulines, and Rydén [91] studied general forms of switching autoregressive processes with a separable compact state space that is not necessarily finite. They proved an ergodic theorem similar to (4.39) for almost sure and L^1 convergence of the normalized conditional log likelihood of the observation sequence [91, Proposition 1]. They also proved (4.42) [91, Proposition 3]. They relied on uniform exponential forgetting of the initial distribution of the inhomogeneous Markov chain representing the states given the observation sequence. It was noted in [91] that application of the approach used in [90] would have required stronger assumptions.

Rydén [273, Lemmas 5, 8] proved an ergodic theorem similar to (4.39) and the conclusion (4.42) for a Markov-modulated Poisson process. The main difference between the HMPs in Leroux [214] and Rydén [273] is that for the former case (4.2) holds while in the latter (4.9) holds as explained in Section IV-B7. In addition, compactification of the parameter set is not possible since $p(y^n;\,\phi)$ does not always vanish at infinity.

We turn now to the general ergodic theorem for relative entropy density developed by Barron [23]. See also [152, Theorem 8.2.1]. Consider a standard alphabet random process $\{Y_t,\, t \geq 1\}$ described by a stationary ergodic distribution P_Y on a sequence measurable Borel space [152, p. 12]. Let Q be a σ-finite Markov measure of order $m \geq 0$ that has stationary transition probabilities and is defined on the same measurable space. Let $P_Y^{(n)}$ and $Q^{(n)}$ denote the n-dimensional distributions induced by P_Y and Q, respectively. Assume that $P_Y^{(n)} \ll Q^{(n)}$ for all n. Let $f(Y^n) = dP_Y^{(n)}/dQ^{(n)}$ denote the Radon–Nikodym derivative or density of $P_Y^{(n)}$ with respect to $Q^{(n)}$. Let

$$f(Y_n|Y^{n-1}) = f(Y^n)/f(Y^{n-1}), \qquad \text{for } n > 1$$

and

$$f(Y_1|Y^0) = f(Y_1).$$

Assume that

$$D_n = E_{P_Y^{(n)}}\{\log f(Y_n|Y^{n-1})\} > -\infty$$

for some $n \geq m$. This condition is automatically satisfied if $Q^{(n)}$ is a finite measure or a probability measure. In the latter case, $D_n \geq 0$. The theorem states that

$$\lim_{n\to\infty} \frac{1}{n} \log f(Y^n) = \overline{D}(P_Y\|Q) \quad P_Y\text{-a.e. and in } L^1 \quad (4.45)$$

where $\overline{D}(P_Y\|Q)$ is the relative entropy rate defined similarly to (4.41)

$$\begin{aligned}\overline{D}(P_Y\|Q) &= \lim_{n\to\infty} \frac{1}{n} E_{P_Y}\left\{\log \frac{dP_Y^{(n)}}{dQ^{(n)}}\right\} \\ &= \lim_{n\to\infty} E_{P_Y}\left\{\log f(Y_n|Y^{n-1})\right\}. \quad (4.46)\end{aligned}$$

Theorem (4.38) for a general HMP could be obtained from (4.45) if $P_Y = P_{\phi^0}$ and $Q^{(n)} = \mu^n$. If

$$E_{\phi^0}\{\min_j \log b(Y_1;\,\theta_j)\} > -\infty$$

then $D_n > -\infty$ and the theorem holds. This condition results from application of Jensen's inequality to (4.4).

An ergodic theorem for $\log f(Y^n)$ when P_Y is AMS and Q is the same Markov measure as above was proved by Barron [23, Theorem 3]. See also Gray [152, Theorem 8.4.1]. Let $\overline{P}_Y$ denote a stationary distribution that asymptotically dominates P_Y. This may be the stationary mean of the AMS process. Let $\overline{f}(Y^n) = d\overline{P}_Y^{(n)}/dQ^{(n)}$. It was shown that if $n^{-1}\log \overline{f}(Y^n) \to h$ $\overline{P}_Y$-a.e. for some shift-invariant measurable function h then also $n^{-1}\log f(Y^n) \to h$ P_Y-a.e. Ergodic theorems for $\log f(Y^n)$ when P_Y is stationary but not ergodic were proved by Barron [23, Theorem 2] and Gray [152, Corollary 8.3.1].

Without the Markov property for the dominating measure Q, convergence of $\log f(Y^n)$ is not guaranteed [23]. When P_Y and Q are two stationary ergodic general HMP distributions, (4.44) provides a version of Barron's theorem with an HMP dominating measure. For finite-alphabet processes, an HMP dominating measure may replace the Markov measure in (4.45) provided that its parameter $\phi \in \Phi_\delta$. This result was first shown by Finesso [116, Theorem 2.3.3] and then by Kehagias [183, Lemma 1]. In particular, Finesso [116, Sec. 2.4] proved that if under P_Y the process $\{Y_t\}$ is stationary ergodic, and $Q = P_\phi$ is an HMP distribution with corresponding n-dimensional pmf $p(y^n;\,\phi)$, then

$$\lim_{n\to\infty} \frac{1}{n} \log p(Y^n;\,\phi) = \overline{H}(P_Y,\,P_\phi) \qquad P_Y\text{-a.s.} \quad (4.47)$$

where $\overline{H}(P_Y,\,P_\phi)$ is defined similarly to (4.40) and convergence is *uniformly* in $\phi \in \Phi_\delta$. This theorem is particularly useful when one wishes to model a stationary ergodic process (P_Y) by an HMP (P_ϕ) and performs ML estimation of its parameter by maximizing $n^{-1}\log p(y^n;\,\phi)$ over $\phi \in \Phi_\delta$. In addition, $-\overline{H}(P_Y,\,P_\phi) = \overline{H}(P_Y) + \overline{D}(P_Y\|P_\phi)$ is the asymptotically minimum average length of a source code designed for the stationary ergodic source P_Y assuming that this source is the HMP P_ϕ [68, Theorem 5.4.3].

E. Finite-Alphabet HMPs

In this subsection, we summarize results for finite-alphabet HMPs and deterministic functions of finite-state Markov chains. We first show that the two classes of processes are closely related. Then we focus on an important subclass of finite-alphabet HMPs known as unifilar sources. This class is amenable to the method of types and hence is particularly attractive. We conclude with some bounds on the entropy rate and rate-distortion function of finite-alphabet HMPs.

Consider an HMP with M states and L letters. Define the Cartesian product $\mathcal{Q} = \mathcal{Y} \times \mathcal{S}$ of observation and state spaces, and a deterministic function g: $\mathcal{Q} \rightarrow \mathcal{Y}$ by $g(y, s) = y$. Rewriting (4.1) redundantly, we have

$$p(y^n, s^n) = p(y_1, s_1) \prod_{t=2}^{n} p(y_t, s_t | y_{t-1}, s_{t-1}). \tag{4.48}$$

Hence, $\{(Y_t, S_t), t \geq 1\}$ is a Markov chain with $L \times M$ states, and $Y_t = g(Y_t, S_t)$. Thus, any finite-alphabet HMP is a deterministic function of a Markov chain with augmented state space [25]. Conversely, if $Y_t = h(S_t)$ for some function h and Markov chain $\{S_t\}$, then $\{Y_t\}$ is an HMP with $P(Y_t = y | S_t = s) = 1$ if $h(s) = y$ and zero otherwise [161]. Thus, any deterministic function of finite-state Markov chain is a trivial HMP.

Let $Y_t = h(S_t)$ for some many-to-one function h and Markov chain $\{S_t\}$. The function h may collapse one or more states of $\{S_t\}$ onto a single letter of $\{Y_t\}$. The process $\{Y_t\}$ is therefore referred to as *aggregated Markov process* [278], [206]. The process $\{Y_t\}$ is not in general a Markov chain and it exhibits long statistical dependencies. It inherits stationarity and ergodicity from the Markov chain [25]. Necessary and sufficient conditions for $\{Y_t\}$ to be a Markov chain were developed by Burke and Rosenblatt [52]. Conditions for stationary processes to be functions of Markov chains were developed by Dharmadhikari [83]–[86], Heller [160], and Erickson [107]. A partial summary of these results appears in [272, pp. 77–78]. These results are not constructive in the sense that they do not lead to an algorithm for producing the Markov chain $\{S_t\}$ and function h for a given stationary process $\{Y_t\}$. Identifiability of a function of Markov chain was first studied by Blackwell and Koopmans [42], Gilbert [136], and Carlyle [54]. Identifiability of a finite-alphabet HMP was studied by Petrie [251]. Identifiability of a function of a nonstationary Markov chain was studied by Ito, Amari, and Kobayashi [167], Rydén [278], and Larget [206]. These results will be further discussed in Section VI-A. See [167] for additional references.

A deterministic function of Markov chain which produces distinct letters when the chain transits from each state i to all states $\{j\}$ with $a_{ij} > 0$ was referred to as a unifilar source by Ash [14]. For unifilar sources, the state s_t is uniquely determined by the previous state s_{t-1} and the current letter y_t. The entire state sequence s^n can be read from the observation sequence y^n provided that the initial state s_0 is known. An important special case of unifilar sources is the mth-order Markov chain $\{X_t\}$ with states defined as $\{S_t = X_{t-m+1}^t\}$.

A more general source was introduced by Gallager who referred to it as *Markov source* [133, Sec. 3.6]. The source is characterized by an initial state s_0, a transition pmf $p(y_t|s_{t-1})$, and a deterministic *next-state function* $s_t = g(y_t, s_{t-1})$ for $t = 1, 2, \ldots$. Given the initial state s_0, an observation y_1 is generated according to $p(y_1|s_0)$ and a new state $s_1 = g(y_1, s_0)$ is chosen. Next, y_2 is generated according to $p(y_2|s_1)$, and so on. The n-dimensional pmf of the source is given by

$$p(y^n|s_0) = \prod_{t=1}^{n} p(y_t|s_{t-1}). \tag{4.49}$$

By construction, s^n is uniquely determined by y^n and the initial state s_0 as we have seen for unifilar sources. The observation y_t, however, is a not a deterministic function of s_t unless $g(\cdot, s_{t-1})$ is a one-to-one function given s_{t-1}. We shall not impose this restriction on g. We shall refer to this source as the *unifilar* source. Other authors have used the more explicit name of *unifilar finite-state source*.

Unifilar sources are mathematically tractable since they are amenable to the *method of types* much like i.i.d. sources and Markov chains. The method of types for i.i.d. sources was developed by Csiszár and Körner [70], [73]. Consider an i.i.d. finite-alphabet source with L letters and pmf $p(\cdot)$. The method of types characterizes the sample space of n-length source sequences by an exhaustive set of empirical distributions called *types*. The set of all n-length source sequences having the same type forms a *type class*. The set of all type classes forms a partition of the sample space of all n-length source sequences. Let y^n denote an observation sequence with empirical pmf $q_n(\cdot)$. Let

$$H(q_n) = -\sum_{y} q_n(y) \log q_n(y)$$

denote the *empirical entropy*. Let

$$D(q_n \| p) = \sum_{y} q_n(y) \log(q_n(y)/p(y))$$

denote the *relative entropy* between $q_n(\cdot)$ and $p(\cdot)$. The following facts were established. We use $\approx$ to denote approximations up to polynomial factors. The pmf of the sequence y^n can be written as

$$p(y^n) = 2^{-n[H(q_n)+D(q_n\|p)]}. \tag{4.50}$$

Hence, all sequences within a given type class are equally likely. There is a polynomial number of types that does not exceed $(n+1)^{L-1}$. There is an exponential number of sequences in each type class given by $\approx 2^{nH(q_n)}$. The probability of a type class is given by $\approx 2^{-nD(q_n\|p)}$.

A summary of the method of types for unifilar sources, which is similar to that for Markov chains, can be found in Csiszár [73]. Let y^n denote an observation sequence from a unifilar source and let s_0^{n-1} denote the state sequence recovered from y^n and s_0. Let $q_n(y, s)$ denote the pmf of the *joint type* of (y^n, s_0^{n-1}). The joint type is given by the relative frequency of appearance of (y, s) among the n pairs $(y_1, s_0), \ldots, (y_n, s_{n-1})$. Let $q_n(y|s)$ denote the empirical transition pmf induced by the joint type. Let

$$H(q_n) = -\sum_{y, s} q_n(y, s) \log q_n(y|s) \tag{4.51}$$

denote the empirical conditional entropy, and let

$$D(q_n \| p) = \sum_{y, s} q_n(y, s) \log \frac{q_n(y|s)}{p(y|s)} \tag{4.52}$$

denote the conditional relative entropy. The following facts were established. The pmf (4.49) has the form of (4.50) with $H(q_n)$ and $D(q_n\|p)$ given by (4.51) and (4.52), respectively. All sequences within a given type class are equally likely. There is a polynomial number of joint types that is larger than $cn^{M(L-1)}$ for some constant c but does not exceed $(n+1)^{M(L-1)}$. The lower bound is due to Alon, cited in [309, Lemma 2]. The cardinality of a type class is $\approx 2^{nH(q_n)}$ and the probability of a type class is $\approx 2^{-nD(q_n\|p)}$.

No extension of the method of types to HMPs is known. Hence, the analysis of HMPs is generally much harder as their statistics cannot be summarized by types. In some problems, this difficulty may be circumvented by defining a conditional type class and lower-bounding its cardinality using the Lempel–Ziv universal codeword length instead of the empirical entropy as for the joint type above. This approach was demonstrated by Ziv and Merhav [330] and Merhav [236]. In addition, any finite-alphabet HMP for which $p(y_t, s_t|s_{t-1}) > \delta > 0$ for all $y_t \in \mathcal{Y}$ and $(s_{t-1}, s_t) \in \mathcal{S}^2$ can be approximated by a unifilar source having sufficiently large number of states as was shown by Zeitouni, Ziv, and Merhav [325, Appendix].

The entropy rate of a unifilar source $\{Y_t\}$ was given by Gallager [133, Theorem 3.6.1]. Let $b_{il} = P(Y_t = l|S_t = i)$ and

$$q_i = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} P(S_t = i).$$

Let

$$H(Y_t|S_t = i) = -\sum_{l=1}^{L} b_{il} \log b_{il}$$

be the conditional entropy of Y_t given that the chain is in state i. The entropy rate of the unifilar source is given by

$$\overline{H}(Y) = \sum_{i=1}^{M} q_i H(Y_t|S_t = i). \tag{4.53}$$

If the Markov chain is irreducible aperiodic with stationary distribution π then $q_i = \pi_i$.

No explicit single-letter expression for the entropy rate of an HMP is known [41]. Sequences of asymptotically tight bounds for the entropy rate of a deterministic function of a stationary Markov chain were first developed by Birch [40]. The same bounds appear in Cover and Thomas [68, p. 69]. Gallager [133, Problem 3.23] provides the same bounds for a finite-alphabet stationary HMP. The bounds are given in terms of conditional entropies of the process. For a process $\{Y_t\}$ with n-dimensional pmf $p(y^n)$, the *conditional entropy* $H(Y_n|Y^{n-1})$ is defined by [68, p. 16]

$$H\left(Y_n|Y^{n-1}\right) = \sum_{y^n} p(y^n) \log p\left(y_n|y^{n-1}\right). \tag{4.54}$$

For a stationary process, this conditional entropy is a monotonically nonincreasing sequence which converges to the entropy rate $\overline{H}(Y)$ of the process. Hence, $H(Y_n|Y^{n-1})$ provides an upper bound for $\overline{H}(Y)$. For the lower bound, the conditional entropy $H(Y_n|Y^{n-1}, S_1)$ is used. This conditional entropy is a monotonically nondecreasing sequence which also converges to the entropy rate $\overline{H}(Y)$. In addition, $H(Y_n|Y^{n-1}, S_1) \le H(Y_n|Y^{n-1})$ [68, p. 27]. Thus, for each n

$$H\left(Y_n|Y^{n-1}, S_1\right) \le \overline{H}(Y) \le H\left(Y_n|Y^{n-1}\right) \tag{4.55}$$

and

$$\lim_{n\to\infty} H\left(Y_n|Y^{n-1}, S_1\right) = \overline{H}(Y) = \lim_{n\to\infty} H\left(Y_n|Y^{n-1}\right). \tag{4.56}$$

The difference between the upper and lower bounds in (4.55) is the *conditional mutual information* [68, p. 22]

$$I\left(Y_n; S_1|Y^{n-1}\right) = H\left(Y_n|Y^{n-1}\right) - H\left(Y_n|Y^{n-1}, S_1\right). \tag{4.57}$$

It signifies the amount of information that can be gained about S_1 from Y_n given Y^{n-1}. The rate at which this difference approaches zero is of theoretical and practical importance. Birch [40] showed that if the transition matrix $A > 0$, then $I(Y_n; S_1|Y^{n-1})$ converges to zero exponentially fast with n.

A lower bound on the rate-distortion function of a finite-alphabet HMP was developed by Gray [146]. The rate-distortion function $R(D)$ provides the minimum possible bit rate R required by *any* encoder to encode the source with average distortion that does not exceed D [68, p. 341]. Consider an HMP with alphabet of size L, n-dimensional pmf $p(y^n)$, and entropy rate $\overline{H}(Y)$. Let $d(y, v)$ be a distortion measure between a letter y and its encoded version v. Assume that $\{d(y, v): y \in \mathcal{Y}\}$ is independent of v. Such a distortion measure is called *balanced*. The Hamming measure $d(y, v) = 1 - \delta_{y,v}$, where $\delta_{y,v}$ is the Kronecker delta, has this property. Let

$$d(y^n, v^n) = n^{-1} \sum_{t=1}^{n} d(y_t, v_t)$$

be the distortion between y^n and v^n. Define the set of conditional pmfs $q(v^n|y^n)$ for all possible encoders that provide average distortion smaller than or equal to D as

$$\mathcal{Q} = \{q(v^n|y^n): E\{d(Y^n, V^n)\} \le D\} \tag{4.58}$$

where expectation is taken with respect to the joint pmf $q(v^n, y^n) = q(v^n|y^n)p(y^n)$. Let $I(Y^n; V^n)$ denote the mutual information between the HMP observation sequence Y^n and its encoded version V^n

$$I(Y^n; V^n) = \sum_{v^n, y^n} q(v^n, y^n) \log \frac{q(v^n, y^n)}{q(v^n)p(y^n)}. \tag{4.59}$$

The *rate distortion function* is defined as [68, p. 341]

$$R(D) = \lim_{n\to\infty} \inf_{q(v^n|y^n)\in\mathcal{Q}} \frac{1}{n} I(Y^n; V^n). \tag{4.60}$$

The bound on the rate-distortion function is given by

$$R(D) \ge \overline{H}(Y) - \log f - \rho D \tag{4.61}$$

where ρ is a constant and

$$f = \sum_{l=1}^{L} \exp[-\rho d(l, 1)]. \tag{4.62}$$

The optimal value of ρ that maximizes the bound is obtained from

$$D = -\frac{\partial}{\partial \rho} \log f(\rho). \tag{4.63}$$

The bound is the same for all HMPs having the same alphabet size and entropy rate. There exists a distortion interval $[0, D_c]$ over which the bound is tight provided that the initial distribution of the Markov chain is positive and either the state transition matrix $A > 0$ or the state-to-observation transition matrix $B > 0$. The value of D_c depends on the number of states.

V. State Estimation

Estimation of the state sequence S^n from an observation sequence y^n is of considerable theoretical and practical importance. Estimation of the state S_t from y^n is a *prediction* problem when $t > n$, a *filtering* problem when $t = n$, and a *smoothing* problem when $t < n$. The state sequence S^n may be estimated under various criteria. The most common criteria are minimum *symbol* error probability and minimum *sequence* error probability. In the first case, an estimate $\tilde{s}_t \in \mathcal{S}$ is chosen by minimizing the probability of error $P(\tilde{S}_t \neq S_t | y^n)$. This results in the maximum *a posteriori* (MAP) *symbol* decision rule

$$\tilde{s}_t = \arg\max_{s_t} p(s_t|y^n) \tag{5.1}$$

and the sequence S^n is estimated as $\tilde{s}^n$. Computationally efficient forward–backward recursions for calculating $p(s_t|y^n)$ were developed by Chang and Hancock [56], Ott [248], Raviv [265], Baum, Petrie, Soules, and Weiss [28], Forney [124], Bahl, Cocke, Jelinek, and Raviv [17], Lindgren [219], Askar and Derin [15], Devijver [81], and Kitagawa [189]. These recursions will be presented in Section V-A. Estimation of the state sequence using the second criterion results in the MAP *sequence* estimate given by

$$\hat{s}^n = \arg\max_{s^n \in \mathcal{S}^n} p(s^n|y^n). \tag{5.2}$$

This problem is solved using dynamic programming [30] or by the well-known Viterbi algorithm [308], [124], [285].

When the states are considered unit vectors in an M-dimensional space, as was done in Section IV-B, they can be estimated in the MMSE sense. The conditional mean estimate in this case is the vector of conditional probabilities $\{p(s_t|y^n),\ s_t = 1, \ldots, M\}$. This approach enables application of nonlinear estimation techniques [99] and will be presented in Section IX. In a related approach, Golubev [143] studied causal conditional mean estimation of S_t given y^t when the finite number of values that S_t can take were assumed real numbers rather than integers. If $S_t \in \{\nu_1, \ldots, \nu_M\}$, then the MMSE estimator of S_t is given by

$$S_t^* = \sum_{j=1}^{M} \nu_j P(S_t = j|y^t).$$

Boguslavskii and Borodovskii [44] proposed rounding the conditional mean estimate of S_t given y^n to the nearest neighbor integer.

Note that the two schemes (5.1) and (5.2) are smoothing approaches except when estimating the nth state from y^n. While the error probability of the MAP symbol decision rule (5.1) cannot exceed that of the MAP sequence decision rule (5.2), there is no guarantee that the sequence estimate $\tilde{s}^n$ is admissible since it may contain transitions that are *a priori* impossible. Both estimators (5.1) and (5.2) require the entire sequence of n observations y^n for estimating a state at time $1 \le t \le n$. A low-delay symbol MAP decoder was proposed in [249]. A hybrid of the two approaches (5.1) and (5.2) was proposed by Brushe, Mahony, and Moore [50]. A forward–backward recursion that depends on a soft-decision parameter v was developed such that symbol decoding (5.1) is obtained when $v \to 0$ and sequence decoding (5.2) is obtained when $v \to \infty$. In particular, this approach shows that the Viterbi algorithm can be implemented in a forward–backward manner.

The forward–backward recursions of Chang and Hancock [56] as well as the Viterbi algorithm were shown by Aji and McEliece [3] to be special cases of a *generalized distributive law* which is used to marginalize a product function such as a product of pdfs. Many other algorithms, such as the turbo decoding algorithm, fall into this category [3].

In some applications, such as automatic speech recognition, it is often desirable to find several state sequences that mostly contribute to the likelihood function $p(y^n) = \sum_{s^n} p(s^n, y^n)$. An algorithm that accomplishes this task was proposed by Foreman [123]. In other applications, lumpable HMPs are encountered. These are generalizations of lumpable Markov chains, whereas states can be grouped together in disjoint sets, and the probability of being in a state in one set is independent of the previous state as long as that state lies in another set. The state filtering problem for lumpable HMPs was studied by White, Mahony, and Brushe [314].

Asymptotic performance of the MAP symbol estimator of S_n from y^n, for an HMP with rare transitions, was studied by Khasminskii and Zeitouni [185]. They assumed a finite-state irreducible aperiodic Markov chain with transition matrix $\{a_{ij}\}$ where $a_{ij} = \epsilon\lambda_{ij}$ when $i \neq j$, $a_{ii} = 1 - \epsilon\lambda_{ii}$, and $\lambda_{ii} = \sum_{j \neq i} \lambda_{ij}$. Let $\{\pi_i\}$ denote the stationary distribution of the chain and let $d_{ij} > 0$ denote the divergence between the observation conditional densities associated with states i and j where $i \neq j$. Let $\hat{P}_e^{\varepsilon,n}$ denote the infimum of $P(\tilde{S}_n \neq S_n)$ over all possible estimators $\tilde{S}_n$. Under some mild assumptions on the observation conditional densities of the HMP, and for any initial distribution, they showed that as $\epsilon \to 0$

$$\lim_{n\to\infty} \hat{P}_e^{\varepsilon,n} = (1 + o(1))\varepsilon \log\left(\epsilon^{-1}\right) \sum_i \pi_i \sum_{j \neq i} \frac{\lambda_{ij}}{d_{ji}}. \tag{5.3}$$

A similar result holds for a continuous-time HMP. When the states are considered real numbers $\{\nu_1, \ldots, \nu_M\}$, Golubev [143] showed under similar assumptions as in [185] that the average MSE

$$R_n^\epsilon = \frac{1}{n} \sum_{t=1}^{n} E\left\{(S_t - S_t^*)^2\right\} \tag{5.4}$$

associated with the conditional mean estimator S_t^* is given by

$$\lim_{n\to\infty} R_n^\epsilon = (1 + o(1))\varepsilon \log(\epsilon^{-1}) \sum_i \pi_i \sum_{j \neq i} \frac{\lambda_{ij}}{d_{ji}} (\nu_i - \nu_j)^2 \tag{5.5}$$

as $\epsilon \to 0$.

Asymptotically optimal recursive estimators for the states of a finite-alphabet HMP, in the minimum probability of error sense, that do not depend on the transition matrix, were derived in [142], [184], and [185].

A. Prediction, Filtering, and Smoothing Recursions

In this subsection, we present recursions for the conditional probabilities $p(s_t|y^{t-1})$, $p(s_t|y^t)$ and $p(s_t|y^n)$, $n > t$.

We begin with the forward–backward recursions of Chang and Hancock [56, eqs. (10), (18)] which were later rediscovered by Baum, Petrie, Soules, and Weiss [28], [29]. These recursions rely on the conditional statistical independence of Y^t and Y_{t+1}^n given s_t, $t = 1, \ldots, n-1$. This property leads to a simple decomposition of $p(s_t, y^n)$. Define the *forward* density by $\alpha(s_t, y^t) = p(s_t, y^t)$ and the *backward* density by $\beta(y_{t+1}^n|s_t) = p(y_{t+1}^n|s_t)$ with $\beta(y_{n+1}^n|s_t) = 1$. For $t = 1, \ldots, n$ we have

$$\begin{aligned} p(s_t, y^n) &= p(s_t, y^t, y_{t+1}^n) \\ &= p(s_t, y^t)p(y_{t+1}^n|s_t) \\ &= \alpha(s_t, y^t)\beta(y_{t+1}^n|s_t). \end{aligned} \tag{5.6}$$

The forward and backward densities satisfy the following recursions:

$$\alpha(s_t, y^t) = \begin{cases} \pi_{s_1} b(y_1|s_1), & t = 1 \\ b(y_t|s_t) \sum\limits_{s_{t-1}=1}^{M} \alpha(s_{t-1}, y^{t-1}) a_{s_{t-1}s_t}, & \\ & t = 2, \ldots, n. \end{cases} \tag{5.7}$$

$$\beta(y_{t+1}^n|s_t) = \begin{cases} 1, & t = n \\ \sum\limits_{s_{t+1}=1}^{M} \beta(y_{t+2}^n|s_{t+1}) a_{s_t s_{t+1}} b(y_{t+1}|s_{t+1}), & \\ & t = n-1, \ldots, 1. \end{cases} \tag{5.8}$$

The conditional probability $p(s_t|y^n)$, $t = 1, \ldots, n$, can be calculated as

$$p(s_t|y^n) = \frac{\alpha(s_t, y^t)\beta(y_{t+1}^n|s_t)}{\sum\limits_{s_t=1}^{M} \alpha(s_t, y^t)\beta(y_{t+1}^n|s_t)}. \tag{5.9}$$

Furthermore, for $t = 2, \ldots, n$

$$\begin{aligned} &p(s_{t-1}, s_t|y^n) \\ &= \frac{\alpha(s_{t-1}, y^{t-1})\beta(y_{t+1}^n|s_t)a_{s_{t-1}s_t}b(y_t|s_t)}{\sum\limits_{s_{t-1}, s_t=1}^{M} \alpha(s_{t-1}, y^{t-1})\beta(y_{t+1}^n|s_t)a_{s_{t-1}s_t}b(y_t|s_t)}. \end{aligned} \tag{5.10}$$

The likelihood function of the observation sequence y^n can be efficiently calculated using the forward recursion as follows:

$$p(y^n) = \sum_{s_n=1}^{M} \alpha(s_n, y^n). \tag{5.11}$$

Evaluation of (5.11) requires an order of nM^2 operations while direct calculation of the likelihood function (4.3) requires an order of nM^n operations.

The forward–backward recursions can be compactly written using matrix notation. Let f_t denote the $1 \times M$ vector whose jth element is $\alpha(j, y^t)$. Let b_t denote the $1 \times M$ vector whose jth element is $\beta(y_{t+1}^n|j)$. Let B_t denote an $M \times M$ diagonal matrix whose (j, j) element is $b(y_t|S_t = j)$. Let $\mathbf{1}$ represent an $M \times 1$ vector of 1's. Recall that A denotes the transition matrix and π denotes a $1 \times M$ vector representing the initial distribution. Let $f_1 = \pi B_1$ and $b_n = \mathbf{1}'$. The matrix forms of (5.7), (5.8) and (5.11), respectively, are given by $f_t = f_{t-1}AB_t$, $b_t = b_{t+1}B_{t+1}A'$, and $p(y^n) = f_n\mathbf{1}$. In particular, we have

$$\begin{aligned} p(y^n) &= \pi \prod_{t=1}^{n} (B_t A)\,\mathbf{1} \\ &= \pi B_1 A B_2 A \cdots B_{n-1} A B_n \mathbf{1}. \end{aligned} \tag{5.12}$$

It is well known that the forward–backward recursions (5.7) and (5.8) are not numerically stable. This has often been observed in practice, see, e.g., [215], [263], [81]. These observations are supported by the ergodic theorem (4.39) as argued by Leroux [212]. For t sufficiently large

$$p(y^t; \phi) = \sum_{s_t=1}^{M} \alpha(s_t, y^t) \approx \exp(t\overline{H}(P_{\phi^0}, P_\phi))$$

with high probability. Furthermore, $\{\alpha(s_t, y^t)\}_{s_t=1}^M$ are typically of the same order of magnitude. Hence, each $\alpha(s_t, y^t)$ tends to zero or infinity exponentially fast as $t \to \infty$.

An embedded iterative scaling procedure for stabilizing the forward–backward recursions was developed by Levinson, Rabiner, and Sondhi [215]. They proposed using

$$N_t = \sum_{s_t=1}^{M} \alpha(s_t, y^t)$$

as a normalizing factor for the forward and backward densities. Starting with a normalized forward density function, say $\overline{\alpha}(s_{t-1}, y^{t-1})$, the recursion (5.7) is executed and normalized by N_t to produce a new normalized updated density $\overline{\alpha}(s_t, y^t)$. For $t = 1$, we have $\overline{\alpha}(s_1, y_1) = p(s_1|y_1)$. Similarly, starting with a normalized backward density function, say $\overline{\beta}(y_{t+2}^n|s_{t+1})$, the recursion (5.8) is executed and normalized by N_{t+1} to produce a new normalized updated density $\overline{\beta}(y_{t+1}^n|s_t)$. For $t = n$, we have $\overline{\beta}(y_{n+1}^n|s_n) = 1$. The conditional probabilities (5.9) and (5.10) may be calculated using the scaled forward and backward densities. Devijver [81, eq. (17)] showed that the scaled forward recursion provides a stable recursion for the conditional probability $p(s_t|y^t)$. The scaled backward recursion does not enjoy such an intuitive interpretation. The recursion for $\overline{\beta}(y_{t+1}^n|s_t)$ is, in fact, a recursion for $p(y_{t+1}^n|s_t)/p(y_{t+1}^n|y^t)$ [81, eqs. (9), (16)]. Furthermore, the state conditional probability $p(s_t|y^n)$ can be obtained from

$$p(s_t|y^n) = \overline{\alpha}(s_t, y^t)\overline{\beta}(y_{t+1}^n|s_t). \tag{5.13}$$

Similar stable recursions were later developed for turbo codes by Barrou, Glavieux, and Thitimajshima [32].

The stable forward recursion for $p(s_t|y^t)$ was provided much earlier than [215] and [81] by Ott [248, eq. (4)], Raviv [265, eqs. (5), (8)], and Lindgren [219, Lemma 2.1]. Denoting $\alpha(s_t|y^t) = p(s_t|y^t)$, this recursion is given by

$$\alpha(s_t|y^t) = \frac{p(s_t|y^{t-1})b(y_t|s_t)}{\sum\limits_{s_t=1}^{M} p(s_t|y^{t-1})b(y_t|s_t)}, \qquad t = 1, \ldots, n \tag{5.14}$$

where $p(s_1|y_1^0) = \pi_{s_1}$ and

$$p(s_t|y^{t-1}) = \sum_{s_{t-1}=1}^{M} a_{s_{t-1}s_t}\alpha(s_{t-1}|y^{t-1}), \qquad t = 2, \ldots, n. \tag{5.15}$$

Lindgren [219, Lemma 2.1] and Askar and Derin [15, Theorem 1] developed equivalent alternative stable backward recursions for calculating $p(s_t|y^n)$ and $p(s_t, s_{t+1}|y^n)$ using only the forward recursions for $\alpha(s_t|y^t)$ and $p(s_t|y^{t-1})$. See also comments in Devijver [81, eq. (21)]. We present here the recursions from [15, Theorem 1] as follows:

$$p(s_t|y^n) = \alpha(s_t|y^t) \sum_{s_{t+1}=1}^{M} \frac{a_{s_t s_{t+1}} p(s_{t+1}|y^n)}{p(s_{t+1}|y^t)} \tag{5.16}$$

$$p(s_t, s_{t+1}|y^n) = \alpha(s_t|y^t) \frac{a_{s_t s_{t+1}} p(s_{t+1}|y^n)}{p(s_{t+1}|y^t)} \tag{5.17}$$

for $t = n-1, n-2, \ldots, 1$, where $p(s_n|y^n) = \alpha(s_n|y^n)$. The recursions (5.16), (5.17) are computationally more efficient than (5.9), (5.10) which use Chang and Hancock's forward–backward recursions [81].

The recursions for $p(s_t|y^{t-1})$ in (5.15), $\alpha(s_t|y^t)$ in (5.14), and $p(s_t|y^n)$ in (5.16) are, respectively, prediction, filtering, and fixed-interval smoothing recursions for estimating S_t from y^n. A recursion for the m-step predictor $p(s_t|y^{t-m})$, $m \geq 1$, can be found in Künsch [203, Lemma 3.1]. These recursions, with sums replaced by integrals, are applicable to discrete-time continuous-range state and observation processes, described by general state-space models of the form $S_t = g_t(S_{t-1}, V_t)$ and $Y_t = h_t(S_t, W_t)$ where g_t and h_t are arbitrary measurable functions and $\{V_t\}$ and $\{W_t\}$ are statistically independent i.i.d. processes [169, eq. (7.84)], [15], [189], [192], [203]. They provide conditional mean estimators for estimation problems that are not necessarily linear or Gaussian. For linear Gaussian state-space models with Gaussian initial conditions, the recursions (5.14)–(5.16) are equivalent to the Kalman filter and fixed-interval Kalman smoother, respectively, [169, Example 7.8], [189], [76], [203, Sec. 3.4.2]. This is easily checked since $p(s_t|y^{t-1})$, $\alpha(s_t|y^t)$, and $p(s_t|y^n)$ are Gaussian and hence characterized by their conditional means and covariance matrices [286, p. 308]. Exponential stability of the filtering and fixed-lag smoothing recursions in finite-alphabet HMPs was demonstrated by Anderson [11].

The stable forward–backward recursions have compact vector forms. Let $\xi_{t|\tau}$ denote the $M \times 1$ vector whose jth element is given by $P(S_t = j|y^\tau)$. Let η_t denote the $M \times 1$ vector whose jth element is given by $b(y_t|S_t = j)$. Let π denote, as usual, the $1 \times M$ vector of initial distribution. Let $\odot$ denote term-by-term multiplication of two vectors and let $(\div)$ denote term-by-term division of two vectors. The vector forms of (5.14) and (5.15) are, respectively, given by [156, eqs. 22.4.5–6],

$$\xi_{t|t} = \frac{\eta_t \odot \xi_{t|t-1}}{\eta_t' \xi_{t|t-1}}, \qquad t = 1, \ldots, n \tag{5.18}$$

where $\xi_{1|0} = \pi'$ and

$$\xi_{t|t-1} = A' \xi_{t-1|t-1}, \qquad t = 2, \ldots, n. \tag{5.19}$$

The vector form of (5.16) is given by [156, eq. 22.4.14]

$$\xi_{t|n} = \xi_{t|t} \odot \left\{ A \left[\xi_{t+1|n} (\div) \xi_{t+1|t} \right] \right\}, \qquad t = n-1, n-2, \ldots, 1. \tag{5.20}$$

The recursions (5.18)–(5.20) hold for the switching autoregressive process (4.10) of which HMPs are special cases, see, e.g., [156, Ch. 22].

We close this subsection with a relation that follows from Lindgren [219, Lemma 2.1]. We have that

$$\begin{aligned} p(s_t|s^{t-1}, y^n) &= p(s_t|s_{t-1}, y^n) \\ &= a_{s_{t-1}s_t} b(y_t|s_t) \frac{\beta(y_{t+1}^n|s_t)}{\beta(y_t^n|s_{t-1})}. \end{aligned} \tag{5.21}$$

This demonstrates the well-known fact that $\{S_t\}$ is a conditionally inhomogeneous Markov chain given $\{Y_t\}$. The transition probabilities are given by (5.21). This important property is often used in analysis of HMPs, see, e.g., [166], [36], and [174, Lemma 4.1]. Properties of the conditionally inhomogeneous Markov chain for switching autoregressive processes, with Markov regime in a separable compact state space that is not necessarily finite, were given by Douc, Moulines, and Rydén [91, Lemma 1].

VI. ML Parameter Estimation

In this section, we address several important aspects of parameter estimation of an HMP. We begin with conditions for identifiability of an HMP and proceed with consistency and asymptotic normality of the ML estimator. This is followed by a brief presentation of the Baum algorithm for local ML estimation of the parameter of an HMP. Next, Louis's formula for estimating the observed information matrix whose inverse provides an estimate of the error covariance of the ML estimator is presented. We conclude this section with Ziv's inequality which provides a tight upper bound on the maximum of the likelihood function of any finite-alphabet HMP.

A. *Identifiability of HMPs*

Consider a stationary HMP with the usual parametrization $\phi = \{\pi, A, \theta\} \in \Phi$ where $\theta = \{\theta_j, j = 1, \ldots, M\}$. Let $p(y^n; \phi)$ denote the n-dimensional density of the HMP. An HMP with true parameter $\phi^0 \in \Phi$ is said to be *identifiable* if for each $\phi \in \Phi$ such that $\phi \neq \phi^0$, $p(y^n; \phi) \neq p(y^n; \phi^0)$ a.e. for some $n > 0$ [274]. Consider the source–channel information-theoretic model of an HMP. If the Markov chain is reducible, there might be infinitely many stationary distributions. In addition, some components of the parameter of the HMP, related to the Markov chain and observation conditional densities, will have no effect on the likelihood function. Similarly, if some of the $\{\theta_j\}$ are identical, there might be an infinite number of stochastic matrices A that induce the same n-dimensional stationary distribution as A^0 does. In both cases, the HMP cannot be identifiable [214], [274]. Note that the states of the HMP can be permuted without affecting its distribution. This trivial ambiguity can be resolved if the states are ordered.

Leroux [214] and Rydén [274], [277] studied identifiability of a general HMP. Leroux observed that the problem is essentially that of identifiability of finite mixtures of product densities, since from (4.8), the density of the HMP can be written as

$$p(y^n;\,\phi) = \sum_{s^n} p(s^n;\,\phi) \prod_{t=1}^{n} b(y_t;\,\theta_{s_t}) \tag{6.1}$$

where $p(s^n;\,\phi) = \pi_{s_1} \prod_{t=2}^{n} a_{s_{t-1}s_t}$. Leroux invoked a result by Teicher [297, Theorem 2] which shows that if the family of all finite mixtures of $\{b(\cdot;\,\theta_j),\, j = 1, \ldots, M\}$ is identifiable, then for every $n \geq 2$, the family of finite mixtures of product densities of the form (6.1) is identifiable. The family of finite mixtures of $\{b(\cdot;\,\theta_j),\, j = 1, \ldots, M\}$ is identifiable if the mixing coefficients can be identified, i.e., if

$$\sum_{j=1}^{M} c_j b(y;\,\theta_j) = \sum_{j=1}^{M} \tilde{c}_j b(y;\,\tilde{\theta}_j) \qquad P_{\phi^0}\text{-a.s.}$$
$$\Rightarrow \sum_{j=1}^{M} c_j \delta_{\theta_j} = \sum_{j=1}^{M} \tilde{c}_j \delta_{\tilde{\theta}_j} \tag{6.2}$$

where δ_{θ_j} denotes the point mass at θ_j, and $\{c_j\}$ and $\{\tilde{c}_j\}$ are distributions. This condition holds, for example, when $b(\cdot;\,\theta_j)$ is Poisson, Gaussian with fixed variance, exponential, and negative exponential. Teicher's theorem combined with the earlier comments lead to the following conclusion. An HMP with the usual parametrization is identifiable for $n \geq 2$ if the Markov chain is irreducible, all $\{\theta_1, \ldots, \theta_M\}$ are distinct, and finite mixtures of the parametric family $\{b(\cdot;\,\theta_j)\}$ are identifiable. In that case, $\phi^0 = \{A^0,\, \theta^0\}$ is uniquely determined from y^n, $n \geq 2$, up to permutations of the states. It should be noted that the finite-dimensional distributions of $\{Y_t\}$ are uniquely determined by the $2M$-dimensional distribution even when not all of the $\{\theta_j\}$ are distinct, Rydén [274], [277, Theorem 1].

Conditions for identifiability of a Markov-modulated Poisson process, defined in Section IV-B7, were given by Rydén [278, Corollary 1]. A Markov-modulated Poisson process is identifiable, up to state permutations, if and only if all Poisson rates are distinct.

Petrie [251, Theorem 1.3] provided conditions for identifiability, up to permutations of the states, of a stationary ergodic finite-alphabet HMP; see also Finesso [116, Theorem 1.4.1]. A complete solution to the identifiability problem of a deterministic function of discrete-time, possibly nonstationary, Markov chain, was developed by Ito, Amari, and Kobayashi [167]. An algebraic approach was used to develop necessary and sufficient conditions for two aggregated Markov processes to be equivalent, i.e., to have equal finite-dimensional distributions. An algorithm for deciding equivalence was also developed. This approach was used by Rydén [278] and Larget [206] to determine equivalence of two continuous-time aggregated Markov processes. These are deterministic functions of continuous-time Markov chains. A unique canonical representation of each equivalence class of aggregated Markov processes that satisfy some mild regularity conditions was developed in [206] for both continuous-time and discrete-time processes. This representation contains a minimal parametrization of all identifiable information for the equivalence class. Equivalence of aggregated Markov processes may be checked in a single direct computation by converting the standard representation of the process to its canonical representation [206].

B. Consistency and Asymptotic Normality

Suppose that an observation sequence y^n was generated by an identifiable HMP with true parameter $\phi^0 \in \Phi$. Let $L_n(\phi) = \log p(y^n;\,\phi)$ denote the log-likelihood function of the HMP where $p(y^n;\,\phi)$ is given by (4.8) for any $\phi \in \Phi$. The ML estimator of ϕ^0 is obtained from

$$\hat{\phi}(n) = \arg\max_{\phi \in \Phi} L_n(\phi). \tag{6.3}$$

This maximization is performed over all $\phi \in \Phi$ such that $\pi(\phi)$ is a distribution, $A(\phi)$ is a stochastic matrix, and $\{\theta_j(\phi)\}$ satisfy appropriate constraints implied by the nature of the observation conditional densities of the HMP. The additional constraint $\pi A = \pi$ must be imposed when $\pi(\phi)$ represents a stationary distribution of the Markov chain. This constraint, however, significantly complicates the maximization problem and is usually ignored since the effect of π is asymptotically negligible as we have seen in Section IV-C3.

An estimator $\hat{\phi}(n)$ of ϕ^0 is said to be *strongly consistent* if

$$\lim_{n\to\infty} \hat{\phi}(n) = \phi^0 \qquad P_{\phi^0}\text{-a.s.} \tag{6.4}$$

Convergence in (6.4) is interpreted in the quotient topology generated by $\sim$. This means that any open subset $G \subset \Phi_{\rm c}$ which contains the equivalence class of ϕ^0 must also contain the equivalence class of $\hat{\phi}(n)$ for n sufficiently large P_{ϕ^0}-a.s. [214]. For an identifiable HMP with parameter ϕ, the equivalence class of the parameter comprises all points in $\Phi_{\rm c}$ induced by permutations of the states of the HMP. The equivalence relation $\sim$ was defined in Section IV-A, and the compactified parameter set $\Phi_{\rm c}$ was defined in Section IV-D.

Strong consistency of the ML estimator $\hat{\phi}(n)$ of the parameter of a finite-alphabet stationary ergodic HMP was proved by Baum and Petrie [25, Theorem 3.4] and by Petrie [251, Theorem 2.8]. Petrie relaxed the assumption that $\phi \in \Phi_\delta$ made in [25]. Strong consistency of the ML estimator $\hat{\phi}(n)$ of the parameter of a general stationary ergodic HMP was proved by Leroux [214, Theorem 3]. He assumed an irreducible aperiodic Markov chain and observation conditional densities that satisfy the mild regularity conditions noted in Section IV-D.

Consistency of the ML estimator was also proved for several extensions of standard HMPs under suitable conditions. In each case, consistency was shown using the corresponding ergodic theorem from Section IV-D. Strong consistency of the ML estimators of the parameters of switching autoregressive processes satisfying (4.14) and (4.15), respectively, was proved by Frencq and Roussignol [127, Theorem 3] and Krishnamurthy and Rydén [198, Theorem 1]. Recall that for a switching autoregressive process, the ML estimator is obtained from maximization of the conditional likelihood function noted in Section IV-D. Weak consistency of the ML estimator of the parameter of an HMP, with a separable compact state space that is not necessarily finite, was proved by Douc and Matias [90, Theorem 2]. The result applies to an HMP with arbitrary initial density, not necessarily a stationary density. Strong consistency of the ML estimator of the parameter of a switching autoregressive process, with a separable compact state space that is not necessarily finite, was proved by Douc, Moulines, and Rydén [91, Theorem 1]. The switching autoregressive process need not be stationary. Strong consistency of the ML estimator of the parameter of a Markov-modulated Poisson process was proved by Rydén [273, Theorem 1].

Consistency of the ML estimator of the parameter of a finite-alphabet HMP, when observations are drawn from a stationary ergodic process that is not necessarily the HMP, was proved by Finesso [116, Theorem 2.2.1]. This situation is described in the last paragraph of Section IV-D. The parameter of the HMP was assumed to satisfy $\phi \in \Phi_\delta$. Almost sure convergence of the set of maximizers of $L_n(\phi)$ over $\phi \in \Phi_\delta$, to the set of parameter values $\{\phi \in \Phi_\delta\}$ that minimize the relative entropy rate $\overline{D}(P_Y \| P_\phi)$ between the observation process (P_Y) and the HMP (P_ϕ), was proved. The relative entropy rate is defined similarly to (4.46) with Q replaced by P_ϕ.

We turn now to asymptotic normality of the ML estimator $\hat{\phi}(n)$. Assume that $\hat{\phi}(n)$ is consistent. Asymptotic normality of the ML estimator $\hat{\phi}(n)$ of the parameter of a stationary ergodic finite-alphabet HMP was proved in 1966 by Baum and Petrie [25] assuming that $\phi \in \Phi_\delta$. Asymptotic normality of the ML estimator $\hat{\phi}(n)$ of the parameter of a stationary ergodic general HMP was proved in 1998 by Bickel, Ritov, and Rydén [36, Theorem 1]. They assumed an irreducible aperiodic Markov chain and observation conditional densities that satisfy some mild regularity conditions. They showed that

$$n^{1/2}\left(\hat{\phi}(n)-\phi^0\right) \to \mathcal{N}\left(0, \mathcal{I}_{\phi^0}^{-1}\right) \quad P_{\phi^0}\text{-weakly as } n \to \infty \tag{6.5}$$

where $\mathcal{I}_{\phi^0}$ is the Fisher information matrix which is assumed nonsingular. This matrix is defined in terms of the score function by [36, eqs. (5) and (6), Lemma 6]

$$\mathcal{I}_{\phi^0} = E_{\phi^0}\{ZZ'\}, \qquad \text{where } Z = \lim_{n\to\infty} D_\phi \log p(Y_1 | Y_{-n}^0;\, \phi)_{|_{\phi=\phi^0}}. \tag{6.6}$$

The ML estimator $\hat{\phi}(n)$ is therefore asymptotically efficient in the sense of Lehmann [211, p. 404]. The crux of the proof in [36] is in establishing a central limit theorem for the score function $D_\phi \log p(Y^n;\, \phi)$ and a law of large numbers for the observed information $-D_\phi^2 \log p(Y^n;\, \phi)$. The proof then follows from the classical approach introduced by Cramér. In proving the limit theorems, the Markov chain $\{S_t\}$ given the observation sequence $\{Y_t\}$ is seen as an inhomogeneous Markov chain, see, e.g., (5.21), and its mixing coefficients are bounded in terms of $\{Y_t\}$.

Asymptotic normality of the ML parameter estimator of a general HMP, using geometric ergodicity of an extended Markov chain, follows from the work of Le Gland and Mevel [210] as described in Section IV-C3. Asymptotic normality of the ML parameter estimator of a general HMP with a separable compact state space that is not necessarily finite, was proved by Jensen and Petersen [174, Theorem 3.3] and by Douc and Matias [90, Theorem 3]. Jensen and Petersen assumed a stationary ergodic HMP and followed the proof of Bickel, Ritov, and Rydén [36]. Douc and Matias relaxed the stationarity assumption by following the approach of Le Gland and Mevel [210]. Asymptotic normality of the conditional ML parameter estimator, of a possibly nonstationary switching autoregressive process, with a separable compact state space that is not necessarily finite, was proved by Douc, Moulines, and Rydén [91, Theorem 4] following the approach of Bickel, Ritov, and Rydén [36].

Asymptotic normality of the ML estimator of the parameter of a general HMP was established only recently in [36] after being an open problem for over 30 years. Local asymptotic normality of an ML estimator defined on a grid of the parameter set was shown in [35]. Consistency and asymptotic normality of a pseudo ML parameter estimator of a stationary ergodic general HMP were proved by Lindgren [219] and Rydén [274]. The estimator maximizes a pseudo likelihood function obtained under the assumption that consecutive blocks of m consecutive observations are statistically independent. This likelihood function is given by

$$q(y^{nm};\, \phi) = \prod_{t=1}^{n} p(y_{m(t-1)+1}^{mt};\, \phi) \tag{6.7}$$

where $p(\cdot;\, \phi)$ is the density of y^m given by (4.3). For an identifiable HMP, any $m \geq 2$ can be chosen. Rydén refers to this estimator as the *maximum split data likelihood estimator* (MSDLE). For an HMP with irreducible aperiodic Markov chain that satisfies some regularity conditions, the MSDLE is consistent and asymptotic normal for fixed m and $n \to \infty$, and it performs as good as the ML estimator [274]. Lindgren [219] used $m = 1$ but did not consider estimation of the transition matrix. Francq and Roussignol [126] specialized these results to HMPs of the form $Y_t = S_t W_t$ described in Section IV-B1. A similar MSDLE estimator was proposed by Rydén [276] for estimating the parameter of a Markov-modulated Poisson process, and proved to be consistent and asymptotically normal. Asymptotic block i.i.d. approximation of the HMP likelihood function was also found useful in [238].

C. The Baum Algorithm

The Baum algorithm is a computationally efficient iterative algorithm for local maximization of the log-likelihood function $L_n(\phi)$ in (6.3). It was developed and proved to converge by Baum, Petrie, Soules, and Weiss [28], [29] in the early 1970s. It is the expectation–maximization (EM) algorithm of Dempster, Laird, and Rubin [80] when applied to HMPs. In this section, we present the Baum algorithm, discuss its relation to the EM algorithm, and provide conditions for local convergence. We assume a general HMP with the usual parametrization.

The rationale of the Baum algorithm is as follows [28, Theorem 2.1]. Suppose that an estimate $\phi_m \in \Phi$ of the parameter ϕ^0 is available at the end of the mth iteration. Let $\hat{\phi} \in \Phi$ denote some other estimate of ϕ^0. Define an *auxiliary function* for the given observation sequence y^n and any pair of parameters $\hat{\phi}$ and ϕ_m in Φ as follows:

$$Q\left(\hat{\phi},\, \phi_m\right) = E_{\phi_m}\left\{ \log p\left(S^n,\, y^n;\, \hat{\phi}\right) \Big| \, y^n \right\}. \tag{6.8}$$

Using Jensen's inequality

$$\begin{aligned} L_n\left(\hat{\phi}\right) - L_n(\phi_m) &= \log \frac{p\left(y^n;\, \hat{\phi}\right)}{p(y^n;\, \phi_m)} \\ &= \log E_{\phi_m}\left\{ \frac{p\left(S^n,\, y^n;\, \hat{\phi}\right)}{p(S^n,\, y^n;\, \phi_m)} \middle| y^n \right\} \\ &\ge E_{\phi_m}\left\{ \log \frac{p\left(S^n,\, y^n;\, \hat{\phi}\right)}{p(S^n,\, y^n;\, \phi_m)} \middle| y^n \right\} \\ &= Q\left(\hat{\phi},\, \phi_m\right) - Q(\phi_m,\, \phi_m) \qquad (6.9) \end{aligned}$$

where expectations are taken over S^n given y^n. Equality in (6.9) holds if and only if

$$p(S^n,\, y^n;\, \hat{\phi}) = p(S^n,\, y^n;\, \phi_m) \quad P_{\phi_m}\text{-a.e.}$$

A new estimate of ϕ^0 at the $m+1$ iteration is obtained from

$$\phi_{m+1} = \arg\max_{\hat{\phi}\in\Phi} Q\left(\hat{\phi},\, \phi_m\right). \qquad (6.10)$$

Since $\phi_m \in \Phi$, the procedure results in $L_n(\phi_{m+1}) \ge L_n(\phi_m)$ as can be seen from (6.9). When $\phi_{m+1} \sim \phi_m$, a *fixed point* is reached and $L_n(\phi_{m+1}) = L_n(\phi_m)$. The Baum algorithm starts with an initial estimate ϕ_0 and alternates between (6.8) and (6.10) until a fixed point is reached or some other stopping criterion is met.

Let $\varphi\colon \phi_m \rightarrow \phi_{m+1}$ denote the mapping defined by (6.8) and (6.10). Baum, Petrie, Soules, and Weiss [28, Theorem 3.1] and Baum [29] showed that if $\log b(y_t;\, \theta_j)$ is strictly concave in θ_j for each y_t and all j, then $\varphi(\phi_m)$ is a single-valued continuous mapping, and $L_n(\phi_{m+1}) > L_n(\phi_m)$ unless ϕ_m is a stationary point of $L_n(\phi)$ or equivalently a fixed point of $\varphi(\phi_m)$. Furthermore, all limit points of $\varphi(\phi_m)$ are stationary points of L_n [28, Proposition 2.1]. The log-concavity condition holds for normal, Poisson, binomia,l and gamma distributions among others, but it fails for the Cauchy distribution. Liporace [221] extended these results to elliptically symmetric multivariate densities which essentially are mixtures of Gaussian densities of which the Cauchy density is a special case.

The Baum algorithm is a particular instance of the EM algorithm of Dempster, Laird, and Rubin [80]. The expectation step (E-step) is given by (6.8) and the maximization step (M-step) by (6.10). In the EM terminology, the state and observation sequences $\{s^n,\, y^n\}$ are the *complete* data while the observation sequence y^n alone is the *incomplete* data. The likelihood function is written as

$$L_n(\phi) = Q(\phi,\, \phi_m) - H(\phi,\, \phi_m) \qquad (6.11)$$

where

$$H(\phi,\, \phi_m) = E_{\phi_m}\{\log p(S^n \,|\, y^n;\, \phi) \,|\, y^n\}. \qquad (6.12)$$

A well-known consequence of Jensen's inequality is that $H(\phi,\, \phi_m) \le H(\phi_m,\, \phi_m)$ for any ϕ and ϕ_m in Φ with equality if and only if $p(S^n|y^n;\, \phi) = p(S^n|y^n;\, \phi_m)$ a.s. [80, Lemma 1]. Hence, $L_n(\phi_{m+1}) \ge L_n(\phi_m)$ if ϕ_{m+1} maximizes $Q(\phi,\, \phi_m)$ over $\phi \in \Phi$.

Convergence of the EM algorithm was established by Wu [316] using the global convergence theorem [226, p. 187]. In particular, it was assumed that i) the level set $\{\phi \in \Phi\colon L_n(\phi) \ge L_n(\phi_0)\}$ is compact for any $\phi_0 \in \Phi$ with $L_n(\phi_0) > -\infty$; ii) $L_n(\phi)$ is continuous in Φ and differentiable in the interior of Φ; iii) $Q(\phi,\, \tilde{\phi})$ is continuous in both ϕ and $\tilde{\phi}$; and iv) all EM instances $\{\phi_m\}$ are in the interior of Φ. Under these conditions, it was shown in [316, Theorem 2] that all the limit points of any instance $\{\phi_m\}$ of the EM algorithm are stationary points of $L_n(\phi)$, and $L_n(\phi_m)$ converges monotonically to $L_n^* = L_n(\phi^*)$ for some stationary point ϕ^*. There exists at least one such limit point. The compactness assumption may be restrictive when no realistic compactification of the original parameter space is possible. Continuity of $Q(\phi,\, \tilde{\phi})$ is satisfied in most practical situations. It is guaranteed for the important family of exponential (Koopman–Darmois) pdfs [211, p. 26], [316]. Wu [316] provided conditions for other convergence theorems, in particular, convergence of limit points of $\{\phi_m\}$ to local maxima of $L_n(\phi)$.

The strict maximization of $Q(\phi,\, \phi_m)$ over ϕ in (6.10) is relaxed in the *generalized EM* algorithm. Any ϕ that satisfies the weaker condition of $Q(\phi,\, \phi_m) > Q(\phi_m,\, \phi_m)$ is admissible. Conditions for local convergence of the generalized EM algorithm were given in Wu [316, Theorem 1]. The generalized EM algorithm was found useful in estimating the parameter of a Markov-modulated Poisson process [273]. An EM algorithm with an explicit M-step for estimating the parameter of a Markov-modulated Poisson process was developed by Rydén [275].

Note that in Section VI-B we were concerned with convergence of the ML estimate sequence $\hat{\phi}(n)$ to the true parameter ϕ^0 when the number of observations $n \rightarrow \infty$. Consistency theorems were provided for observation sequences generated by the HMP or by any other stationary ergodic process in the case of a finite-alphabet HMP. In this section, we considered convergence of an instance of the Baum algorithm, $\{\phi_m\}$, for fixed n and observation sequence y^n, when the iteration number $m \rightarrow \infty$. In this discussion, the observation sequence y^n need not be generated by the HMP as the EM algorithm can be applied to any observation sequence. When y^n is generated by an HMP with parameter ϕ^0, convergence of an EM instance $\{\phi_m(n)\}$ as $m,\, n \rightarrow \infty$ may not be to the ML estimate of ϕ^0, since only local convergence is guaranteed.

1) The Re-Estimation Formulas: Maximization of the auxiliary function $Q(\phi,\, \phi_m)$ in (6.8) for a given observation sequence y^n results in re-estimation formulas for the parameter of the HMP. They generate a new parameter estimate from an old parameter estimate. To demonstrate how the Baum algorithm works, we shall provide here the re-estimation formulas for two important HMPs, those with Gaussian and Poisson observation conditional densities. In both cases, maximization of (6.8) over ϕ for a given ϕ_m results in an explicit estimate ϕ_{m+1} at the end of the $m+1$st iteration. The re-estimation formulas require the conditional probabilities $p(s_t|y^n;\, \phi_m)$ and $p(s_{t-1},\, s_t|y^n;\, \phi_m)$ which can be efficiently calculated as shown in Section V.

Using (4.8), the auxiliary function $Q(\phi, \phi_m)$ in (6.8) is written as [29]

$$\begin{aligned} Q(\phi, \phi_m) = & \sum_{j=1}^{M} P(S_1 = j | y^n;\, \phi_m) \log \pi_j \\ & + \sum_{i,j=1}^{M} \sum_{t=2}^{n} P(S_{t-1} = i, S_t = j | y^n; \phi_m) \log a_{ij} \\ & + \sum_{j=1}^{M} \sum_{t=1}^{n} P(S_t = j | y^n;\, \phi_m) \log b(y_t;\, \theta_j). \end{aligned} \tag{6.13}$$

Maximization of (6.13) over the distribution $\pi = \{\pi_j\}$ and the stochastic matrix $A = \{a_{ij}\}$ gives

$$\pi_j(m+1) = P(S_1 = j | y^n;\, \phi_m) \tag{6.14}$$

$$a_{ij}(m+1) = \frac{\sum_{t=2}^{n} P(S_{t-1} = i,\, S_t = j | y^n;\, \phi_m)}{\sum_{t=2}^{n} P(S_{t-1} = i | y^n;\, \phi_m)}. \tag{6.15}$$

These re-estimation formulas are intuitively appealing. The initial state probability estimate (6.14) is the conditional probability of the state given the observations. The estimate of the transition probability a_{ij} in (6.15) is the ratio of the Cesáro mean of the conditional probabilities of visiting state i and then j and the Cesáro mean of the conditional probabilities of visiting state i. The conditional probabilities are calculated under the current estimate of the HMP.

The stationary distribution of the Markov chain is commonly estimated as [219]

$$\pi_j = \frac{1}{n} \sum_{t=1}^{n} P(S_t = j | y^n;\, \phi_m). \tag{6.16}$$

For an HMP with Gaussian observation conditional densities, the re-estimation formula for the mean vector ς_j is given by

$$\varsigma_j(m+1) = \frac{\sum_{t=1}^{n} P(S_t = j | y^n;\, \phi_m) y_t}{\sum_{t=1}^{n} P(S_t = j | y^n;\, \phi_m)}. \tag{6.17}$$

The re-estimation formula for the covariance matrix R_j is given by (6.18) shown at the bottom of the page. For an HMP with Poisson observation conditional pmfs, the re-estimation formula for the mean parameter λ_j is given by (6.17).

D. Observed Information Matrix

Unlike the Kalman filter, the Baum algorithm does not provide the error covariance matrix of the estimated parameter in each iteration. While this matrix is an integral part of the Kalman recursion, it is not needed by Baum's re-estimation formulas. An estimate of this matrix can provide some idea about the quality of parameter estimates obtained by the Baum algorithm. The actual error covariance associated with the Baum algorithm is not known. For consistent ML estimation, however, it is known from (6.5) that the asymptotic error covariance is given by the inverse of the Fisher information matrix $\mathcal{I}_{\phi^0}$. An estimate of this matrix is given by the *observed information matrix* which is the negative Hessian matrix

$$I_Y = -D_\phi^2 \log p(y^n;\, \phi). \tag{6.19}$$

Under some mild regularity conditions, the observed information matrix of a stationary ergodic HMP was shown by Bickel, Ritov, and Rydén [36, Lemma 2] to be a consistent estimate of the Fisher information matrix. Specifically, for any consistent estimate $\hat{\phi}(n)$ of ϕ^0 it holds that

$$\lim_{n \to \infty} \frac{1}{n} \left[-D_\phi^2 \log p(Y^n;\, \phi) \right]_{\phi = \hat{\phi}(n)} = \mathcal{I}_{\phi^0} \quad \text{in } P_{\phi^0}\text{-probability.} \tag{6.20}$$

Louis [225] developed a formula for calculating I_Y from the complete data comprising the state and observation sequences. Let I_{SY} denote the complete data observed information matrix given by

$$I_{SY} = -D_\phi^2 \log p(S^n,\, y^n;\, \phi). \tag{6.21}$$

Let $I_{S|Y}$ denote the conditional complete data observed information matrix given by

$$I_{S|Y} = -D_\phi^2 \log p(S^n | y^n;\, \phi). \tag{6.22}$$

The formula is given by

$$I_Y = E_\phi\{I_{SY} | y^n\} - E_\phi\{I_{S|Y} | y^n\}. \tag{6.23}$$

The formula follows from a relation between the score function of the incomplete data $G_Y = D_\phi \log p(y^n;\, \phi)$ and the score function of the complete data $G_{SY} = D_\phi \log p(S^n,\, y^n;\, \phi)$. This relation is given by

$$G_Y = E_\phi\{G_{SY} | y^n\} \tag{6.24}$$

where expectation is over S^n given y^n. The second term in (6.23) can be written as

$$\begin{aligned} E_\phi\{I_{S|Y} | y^n\} = & E_\phi\{G_{SY} G'_{SY} | y^n\} \\ & - E_\phi\{G_{SY} | y^n\} E_\phi\{G'_{SY} | y^n\} \end{aligned} \tag{6.25}$$

which implies its nonnegative definiteness. Hence $I_Y - E_\phi\{I_{SY} | y^n\}$ in non-positive definite.

A method for calculating the observed information matrix of an HMP from (6.23) and (6.25) was proposed by Hughes [166]. The term $E_\phi\{I_{SY} | y^n\}$ was evaluated similarly to Baum's auxiliary function (6.13) using the forward–backward formulas. From (6.24), $E_\phi\{G_{SY} | y^n\} = 0$ for any local ML estimate of ϕ. The remaining term $E_\phi\{G_{SY} G_{SY} | y^n\}$ in (6.25) is the hardest to calculate since it involves double summations

$$R_j(m+1) = \frac{\sum_{t=1}^{n} P(S_t = j | y^n;\, \phi_m)[(y_t - \varsigma_j(m+1))(y_t - \varsigma_j(m+1))']}{\sum_{t=1}^{n} P(S_t = j | y^n;\, \phi_m)}. \tag{6.18}$$

of cross-product terms over pairs of states at distinct time instants. Hughes used the fact that the state sequence is a conditionally inhomogeneous Markov chain given the observation sequence, and provided some mild conditions for the sequence to be mixing with exponentially decreasing coefficients. This enabled dropping cross-product terms involving state variables that are well separated in time.

The observed information matrix was calculated using Monte Carlo simulations by Diebolt and Ip [88] for general EM applications and by Turner and Cameron [303] for HMPs.

E. Upper Bound on Likelihood of Finite-Alphabet HMPs

Algorithms for global maximization of the likelihood function $\log p(y^n;\,\phi)$ over $\phi \in \Phi$ are not known for most interesting HMPs. An upper bound on the global maximum of the likelihood function exists for any finite-alphabet HMP. The bound uses universal coding of the observation sequence and its nonvanishing term is independent of the number of states and the underlying parameter ϕ. The bound is tight with high probability and hence can be used to assess the closeness of $\log p(y^n;\,\hat{\phi})$ to the global maximum of $\log p(y^n;\,\phi)$ for any estimator $\hat{\phi}$.

The upper bound is provided by the Ziv inequality which was first derived for Markov chains in [329], see also [68, Lemma 12.10.3]. The bound was extended to finite-alphabet HMPs by Plotnik, Weinberger, and Ziv in [253, p. 68]. The bound is essentially given by $-u(y^n)$ where $u(y^n)$ is the length of the binary codeword for y^n in the Lempel–Ziv universal data compression algorithm [326]. This algorithm sequentially parses the sequence y^n into $c-1$ distinct phrases $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{c-1}$ of variable length, and an additional, possibly incomplete, phrase $\boldsymbol{z}_c$ that may coincide with one of the other $c-1$ phrases. Each phrase comprises a concatenation of a phrase that appeared previously in the sequence and an additional symbol that distinguishes the newly created phrase from any previously defined phrase. For example, the binary sequence $y^{20} = 10110001011101110000$ is parsed as $1, 0, 11, 00, 01, 011, 10, 111, 000, 0$ where $c = 10$ and the first nine phrases are distinct. The number of phrases c depends on the sequence y^n and may be expressed more explicitly as $c(y^n)$. The length of the codeword for y^n, or the number of bits required to represent y^n in the Lempel–Ziv algorithm, is given by $u(y^n) = c(y^n)[\log c(y^n) + 1]$. The algorithm asymptotically outperforms any finite-state coding scheme in compressing any individual sequence not necessarily from an HMP. It asymptotically achieves the entropy rate $\overline{H}(Y)$ in compressing any stationary ergodic finite-alphabet source $\{Y_t,\, t \geq 1\}$, i.e., $(1/n)u(Y^n) \to \overline{H}(Y)$ with probability 1 as $n \to \infty$ [326], see also [68, Theorem 12.10.2]. Lempel–Ziv is the standard compression algorithm in UNIX and operating systems for PCs.

The upper bound for any stationary ergodic finite-alphabet HMP with M states, L letters, and parameter ϕ^0, and for any observation sequence y^n, is given by [253, p. 68]

$$\frac{1}{n}\log p(y^n;\,\phi^0) \leq -\frac{c(y^n)}{n}\log\frac{c(y^n)}{M^2} + h\left(\frac{c(y^n)}{n}\right) \quad (6.26)$$

where $h(\cdot)$ denotes the binary entropy function given by

$$h(q) = -q\log q - (1-q)\log(1-q), \qquad \text{for } 0 \leq q \leq 1.$$

Since the number of phrases satisfies [326, eq. (4)], [68, Lemma 12.10.1]

$$\frac{c(y^n)}{n} \leq \frac{\log L}{(1-\varepsilon_n)\log n} \quad (6.27)$$

where $\varepsilon_n = \min\left\{1,\, \frac{\log(\log n)+4}{\log n}\right\}$, the bound can be written as

$$\frac{1}{n}\log p(y^n;\,\phi^0) \leq -\frac{1}{n}\,u(y^n) + \delta_n \quad (6.28)$$

where $\delta_n = O\left(\frac{\log(\log n)}{n}\right)$. Hence, $\delta_n \to 0$ uniformly for every y^n as $n \to \infty$. For $n \to \infty$, the bound becomes $-\overline{H}(Y)$. Since (6.28) holds for any $\phi^0 \in \Phi$, it also holds for the maximizing $\phi \in \Phi$ as follows:

$$\max_{\phi}\, \log p(y^n;\,\phi) \leq -u(y^n) + n\delta_n. \quad (6.29)$$

A partial converse to Ziv's inequality is obtained as follows. Let

$$\Psi = \left\{y^n\colon p(y^n;\,\phi^0) < 2^{-u(y^n)-n\epsilon}\right\} \quad (6.30)$$

for some $\epsilon > 0$. From the Kraft inequality [68, Sec. 5.2]

$$\begin{aligned} P(y^n \in \Psi) &= \sum_{y^n \in \Psi} p(y^n;\,\phi^0) \\ &< \sum_{y^n \in \Psi} 2^{-u(y^n)-n\epsilon} \\ &\leq 2^{-n\epsilon}\sum_{y^n} 2^{-u(y^n)} \leq 2^{-n\epsilon}. \end{aligned} \quad (6.31)$$

Hence $P(y^n \in \Psi^c) \geq 1 - 2^{-n\epsilon}$ and the probability that

$$\max_{\phi}\, \log p(y^n;\,\phi) \geq -u(y^n) - n\epsilon$$

approaches one as $n \to \infty$. Ziv's inequality was used in many applications including order estimation [330] and source coding [236] of finite-alphabet HMPs. These applications are reviewed in Sections VIII and XIV, respectively.

A stronger result holds for unifilar sources defined in Section IV-E. From the analog of (4.50) for unifilar sources

$$\max_{p(y|s)}\, \frac{1}{n}\log p(y^n|s_0) = -H(q_n) \leq -\frac{1}{n}\,u(y^n) + \delta_n \quad (6.32)$$

where $H(q_n)$ is the conditional empirical entropy defined in (4.51) for a given next-state function g. If g is not known, the left-hand side of (6.32) is maximized over g. There are $M^{L \cdot M}$ such functions for a unifilar source with M states and L letters [330].

VII. Joint State and Parameter Estimation

In this section, we review joint estimation of the state sequence and the parameter of an HMP. We first describe the Baum–Viterbi algorithm and its relations to the Baum algorithm and to the generalized Lloyd algorithm for designing vector quantizers. The relation of the Baum–Viterbi algorithm to the minimum discrimination information parameter estimation approach is given in Section XIV-B. We then present a noniterative algorithm for global maximization of the joint likelihood function of a left–right HMP. We conclude by reviewing Bayesian Gibbs sampling approaches.

A. The Baum–Viterbi Algorithm

The Baum–Viterbi algorithm jointly estimates the parameter and state sequence of an HMP. The state sequence is estimated in the minimum probability of error sense by the Viterbi algorithm. Recall that the Baum algorithm uses the state conditional probabilities in estimating the parameter. When the states are considered unit vectors in a Euclidean space, these conditional probabilities are the MMSE estimate of the state. The Baum–Viterbi algorithm was proven useful when the observations are vectors of sufficiently high dimension. In that case, the Baum–Viterbi algorithm provides parameter estimates that are almost as good as those obtained by the Baum algorithm. The algorithm has an intuitive appeal and is computationally more stable than the Baum algorithm. The two algorithms require about the same amount of computation.

When the observations of the HMP are scalar, or vectors of fixed dimension, say k, the Baum–Viterbi algorithm provides inconsistent estimates of the state sequence and parameter as the number of observations $n \to \infty$. This was shown in [51], [299] for mixture processes which are special cases HMPs. The asymptotic mode considered here of $k \to \infty$ and fixed n is motivated by applications in automatic speech recognition where HMPs with vector observations of relatively large dimensions are often used and estimation is performed from a fixed number of observations. The reason for using vector observations is that states representing articulatory cues mix at significantly lower rate than the sampling rate of the signal itself which is typically about 8000 Hz. Thus, the state process of a speech signal has significantly lower bandwidth than that of the signal itself.

The Baum–Viterbi algorithm was first introduced in 1976 by Jelinek and his colleagues at IBM [172] and was termed Viterbi extraction. The algorithm was further studied by Rabiner, Wilpon, and Juang [261], [262], [176], where it was referred to as segmental k-means. Asymptotic equivalence of parameter estimates obtained by the Baum algorithm and by the Baum–Viterbi algorithm for fixed n and $k \to \infty$ was shown by Merhav and Ephraim [234]. We opted for the name *Baum–Viterbi* since each iteration of the algorithm involves Baum's re-estimation iteration and application of the Viterbi algorithm.

Consider an HMP with vector observations $\{y_t\}$, $y_t \in \mathcal{R}^k$, and true parameter $\phi^0 \in \Phi$. The Baum–Viterbi algorithm estimates ϕ^0 from

$$\max_{\phi\in\Phi} \max_{s^n\in\mathcal{S}^n} p(s^n, y^n; \phi) \tag{7.1}$$

where the double maximization is alternately performed over s^n and ϕ. For a given parameter estimate $\phi_m \in \Phi$ at the end of the mth iteration, the most likely state sequence is estimated by maximizing $p(s^n, y^n; \phi_m)$ over s^n. This maximization is performed using the Viterbi algorithm. Let $s^n(\phi_m)$ denote the maximizing state sequence. Next, a new estimate ϕ_{m+1} of the parameter is obtained by maximizing $p(s^n(\phi_m), y^n; \phi)$ over $\phi \in \Phi$. The alternate maximization procedure produces a sequence of estimates $\{\phi_m\}$ with nondecreasing joint likelihood values. The algorithm is terminated if a fixed point is reached or when a stopping criterion is met. Local convergence of the algorithm can be established in a manner similar to that used for the EM algorithm [316], [176]. Note that a byproduct of the algorithm is an estimate of the most likely state sequence. This is analogous to the byproduct of conditional state probabilities given the observation sequence provided by the Baum algorithm.

Maximization of $p(s^n(\phi_m), y^n; \phi)$ over ϕ is equivalent to maximizing the auxiliary function

$$Q_1(\phi, \phi_m) = \sum_{s^n} \delta(s^n - s^n(\phi_m)) \log p(s^n, y^n; \phi) \tag{7.2}$$

where $\delta(\cdot)$ is the Kronecker delta function that is equal to one when $s^n = s^n(\phi_m)$ and is zero otherwise. Recall that in the Baum algorithm, a new estimate ϕ_{m+1} is obtained from maximization over ϕ of the auxiliary function

$$Q(\phi, \phi_m) = \sum_{s^n} p(s^n|y^n; \phi_m) \log p(s^n, y^n; \phi). \tag{7.3}$$

Comparing (7.2) with (7.3) shows that ϕ_{m+1} in the Baum–Viterbi algorithm can be obtained from the re-estimation formulas of the Baum algorithm by substituting $p(s^n|y^n; \phi_m)$ by $\delta(s^n - s^n(\phi_m))$. These formulas are given by (6.14), (6.15) and by (6.17), (6.18) for HMPs with Gaussian observation conditional densities. The re-estimation formulas for the Baum–Viterbi algorithm are rather intuitive. For example, the estimate for a_{ij} is the ratio between the number of transitions from state i to j and the number of transitions from state i to any other state on the most likely sequence $s^n(\phi_m)$. Similarly, the new estimate for the mean and covariance matrices of the Gaussian density in the jth state are obtained from sample averages of observation vectors assigned to state j by $s^n(\phi_m)$. Alternatively, the observation vectors in y^n are clustered into M subsets by the most likely sequence $s^n(\phi_m)$ and the parameter of the observation conditional densities are obtained from these clusters.

It follows from (7.2) and (7.3) that the Baum algorithm and the Baum–Viterbi algorithm yield the same sequence of estimates $\{\phi_m\}$ when started from the same initial estimate ϕ_0 if

$$p(s^n|y^n; \phi_m) = \delta(s^n - s^n(\phi_m)) \tag{7.4}$$

for every m. Convergence of $p(s^n|y^n; \phi_m)$ to $\delta(s^n - s^n(\phi_m))$ P_{ϕ^0}-a.s., when $k \to \infty$, was proved in [234]. It was assumed that the transition matrix satisfies $A \geq \delta > 0$, and an ergodic theorem holds for $k^{-1} \log b(Y_t|\theta_j(\phi))$ for any $\phi \in \Phi$. The required ergodic property was demonstrated in [234] for HMPs with Gaussian observation conditional densities. The required ergodic theorem under more general conditions is implied from (4.39). The result is not surprising since states are detectable when a sufficient number of consecutive observations is available from each state. When $k \to \infty$, the most likely state sequence $s^n(\phi)$ is given by

$$s_t(\phi) = \arg\max_j b(y_t; \theta_j(\phi)), \qquad t = 1, \ldots, n \tag{7.5}$$

for any parameter $\phi \in \Phi$.

Bounds on the log-likelihood difference resulting from (6.3) and (7.1) were derived in [234]. Let $\tilde{\phi}$ denote the maximizer over ϕ of $\max_{s^n} p(s^n, y^n; \phi)$ and let $\hat{\phi}$ denote the maximizer over

ϕ of $p(y^n; \phi)$. Let y^n denote a sequence of n k-dimensional observation vectors and let $N = nk$ denote the total number of observations. Then

$$0 \le \frac{1}{N} \log p(y^n; \hat{\phi}) - \frac{1}{N} \log \max_{s^n} p(s^n, y^n; \tilde{\phi}) \le \frac{1}{k} \log M \tag{7.6}$$

$$0 \le \frac{1}{N} \log p(y^n; \hat{\phi}) - \frac{1}{N} \log p(y^n; \tilde{\phi}) \le \frac{1}{k} \log M. \tag{7.7}$$

Thus the difference between the normalized likelihood values associated with $\hat{\phi}$ and $\tilde{\phi}$ can never exceed $(1/k) \log M$. This bound can be made sufficiently small compared to the likelihood values if $k \gg M$. This is often the situation in isolated-word speech recognition applications where typically $M = 5$–30 and $k = 256$–512 [234]. Note that these inequalities are not sufficient to guarantee closeness of the parameter estimates obtained by the Baum and the Baum–Viterbi algorithms, since both algorithms perform local rather than global maximization. Moreover, these inequalities do not imply that a dominant state sequence exists since they hold even when all state sequences are equally likely.

A theoretical approach for a sequential Baum–Viterbi algorithm was proposed by Kogan [191]. The approach is based on the observations that stopping times for the most likely state sequence appear infinitely often if the Markov chain is irreducible and aperiodic, and the most likely state sequence at time instants smaller than the stopping time is independent of future observations.

B. The Generalized Lloyd Algorithm

The Baum–Viterbi algorithm is closely related to the generalized Lloyd algorithm for designing vector quantizers for parametric processes [135], [234]. The generalized Lloyd algorithm is also known as the Linde–Buzo–Gray (LBG) algorithm [218]. A vector quantizer partitions the parameter set of a process into a finite number of cells, say M, and chooses a parameter representative from each cell. The design of vector quantizers requires a distortion measure that quantifies the similarity of one parameter with respect to another. In the context of this section, the distortion measure is between a vector y_t of the process, which has some underlying parameter, and a parameter $\theta_j \in \Theta$. A vector quantizer is designed by minimizing the expected value of the distortion measure over all partitions and parameter representatives. The generalized Lloyd algorithm performs this minimization iteratively, once over the partition for a given set of parameter representatives, and then over the parameter representatives using the estimated partition. The process proceeds until a fixed point is reached or otherwise a stopping criterion is satisfied. In practice, the expected value of the distortion measure is replaced by the sample mean of a training sequence of observations. Convergence properties of the generalized Lloyd algorithm were established by Sabin and Gray [283]. A comprehensive overview of quantization theory and its applications can be found in Gray and Neuhoff [153].

An important application of vector quantization is in coding of speech signals in cellular communication. The signal is modeled as an autoregressive process with a time-varying parameter. A finite number of parameter representatives is estimated and used in encoding the speech signal at a relatively low bit rate [135, pp. 387–393], [259, Sec. 10.4].

The relation of the Baum–Viterbi algorithm to the generalized Lloyd algorithm becomes clear when k is large and $-k^{-1} \log b(y_t; \theta_j)$ is interpreted as the distortion measure between the vector y_t and a parameter $\theta_j \in \Theta$. Almost sure convergence of $-k^{-1} \log b(Y_t; \theta_j)$ when $k \to \infty$ is implied from (4.39). It was demonstrated in [234] for HMPs with Gaussian observation conditional densities where explicit expressions for the limit were given. This distortion measure may take negative values but this does not affect the generalized Lloyd algorithm as long as the distortion is greater than $-\infty$. Let y^n denote a training sequence of n k-dimensional observation vectors. Assuming $A \ge \delta > 0$ and large k, the sample mean of the distortion measure is given by

$$\frac{1}{n} \sum_{t=1}^{n} -\frac{1}{k} \log b(y_t; \theta_{s_t}) \approx -\frac{1}{n} \frac{1}{k} \log p(s^n, y^n; \phi). \tag{7.8}$$

Estimation of $\{\theta_1, \ldots, \theta_M\}$ by the iterative Baum–Viterbi algorithm is equivalent to estimating these components of the parameter by minimizing the average distortion in the left-hand side of (7.8). The most likely state sequence (7.5) in the Baum–Viterbi algorithm provides the optimal partition or classification of the vectors $\{y_t\}$ in the generalized Lloyd algorithm. This, in turn, provides the optimal partition of the underlying parameter space of these vectors. This partition rule is referred to as the *nearest neighbor* rule in vector quantization terminology. Estimation of each θ_j by minimizing the average of the distortion measure over all vectors assigned to the jth state, as in the Baum–Viterbi algorithm, provides the best parameter representative in the generalized Lloyd algorithm. This estimate is referred to as the *centroid* of the partition cell in vector quantization terminology. Thus, each iteration of the Baum–Viterbi algorithm parallels an iteration of the generalized Lloyd algorithm. Note that the generalized Lloyd algorithm provides estimates of the parameter of the observation conditional densities only. An estimate of the transition matrix can be found from the nearest neighbor state sequence (7.5) as in the Baum–Viterbi algorithm.

C. Initialization of the Baum Algorithm

The likelihood function of an HMP may have multiple local maxima while the Baum algorithm converges at best to a local maximum in the neighborhood of the initial guess of the parameter. Local convergence was demonstrated in [95] for a binary HMP with a binary Markov chain. Initialization of the Baum algorithm has therefore a significant impact on the optimality of the parameter estimate.

Several initialization strategies were proposed. For HMPs with ordered states that are allowed self-transitions and next-state transitions only, commonly used in automatic speech recognition applications, it was suggested to segment the acoustic signal from each word into M segments of approximately equal length, and to estimate the parameter of each state from the observations in the corresponding segment [262]. For HMPs with $A > 0$, the generalized Lloyd algorithm may be used to cluster the observations into M sets from which the parameter of the HMP can be estimated. The generalized

Lloyd algorithm applies to scalar as well as vector observation processes. Similar clustering techniques for initialization of the Baum algorithm were proposed in [212] and [232, Sec. 1.7]. Simulated annealing may also be used as discussed in Section VII-E.

D. Global Likelihood Maximization for Left–Right HMPs

An HMP is said to be *left–right* if its transition matrix is an upper triangular matrix. An HMP is said to be *linear* if its transition matrix has nonzero entries only on the main diagonal and first off-diagonal. In this subsection, we present a noniterative algorithm for global maximization of $p(s^n, y^n; \phi)$ over s^n and ϕ for a left–right HMP. The algorithm was developed by Faregó and Lugosi [111] in 1989 for a finite-alphabet HMP but it applies to a general HMP as well. Parameter estimation for a left–right HMP can be reduced to parameter estimation of a linear HMP [111]. Hence it suffices to describe the algorithm for a linear HMP. The practical importance of left–right HMPs is discussed in Section XIV-A. The rationale for maximizing $p(s^n, y^n; \phi)$ was detailed in Section VII-A. The key idea of this algorithm is that the state sequence in a linear HMP is uniquely determined by the state occupancy durations. Global noniterative maximization is achieved by explicit estimation of the parameter of the HMP for a given state sequence, and substituting that estimate back in the likelihood function. The resulting likelihood function depends only on the M-state occupancy durations. This function is maximized by the Viterbi algorithm which is applied to a specially constructed trellis scheme.

Assume that the number of states M is smaller than the length of the observation sequence n; otherwise, the estimation problem is trivial. Furthermore, consider only state sequences that start in the first state $i = 1$ and end in the last state $i = M$, since higher likelihood cannot be achieved with partial state sequences. For a linear HMP, $a_{ij} = 0$ if $j \notin \{i, i+1\}$. Let $\alpha_i = a_{i,i+1}$. Let K_i denote the number of time units the chain spends in state i. The probability of spending k_i time units in state $i < M$ and then moving to state $i+1$ is $(1-\alpha_i)^{k_i}\alpha_i$. Hence, the pmf of a state sequence s^n is given by

$$p(s^n; \phi) = \prod_{i=1}^{M-1} (1-\alpha_i)^{k_i}\alpha_i. \tag{7.9}$$

Let $l_i = \sum_{j=1}^{i} k_j$ denote the total number of time units spent in the first i states. Thus, $k_i = l_i - l_{i-1}$, where $l_0 = 0$. The sequence of observations from state i is given by $Y_{l_{i-1}+1}, \dots, Y_{l_i}$, and by assumption, these random variables are statistically independent. In addition, observations from different states are also statistically independent. Hence

$$p(y^n|s^n; \phi) = \prod_{i=1}^{M} \prod_{\tau=l_{i-1}+1}^{l_i} p(y_\tau|S_\tau = i; \phi). \tag{7.10}$$

From (7.9) and (7.10)

$$\begin{aligned}\log p(s^n, y^n; \phi) = &\sum_{i=1}^{M-1} (l_i - l_{i-1}) \log(1-\alpha_i) + \log(\alpha_i) \\ &+ \sum_{i=1}^{M} \sum_{\tau=l_{i-1}+1}^{l_i} \log p(y_\tau|S_\tau = i, \theta_i). \end{aligned} \tag{7.11}$$

The parameter is estimated from maximization of (7.11), first over ϕ, and then over $\{l_i\}$. The maximization over ϕ can be independently performed for each i. Maximizing over α_i gives

$$\alpha_i = \frac{1}{l_i - l_{i-1} + 1}, \qquad i = 1, \dots, M-1 \tag{7.12}$$

where $\alpha_M = 0$. Estimation of θ_i depends on the specific form of the observation conditional density. For an HMP with finite-alphabet of L letters, maximization of (7.11) over the state-to-observation transition matrix gives

$$P(Y_t = y|S_t = i) = f(y;\, l_{i-1}+1,\, l_i), \qquad \begin{array}{l} y = 1, \dots, L \\ t = 1, \dots, n \end{array} \tag{7.13}$$

where $f(y;\, l_{i-1}+1,\, l_i)$ denotes the relative frequency of occurrences of the symbol y in the sequence $y_{l_{i-1}+1}, \dots, y_{l_i}$. Substituting (7.12) and (7.13) in (7.11) gives

$$\begin{aligned}\log p(s^n, y^n; \phi) \\ = &\sum_{i=1}^{M-1} \log\left(\left(1 - \frac{1}{l_i - l_{i-1}+1}\right)^{l_i - l_{i-1}} \frac{1}{l_i - l_{i-1}+1}\right) \\ &+ \sum_{i=1}^{M} \sum_{\tau=l_{i-1}+1}^{l_i} \log f(y_\tau;\, l_{i-1}+1,\, l_i). \end{aligned} \tag{7.14}$$

Maximization of (7.14) over $\{l_i\}$ provides the optimal values that can be used in (7.12) and (7.13) to obtain the parameter ϕ that globally maximizes $p(s^n, y^n; \phi)$. A detailed algorithm is provided in [111] that shows how maximization of (7.14) can be performed using the Viterbi algorithm.

The algorithm extends to parameter estimation from multiple statistically independent training sequences that share a common state sequence. Parameter estimation from multiple training sequences with no restrictions on their individual state sequences does not appear feasible with this noniterative approach. Estimation from multiple training sequences is essential for left–right HMPs and is commonly performed when the Baum–Viterbi algorithm is used in applications such as automatic speech recognition.

E. Bayesian Parameter Estimation

Bayesian estimation of the parameter of an HMP was studied by Robert, Celeux, and Diebold [269]. The approach generalizes Bayesian estimation of mixture processes [87], [270]. It is based on Gibbs sampling of the parameter which is assumed random with a given prior. Usually conjugate priors are used. In [269], the M rows of the transition matrix were assumed statistically independent and a product of M Dirichlet priors was assumed. The observation conditional densities $\{b(y_t|\theta_{s_t})\}$ were assumed members of the exponential family for which a conjugate prior for each θ_j exists. The parameter ϕ can, in principle, be estimated by sampling from the conditional density of the parameter $p(\phi|y^n)$. This, however, appears impractical as this conditional density involves the sum of an exponentially growing number of terms with n. On the other hand, sampling from $p(\phi|s^n, y^n)$ is much simpler since this density constitutes only one term of that sum. Thus, the Gibbs sampling approach proposed in [269] is based on alternative samplings from $p(\phi|s^n, y^n)$ and from $p(s^n|y^n, \phi)$. The first sampling produces an estimate of the parameter ϕ which is then used in the second sampling to estimate the state sequence s^n. Further simplification was obtained by performing n samplings for each ϕ from

$$\begin{aligned} p(s_t|y^n, \{s_\tau, \tau \neq t\}, \phi) &= p(s_t|y_t, s_{t-1}, s_{t+1}, \phi) \\ &= \frac{a_{s_{t-1}s_t}(\phi) b(y_t|\theta_{s_t}(\phi)) a_{s_t s_{t+1}}(\phi)}{\sum_{s_t} a_{s_{t-1}s_t}(\phi) b(y_t|\theta_{s_t}(\phi)) a_{s_t s_{t+1}}(\phi)} \end{aligned} \tag{7.15}$$

instead of a single sampling from $p(s^n|y^n, \phi)$ which requires forward–backward recursions. The Gibbs sampling algorithm produces a sequence $\{\phi_m, s^n(m)\}$ where ϕ_m and $s^n(m)$ denote, respectively, the parameter and state sequence estimates at the end of the mth iteration.

Convergence properties of the Gibbs sampler were studied in [269] and a summary of the results was also given by Rydén and Titterington [280]. It was shown that the sequence $\{\phi_m, s^n(m)\}$ is geometrically ergodic φ-mixing homogeneous Markov chain with a unique stationary distribution given by $p(\phi, s^n|y^n)$. The sequence $\{s^n(m)\}$ is geometrically ergodic φ-mixing Markov chain with a unique stationary distribution given by $p(s^n|y^n)$. The sequence $\{\phi_m\}$, which is not a Markov chain, is ergodic, φ-mixing, and converges weakly as $m \to \infty$ at geometric rate to a stationary distribution given by $p(\phi|y^n)$. It follows from [269, Theorem 1] that the conditional expectation of any function $g(\cdot)$ of the parameter ϕ, given y^n, can be approximated by the corresponding sample average from a realization of $\{\phi_m\}$. A central limit theorem for such an average is given in [269, Corollary 2].

A simulated annealing approach for estimating the parameter of an HMP was developed byAndrieu and Doucet [13]. Each iteration of the algorithm includes the above described iteration and an additional step which aims at accepting or rejecting the new parameter estimate. The decision is based on a probabilistic scheme involving a deterministic cooling schedule. Convergence in probability of $p(\phi_m|y^n)$ to $p(\hat{\phi}|y^n)$ where $\hat{\phi}$ is a MAP estimate of ϕ was shown under some mild regularity conditions.

A Bayesian approach for iterative estimation of the parameter of a switching autoregressive moving average (ARMA) process was developed by Billio, Monfort, and Robert [39]. Several versions of the Gibbs sampler presented earlier, that are particularly suitable for hidden Markov fields, were studied by Qian and Titterington [260] and by Rydén and Titterington [280]. Here sampling is performed from a tractable pseudo-likelihood function of the underlying Markov process. Reparametrization of HMPs with Gaussian and Poisson observation conditional densities, using less informative priors, was studied by Robert and Titterington [271].

VIII. Order Estimation

The *order* is the number of states of the HMP. Algorithms for estimating the parameter of an HMP assume that the order is known. In many applications this is not the case. For example, in blind deconvolution of unknown communication channels, the received signal is an HMP, but its order determined by the memory length of the channel is not known. This application is further discussed in Section XIV-C. In addition, HMPs are not identifiable if their order is overestimated [116], [156, Ch. 22], [282]. Information-theoretic approaches for order estimation of a finite-alphabet HMP were developed by Finesso [116], Ziv and Merhav [330], Kieffer[187], and Liu and Narayan [223]. An order estimation approach for a general HMP was developed by Rydén [277]. These approaches are reviewed in this section.

Let M be the true order and let ϕ^0 be the true parameter of an HMP $\{Y_t\}$. Let $\hat{M}_n$ denote an estimate of M from an observation sequence y^n. Let ϕ_j denote the parameter of an HMP with assumed order j. Let $\Phi^{(j)}$ denote the parameter set. For a finite-alphabet HMP of assumed order j and parameter in Φ_δ, we denote the parameter set by $\Phi_\delta^{(j)}$. Also, L denotes the size of the alphabet. Let $\mathcal{P}_{\phi_j}$, $j = 1, 2, \ldots$, denote the sequence of nested HMP densities. All but the order estimator of [223] use the ML estimate of ϕ_j. Let

$$\hat{\phi}_j = \arg \max_{\phi \in \Phi^{(j)}} \log p(y^n; \phi). \tag{8.1}$$

The order estimator for a finite-alphabet HMP proposed by Finesso is given by [116]

$$\hat{M}_n = \min \left\{ \arg\min_{j \geq 1} \left\{ -\frac{1}{n} \log p\left(y^n; \hat{\phi}_j\right) + 2c_j^2 \frac{\log n}{n} \right\} \right\} \tag{8.2}$$

where $\hat{\phi}_j$ is the ML estimator over $\Phi_\delta^{(j)}$ and $c_j = j(j+L-2)$. This penalized ML estimator was proved strongly consistent when $\phi^0 \in \Phi_\delta$ and $-D_\phi^2 \bar{H}(P_{\phi^0}, P_\phi)_{|_{\phi=\phi^0}} > 0$, where $\bar{H}(P_{\phi^0}, P_\phi)$ is defined in (4.40) [116, Theorem 4.5.2]. The order estimator uses an estimate of the rate of growth of the maximized log-likelihood ratio $\log p(y^n; \hat{\phi}_j)/p(y^n; \phi^0)$ which was found to be in the order of $\log n$ a.s. [116, Theorem 4.4.1].

The order estimator for a finite-alphabet HMP proposed by Ziv and Merhav was derived using a Neyman–Pearson type criterion [330]. It minimizes the underestimation probability $\Pr(\hat{M}_n < M)$, uniformly for all HMPs in $\mathcal{P}_{\phi_M}$, subject to an exponential decay of the overestimation probability given by

$$\liminf_{n\to\infty} -\frac{1}{n} \log P\left(\hat{M}_n > M\right) > \lambda \tag{8.3}$$

for all HMPs in $\mathcal{P}_{\phi_M}$. The estimator is given by

$$\hat{M}_n = \min \left\{ j\colon -\frac{1}{n} \log p\left(y^n; \hat{\phi}_j\right) - \frac{1}{n} u(y^n) < \lambda \right\} \tag{8.4}$$

where $u(y^n)$ is the length of the binary codeword for y^n in the Lempel–Ziv universal data compression scheme. This length function was defined in Section VI-E. If $-\log p(y^n; \hat{\phi}_j)$ is interpreted as a model-based codeword length for y^n [68, p. 85], then (8.4) seeks the shortest model-based binary codeword length that is sufficiently close to the universal codeword length $u(y^n)$. Alternatively, using Ziv's inequality (6.28), the order estimator (8.4) is a likelihood ratio test in which $-\log p(y^n; \phi^0)$ is replaced by $u(y^n)$. Unlike some other estimators presented in this section, (8.4) does not require knowledge of an upper bound on the order M.

It was pointed out in [223], [187] that the estimator (8.4) tends to underestimate the order of the HMP and hence is not consistent. Liu and Narayan [223] proposed a slightly modified estimator and proved its consistency for a stationary ergodic HMP that satisfies some mild regularity conditions. The estimator assumes knowledge of an upper bound $\overline{M}$ on M. It uses the binary codeword length $v(y^n)$ for encoding y^n in the Wyner–Ziv asymptotically optimal universal compression scheme [319]. The estimator is given by

$$\hat{M}_n = 1 + \max\left\{1 \le j \le \overline{M}: -\frac{1}{n}\log p(y^n;\, \hat{\phi}_j) - \frac{1}{n}v(y^n) > \lambda_n\right\} \tag{8.5}$$

provided the set is not empty, otherwise, $\hat{M}_n = 1$. The sequence λ_n must satisfy $\lim_{n\to\infty}\lambda_n = 0$ and $\lim_{n\to\infty} n\lambda_n = \infty$. The estimator is strongly consistent if $\sum_{n=1}^{\infty} 2^{-n\lambda_n} < \infty$ and is weakly consistent otherwise.

Liu and Narayan [223] proposed another strongly consistent order estimator for a stationary ergodic finite-alphabet HMP. They assumed that the parameter ϕ_j is random with prior η, and the pmf of y^n is the mixture

$$q_j(y^n) = \int p(y^n|\phi_j)\eta(d\phi_j). \tag{8.6}$$

Dirichlet priors were assumed for the entries of A and B. Estimation of $q_j(y^n)$ in terms of relative frequencies of states and observation symbols is outlined in [223, Appendix]. Avoiding ML estimation is desirable since only local ML estimation procedures are available. The mixture model (8.6), however, is not trivial to estimate. Aspects of data modeling in the minimum description length (MDL) sense using mixture densities and ML estimated parameters were studied by Barron, Rissanen, and Yu [24]. They provided sufficient conditions for asymptotic equivalence of the two approaches. The order estimator of Liu and Narayan [223] is given by

$$\hat{M}_n = \max\left\{1 \le j \le \overline{M}: q_j(y^n) - q_{j-1}(y^n) > c_j \log n\right\} \tag{8.7}$$

where $c_j = (j(j+L-2)+5)/2$ and $q_0(y^n) = 1$. If the set in (8.7) is empty then $\hat{M}_n = 1$. This estimator provides exponentially decaying underestimation probability and polynomially (as $1/n^3$) decaying overestimation probability.

Kieffer [187, Theorem 2] proposed a code-based order estimator for a class of stationary ergodic *constrained finite-state sources* and proved strong consistency of the estimator. Stationary ergodic finite-alphabet HMPs are special cases of that class. Let φ_j denote a code designed for $\mathcal{P}_{\phi_j}$ source sequences. The code φ_j is a mapping of source sequences into binary strings such that $\varphi_j(y^n)$ is not a prefix of $\varphi_j(z^n)$ if y^n and z^n are two distinct sequences. Let $w(\varphi_j(y^n))$ denote the length of the binary string $\varphi_j(y^n)$. Kieffer used ML codes $\varphi_j(y^n)$ whose lengths are determined by $-\log p(y^n;\, \hat{\phi}_j)$. The estimator is given by

$$\hat{M}_n = \arg\min_j\{w(\varphi_j(y^n)) + k_j \log n\} \tag{8.8}$$

where k_j is a subsequence of the positive integers Z^+ that satisfies $k_{j+1} \ge 2(k_j+1)$ and $\sum_{y^n}\sup_{\phi\in\Phi_j} p(y^n;\, \phi) \le n^{k_j}$ for $n \ge 2$ and all $j \in Z^+$. For sufficiently large n, this estimator takes the approximate form

$$\hat{M}_n \approx \arg\min_j\left\{-\log p\left(y^n;\, \hat{\phi}_j\right) + c_j \log n\right\} \tag{8.9}$$

where $\{c_j\}$ is a nondecreasing sequence of positive constants that is determined from the model classes $\{\mathcal{P}_{\phi_j}\}$. This estimator resembles the MDL code-based order estimator of Rissanen [267] or the Bayesian information criterion (BIC) based order estimator derived independently by Schwarz [287]. The MDL order estimator uses a code for the class $\mathcal{P}_{\phi_j}$ whose expected redundancy grows at the minimum possible rate for almost all sequences modeled by members of $\mathcal{P}_{\phi_j}$. This results in positive constants $d_j \le c_j$ for all j. It is noted that relatively small penalty terms may not provide consistent order estimators as overfitting of the data may prevail. Sufficient conditions for consistency of the MDL order estimator and examples of model classes for which the MDL estimator is consistent were given by Kieffer [187], Barron, Rissanen, and Yu [24] and Csiszár and Shields [74].

Rydén [277] proposed an order estimator for a stationary ergodic general HMP. The estimator is based on the MSDLE obtained from maximization of $q(y^{nm};\, \phi)$ in (6.7). When the HMP is identifiable, in particular, when all $\{\theta_j\}$ are distinct, any $m \ge 2$ may be used and there is no need for an estimate of the largest possible order of the HMP. Otherwise, an upper bound $\overline{M}$ on M is required, and $m \ge 2\overline{M}$ must be used, since finite-dimensional distributions of the HMP are uniquely determined by the $2\overline{M}$-dimensional distribution [277, Theorem 1]. The order estimator is given by

$$\hat{M}_n = \arg\max_j\left\{\log q\left(y^{nm};\, \tilde{\phi}_j\right) - c_{j,n}\right\} \tag{8.10}$$

where $\tilde{\phi}_j$ is the maximizer of $q(y^{nm};\, \phi)$ over $\phi \in \Phi_j$, and $c_{j,n}$ is a nondecreasing sequence of real numbers that penalize the likelihood and thus prevent overestimation of the model order. When $\overline{M}$ is required, maximization in (8.10) is over $1 \le j \le [m/2]$ where $[\cdot]$ denotes an integer part. The sequence $c_{j,n}$ satisfies $c_{j+1,n} \ge c_{j,n}$ for all n and $\limsup_n c_{j,n}/n = 0$. Under these and some additional regularity conditions, it was shown in [277, Theorem 2] that the estimator (8.10) does not underestimate the order of the HMP asymptotically as $n \to \infty$, with probability one. The regularity conditions hold, for example, for HMPs with observation conditional densities from the Poisson, negative exponential, and normal with fixed variance families. The conditions on $c_{j,n}$ are satisfied by the penalizing terms used in the Akaike information criterion (AIC) [4] and in MDL [267] or BIC [287]. Thus, these estimators never underestimate the HMP order when n is sufficiently large. The AIC choice is $c_{j,n} = \dim(\Phi_j)$ and the BIC choice is $c_{j,n} = (1/2)\dim(\Phi_j)\log(n)$ where $\dim(\Phi_j)$ denotes the dimension of the parameter space of the jth-order HMP. An earlier similar result on order estimation of mixture processes obtained from maximization of a penalized likelihood function was proved by Leroux [213, Theorem 4]. Additional references on consistent order estimators for mixture processes can be found in Rydén [277].

IX. Dynamical System Approach

We have seen in Section IV-B6 that a finite-alphabet HMP has a dynamical system representation in the sense of control theory. Similar representations exist for other types of HMPs with discrete as well as continuous time and discrete as well as continuous range state and observation processes. Elliott, Aggoun, and Moore [99] provide a comprehensive study of HMPs in the dynamical system setup. They develop conditional mean estimators for the states, the number of jumps from one state to another, the state occupation time, and for some statistics reflecting the assignment of observations among the various states. The estimators are then used in the EM algorithm for ML estimation of the parameter of the HMP. As is well known, conditional mean estimation of continuous-time signals usually results in infinite-dimensional filters for nonlinear non-Gaussian problems. The book [99] contains almost all known estimation problems for which finite-dimensional conditional mean estimators exist. In this section, we demonstrate the approach for a discrete-time HMP with Gaussian observation conditional densities. The approach requires forward recursions only. Its main advantage is that it generalizes to continuous-time HMPs.

Application of the EM algorithm for estimating the parameter of a discrete-time dynamical system using Kalman smoothers was first performed by Shumway and Stoffer [292]. The approach was then expanded by several authors. Zeitouni and Dembo [324] studied finite-state continuous-time Markov chains observed in white noise. They developed a finite-dimensional conditional mean causal estimator for the number of jumps from one state to another. The estimator was used in an extended EM algorithm for ML estimation of the transition matrix of the Markov chain. The extension of the EM algorithm to continuous-time processes and its convergence properties were established by Dembo and Zeitouni [77]. They also applied the EM algorithm to a wide class of diffusion processes which resulted in iterative applications of finite-dimensional Kalman smoothers. A finite-dimensional conditional mean causal estimator for the states of the chain was first developed by Wonham [315]. Finite-dimensional conditional mean estimators for the state occupation time and for a stochastic integral related to the drift in the observation process were derived by Elliott [98]. MAP estimators of a randomly, slowly varying parameter, of a continuous-time and a discrete-time ARMA processes, were developed by Dembo and Zeitouni in [78] and [79], respectively. ML estimation of the parameter of a discrete-time dynamical system using Kalman filters rather than smoothers in conjunction with the EM approach was developed by Elliott and Krishnamurthy [100]. Robust time discretization of the continuous-time filters and smoothers for estimating the parameter of an HMP was studied by James, Krishnamurthy, and Le Gland [168].

The central theme in [99] is to derive conditional mean estimators for statistics of the HMP which are required for ML estimation of its parameter by the EM algorithm. The conditional mean estimators are developed using a generalized Bayes rule. This is a standard technique used, for example, in [324]. This rule, or formula, enables evaluation of a conditional mean under one probability measure using another more convenient probability measure. This is done as follows. Let $P_1 \ll P_0$ be two probability measures on the measurable space $(\Omega, \mathcal{F})$. Let $\Lambda(\omega) = dP_1(\omega)/dP_0(\omega)$, $\omega \in \Omega$, denote the Radon–Nikodym derivative or density of P_1 with respect to P_0. Let $\mathcal{G} \subseteq \mathcal{F}$ denote a sub-σ-field of $\mathcal{F}$. Let X denote a random variable on $\{\Omega, \mathcal{F}\}$. Let $E_1\{X|\mathcal{G}\}$ denote the desired conditional mean of X under P_1. Let $E_0\{X|\mathcal{G}\}$ denote the conditional mean of X under P_0. The *generalized Bayes rule* [247, Lemma 8.6.2], or the Kallianpur–Striebel formula [222, Lemma 7.4], is given by

$$E_1\{X|\mathcal{G}\} = \frac{E_0\{\Lambda X|\mathcal{G}\}}{E_0\{\Lambda|\mathcal{G}\}} \tag{9.1}$$

for all ω such that $E_0\{\Lambda|\mathcal{G}\} \neq 0$, otherwise, $E_1\{X|\mathcal{G}\}$ can be arbitrarily chosen. The approach can be applied when P_1 is the probability measure of the HMP and P_0 is the probability measure of an i.i.d. process that is independent of the Markov chain. The approach is demonstrated here for a discrete-time HMP with Gaussian observation conditional densities. Our discussion follows [99, Ch. 3].

Let $(\mathcal{A}^\infty, \mathcal{B}_{\mathcal{A}}^\infty)$ denote a sequence measurable space where $\mathcal{A}^\infty = \{s_0^\infty, y_1^\infty\}$ is the set of all state and observation sequences, and $\mathcal{B}_{\mathcal{A}}^\infty$ denotes the Borel product σ-field. Let P_1 be the distribution of the HMP on $(\mathcal{A}^\infty, \mathcal{B}_{\mathcal{A}}^\infty)$. For the Markov chain we use the same representation as in (4.19). Specifically, we assume a Markov chain $\{S_t\}$ with state space $\mathcal{S} = \{e_j, j = 1, \ldots, M\}$ where e_j is a unit vector in $\mathcal{R}^M$, a transition matrix A, and a stationary martingale difference sequence $\{V_t\}$. The observation process $\{Y_t\}$ is characterized by a sequence $\{W_t\}$ of i.i.d. standard Gaussian random variables independent of $\{S_t\}$, and two M-dimensional vectors c and σ representing the means and standard deviations of the Gaussian observation conditional densities in the M states. All components of σ are assumed positive. The dynamical system representation of the HMP under P_1 is given by

$$\begin{aligned} S_{t+1} &= A'S_t + V_{t+1} \\ Y_{t+1} &= c'S_t + (\sigma'S_t)W_{t+1}, \qquad t = 0, 1, 2, \ldots. \end{aligned} \tag{9.2}$$

Let P_0 denote a second distribution on $(\mathcal{A}^\infty, \mathcal{B}_{\mathcal{A}}^\infty)$. Under P_0, $\{S_t\}$ has the same distribution as under P_1, $\{W_t\}$ is an i.i.d. sequence of standard Gaussian random variables, and $\{S_t\}$ and $\{W_t\}$ are statistically independent. The dynamical system representation of the HMP under P_0 is given by

$$\begin{aligned} S_{t+1} &= A'S_t + V_{t+1} \\ Y_{t+1} &= W_{t+1}, \qquad t = 0, 1, 2, \ldots. \end{aligned} \tag{9.3}$$

Let $P_1^{(n)}$ and $P_0^{(n)}$ denote the n-dimensional distributions induced by P_1 and P_0, respectively. Clearly, $P_1^{(n)}$ and $P_0^{(n)}$ possess densities with respect to $\kappa^n \times \mu^n$, where μ here is the Lebesgue measure, and $P_1^{(n)} \ll P_0^{(n)}$. Let $p_1(\cdot)$ and $p_0(\cdot)$ denote the n-dimensional densities corresponding to $P_1^{(n)}$ and $P_0^{(n)}$, respectively. Assume that $\kappa^n \times \mu^n \ll P_0^{(n)}$. The Radon–Nikodym derivative of $P_1^{(n)}$ with respect to $P_0^{(n)}$ is given by

$$\begin{aligned} \Lambda(s_0^{n-1}, y^n) &= \frac{p_1(y^n|s_0^{n-1})p(s_0^{n-1})}{p_0(y^n)p(s_0^{n-1})} \\ &= \prod_{t=0}^{n-1} \frac{p_1(y_{t+1}|s_t)p(s_t|s_{t-1})}{p_0(y_{t+1})p(s_t|s_{t-1})} \\ &= \prod_{t=0}^{n-1} \frac{g\left(\frac{y_{t+1}-c's_t}{\sigma's_t}\right)\frac{1}{\sigma's_t}}{g(y_{t+1})} \end{aligned} \tag{9.4}$$

where $p(s_0|s_{-1}) = p(s_0)$ and $g(y) = (2\pi)^{-1/2}\exp\{-y^2/2\}$ denotes the standard normal pdf.

To state the generalized Bayes rule for the systems (9.2) and (9.3) let $\mathcal{G}_n = \sigma(S_0^{n-1}, Y^n)$ denote the smallest σ-field generated by (S_0^{n-1}, Y^n). The sequence $\{\mathcal{G}_n\}$ forms a filtration.

Similarly, let $\mathcal{Y}_n = \sigma(Y^n)$ denote the smallest σ-field generated by Y^n. Let $\{X_n\}$ be a sequence of scalar integrable random variables adapted to $\{\mathcal{G}_n\}$. From (9.1)

$$E_1\{X_n|\mathcal{Y}_n\} = \frac{E_0\left\{\Lambda\left(S_0^{n-1}, Y^n\right)X_n|\mathcal{Y}_n\right\}}{E_0\left\{\Lambda\left(S_0^{n-1}, Y^n\right)|\mathcal{Y}_n\right\}}. \tag{9.5}$$

This equation can be verified using the first line of (9.4) without resorting to measure theoretic arguments. We emphasize that $\{S_t\}$ and $\{Y_t\}$ are statistically independent under P_0. In a more general situation of a finite-energy continuous-time continuous-range signal $\{S_t\}$ observed in white noise, the Radon–Nikodym derivative of P_1 with respect to P_0 is given by Girsanov theorem [222, Theorem 6.3], [247, Theorem 8.6.3]. This form involves a stochastic integral.

Let

$$\gamma_n(X_n) = E_0\left\{\Lambda\left(S_0^{n-1}, Y^n\right)X_n|\mathcal{Y}_n\right\} \tag{9.6}$$

be the nonnormalized version of $E_1\{X_n|\mathcal{Y}_n\}$ and rewrite (9.5) as

$$E_1\{X_n|\mathcal{Y}_n\} = \frac{\gamma_n(X_n)}{\gamma_n(1)}. \tag{9.7}$$

It is easier to derive recursions for $\gamma_n(X_n)$ than for $E_1\{X_n|\mathcal{Y}_n\}$. Hence, (9.7) is the basic equation we shall be working with.

Of interest are special cases of X_n that provide sufficient statistics for an EM iteration in ML estimation of the HMP parameter. These are as follows.

i) $X_n = J_{ij}(n)$. This is the number of jumps from state i to state j during n transitions of the chain. It is given by

$$J_{ij}(n) = \sum_{t=1}^{n}(S'_{t-1}e_i)(S'_t e_j), \qquad j = 1, \ldots, M. \tag{9.8}$$

ii) $X_n = O_j(n)$. This is the occupation time of state j in n chain transitions given by

$$O_j(n) = \sum_{t=1}^{n}(S'_{t-1}e_j), \qquad j = 1, \ldots, M. \tag{9.9}$$

iii) $X_n = T_j^f(n)$, for some deterministic function $f(\cdot)$. This random variable represents the sum of elements of $\{f(Y_t), t = 1, \ldots, n\}$ assigned to state j during n transitions of the chain. Of interest here are the functions $f(y) = y$ and $f(y) = y^2$. $T_j^f(n)$ is defined by

$$T_j^f(n) = \sum_{t=1}^{n}(S'_{t-1}e_j)f(Y_t), \qquad j = 1, \ldots, M. \tag{9.10}$$

It turns out that a recursion for the M-dimensional vector $\gamma_n(X_nS_n)$ can be developed from which the desired $\gamma_n(X_n)$ can be obtained simply by taking the inner product $\gamma'_n(X_nS_n)\mathbf{1}$ where $\mathbf{1}$ denotes an $M \times 1$ vector of 1's. We shall therefore focus on the development of the recursions for calculating $\gamma_{n+1}(X_{n+1}S_{n+1})$ from $\gamma_n(X_nS_n)$ for $n \geq 1$. This will also provide a recursion for estimating the state vector S_n at time n simply by assigning $X_n = 1$ in $\gamma_n(X_nS_n)$. A general recursion for $\gamma_{n+1}(X_{n+1}S_{n+1})$ when X_{n+1} is any of the above defined four random variables was given in [99, Theorem 3.5.3]. The recursion is given in terms of $d_k = A'e_k$ and

$$\lambda_k(y_{n+1}) = \frac{g\left(\frac{y_{n+1}-c'e_k}{\sigma'e_k}\right)\frac{1}{\sigma'e_k}}{g(y_{n+1})}\,e_k. \tag{9.11}$$

Note that $\lambda_k(y_{n+1})$ depends on the observation y_{n+1} as well as the parameter of the HMP. It constitutes the product of the kth-unit vector e_k and the last multiplicative term of $\Lambda(S_0^n, y^{n+1})$ for $S_n = e_k$. The identity $\sum_{k=1}^{M} S'_n e_k = 1$ was found useful in deriving the recursions. For example, using this identity and the state equation from (9.3), it is easy to verify the following recursion for estimating the state vector:

$$\gamma_{n+1}(S_{n+1}) = \sum_{k=1}^{M}\left[\gamma'_n(S_n)\lambda_k(y_{n+1})\right]d_k. \tag{9.12}$$

The recursions for estimating the other statistics represented by X_n are given by

$$\begin{aligned}\gamma_{n+1}(J_{ij}(n+1)S_{n+1}) = &\sum_{k=1}^{M}[\gamma'_n(J_{ij}(n)S_n)\lambda_k(y_{n+1})]\,d_k\\ &+ [\gamma'_n(S_n)\lambda_i(y_{n+1})]a_{ij}e_j\end{aligned} \tag{9.13}$$

$$\begin{aligned}\gamma_{n+1}(O_j(n+1)S_{n+1}) = &\sum_{k=1}^{M}[\gamma'_n(O_j(n)S_n)\lambda_k(y_{n+1})]\,d_k\\ &+ [\gamma'_n(S_n)\lambda_j(y_{n+1})]\,d_j\end{aligned} \tag{9.14}$$

$$\begin{aligned}\gamma_{n+1}(T_j^f(n+1)S_{n+1}) = &\sum_{k=1}^{M}[\gamma'_n(T_j^f(n)S_n)\lambda_k(y_{n+1})]\,d_k\\ &+ [\gamma'_n(S_n)\lambda_j(y_{n+1})]f(y_{n+1})\,d_j.\end{aligned} \tag{9.15}$$

These recursions can now be used to obtain the conditional mean estimates $E_1\{X_{n+1}|\mathcal{Y}_{n+1}\}$. For $X_n = S_n$ we use (9.12) to recursively calculate

$$\begin{aligned}\hat{S}_{n+1} &= E_1\{S_{n+1}|\mathcal{Y}_{n+1}\}\\ &= \frac{\gamma_{n+1}(S_{n+1})}{\gamma'_{n+1}(S_{n+1})\mathbf{1}}.\end{aligned} \tag{9.16}$$

Note that $\gamma_{n+1}(S_{n+1})$ is the vector of nonnormalized conditional probabilities of S_{n+1} given $\mathcal{Y}_{n+1}$ since $\hat{S}_{n+1}$ is the M-dimensional vector whose jth component is $P(S_{n+1} = e_j|\mathcal{Y}_{n+1})$. Equations (9.12) and (9.16) coincide with (5.14). A smoothed estimator for S_{n+1} was derived in [99, eq. 3.6.2]. For $X_n = J_{ij}(n)$ we use (9.13) to recursively calculate

$$\begin{aligned}\hat{J}_{ij}(n+1) &= E_1\{J_{ij}(n+1)|\mathcal{Y}_{n+1}\}\\ &= \frac{\gamma'_{n+1}(J_{ij}(n+1)S_{n+1})\mathbf{1}}{\gamma'_{n+1}(S_{n+1})\mathbf{1}}\\ &= \frac{\gamma_{n+1}(J_{ij}(n+1))}{\gamma_{n+1}(1)}.\end{aligned} \tag{9.17}$$

Estimation of the other two random variables $O_j(n+1)$ and $T_j^f(n+1)$ can be performed similarly from (9.14) and (9.15), respectively.

Estimation of the parameter ϕ^0 of the HMP (9.2) from n observations can be iteratively performed using the EM algorithm. In the context of this section, the parameter in each iteration can be estimated from maximization of the following function over $\hat{\phi}$ [77]

$$\Delta Q_n\left(\hat{\phi}, \phi_m\right) = E_{\phi_m}\left\{\log \frac{dP^{(n)}_{\hat{\phi}}}{dP^{(n)}_{\phi_m}}\left(S_0^{n-1}, y^n\right) \big| \mathcal{Y}_n\right\}. \tag{9.18}$$

This function is analogous to the right-hand side of (6.9). The usual parametrization is assumed. Maximization of (9.18) subject to natural constraints gives the following estimates at the $m+1$th iteration. For $\{i, j\} \in \{1, \ldots, M\}$

$$\hat{a}_{ij}(n) = \frac{\hat{J}_{ij}(n)}{\hat{O}_i(n)} = \frac{\gamma_n(J_{ij}(n))}{\gamma_n(O_i(n))} \tag{9.19}$$

$$\hat{c}_j(n) = \frac{\hat{T}_j^y(n)}{\hat{O}_j(n)} = \frac{\gamma_n(T_j^y(n))}{\gamma_n(O_j(n))} \tag{9.20}$$

$$\begin{aligned}\hat{\sigma}_j(n) &= \frac{1}{\hat{O}_j(n)}\left[\hat{T}_j^{y^2}(n) - 2\hat{c}_j(n)\hat{T}_j^y(n) + \hat{c}_j^2(n)\hat{O}_j(n)\right] \\ &= \frac{1}{\gamma_n(O_j(n))}\Big[\gamma_n\left(T_j^{y^2}(n)\right) - 2\hat{c}_j(n)\gamma_n\left(T_j^y(n)\right) \\ &\qquad + \hat{c}_j^2(n)\gamma_n(O_j(n))\Big].\end{aligned} \tag{9.21}$$

The recursions for estimating $J_{ij}(n)$, $O_i(n)$, $T_j^y(n)$, and $T_j^{y^2}$ are calculated based on the available parameter ϕ_m and a fixed number of observations n. These re-estimation formulas may be interpreted similarly to the re-estimation formulas (6.15), (6.17), and (6.18), respectively. Note that only forward recursions are used in (9.19)–(9.21). Furthermore, the parameter estimates can be straightforwardly updated when the number of observations is increased from n to $n+1$.

X. Recursive Parameter Estimation

Recursive estimation of the parameter of an HMP is of great practical and theoretical importance since one always wishes to be able to update the parameter estimate when new observations become available. Consider, for example, hidden Markov modeling of speech signals in automatic speech recognition applications. Here, an affirmative human feedback can be used by the recognizer to improve the modeling of a particular word using the speech utterance entered by the user. This, of course, could not be done with the Baum algorithm which requires the entire observation sequence in each iteration. Recursive estimation is also desired when adapting to time-varying parameter of an HMP. This situation occurs in automatic speech recognition, neurophysiology, and data communications when the underlying HMP changes with time. These applications are discussed in Section XIV. Recursive estimation may also be computationally more efficient and require less storage than the Baum algorithm.

Recursive estimation of the parameter of an HMP was studied as early as 1970 by Kashyap [182]. A stochastic descent recursion was developed for estimating the transition matrix of a Markov chain observed through arbitrary noise with independent samples and some unknown finite variance. Convergence of the recursion with probability one and in mean square was shown under some conditions.

With the introduction of the EM algorithm in 1977 there has been renewed interest in recursive estimation from incomplete data. Although HMPs fall into this category, recursions for general incomplete data models are not immediately applicable to HMPs. Recursions for parameter estimation from incomplete data often aim at least at local minimization of the relative entropy

$$K(\phi) = E_{\phi^0}\left\{\log \frac{p(Y_{n+1}; \phi^0)}{p(Y_{n+1}; \phi)}\right\} \tag{10.1}$$

over $\phi \in \Phi$ where ϕ^0 is the true parameter. The relative entropy attains its global minimum of zero for $\phi \sim \phi^0$. To describe a recursion with this goal, let $h(y_{n+1}; \phi) = D_\phi \log p(y_{n+1}; \phi)$ denote the score function and let F_n denote a matrix of suitable dimension. The recursion has the form of

$$\hat{\phi}_{n+1} = \hat{\phi}_n + \frac{1}{n+1}F_n h\left(y_{n+1}; \hat{\phi}_n\right) \tag{10.2}$$

where the specific form of the adaptive matrix F_n significantly affects convergence properties of the recursion. Of particular interest is the inverse of the information matrix for the incomplete data given by $I_n(\phi) = E_\phi\{h(Y_n; \phi)h'(Y_n; \phi)\}$. For $F_n = I_n^{-1}(\hat{\phi}_n)$, and under suitable regularity conditions, the recursion can be shown to be consistent asymptotically normal and efficient in the sense of achieving equality in the Cramér–Rao inequality [110], [281]. Rydén [281] showed that some of these conditions, however, do not hold for mixture processes and hence cannot hold for HMPs. The recursion (10.2) with $F_n = I_n^{-1}(\hat{\phi}_n)$ is also difficult to implement since explicit form of the incomplete data information matrix is rarely available. Titterington [300, eq. 9] proposed to use instead the information matrix for the complete data. The recursion was related to an EM iteration and proved under some conditions to be consistent and asymptotically normal for i.i.d. data. This recursion, however, is never efficient and its convergence for mixture processes was not proved [281]. Weinstein, Feder, and Oppenheim [310, eqs. (19)–(21)] derived a similar EM related recursion for stationary ergodic processes but did not study its properties.

Another recursion with the same goal of minimizing the relative entropy (10.1) proposed in [310, eq. (4)] is given by

$$\hat{\phi}_{n+1} = \hat{\phi}_n + \gamma_n h\left(y_{n+1}; \hat{\phi}_n\right) \tag{10.3}$$

where the sequence $\{\gamma_n\}$ satisfies

$$\lim_{n\to\infty} \gamma_n = 0$$

$$\sum_{n=1}^{\infty} \gamma_n = \infty$$

and

$$\sum_{n=1}^{\infty} \gamma_n^2 < \infty.$$

It was suggested that $h(y_{n+1};\,\hat{\phi}_n)$ may be calculated from the complete data using a one-dimensional version of the identity (6.24) given by

$$h(y_{n+1};\,\phi) = E_\phi\{D_\phi \log p(y_{n+1},\, S_{n+1};\,\phi)\,|\,y_{n+1}\}. \quad (10.4)$$

For HMPs, the alternative (10.4) does not offer computational savings over direct calculation of $h(y_{n+1};\,\hat{\phi}_n)$ using (4.5), particularly when estimating the transition matrix of the Markov chain. Another form for calculating $h(y_{n+1};\,\hat{\phi}_n)$, presented below, is more suitable for HMPs. It was argued in [310] that the recursion (10.4) is consistent in the strong sense and in the mean-square sense for stationary ergodic processes that satisfy some regularity conditions. Some of these conditions, however, are in general violated for i.i.d. observations from a finite mixture density and hence by HMPs [279], [281]. This problem can be circumvented if minimization of $K(\phi)$ is constrained to a compact convex subset $G \subseteq \Phi$ by projecting $\hat{\phi}_{n+1}$ onto G in each iteration [279], [281]. Of course, $\phi^0 \in G$. The estimator (10.3) with $\gamma_n = n^{-\alpha}$, $1/2 < \alpha < 1$, is asymptotically efficient if post-averaging of parameter estimates is applied [281]. A consistent asymptotically efficient estimator in the sense of [211, p. 404] for i.i.d. data with better finite-sample properties was proposed by Rydén [281, Theorem 3]. The estimator has the form of (10.2), where F_n^{-1} is an empirical estimate of the incomplete data information matrix and parameter estimates are recursively projected onto G. These ideas were also found useful for HMPs as will be seen shortly.

Holst and Lindgren [163, eq. 16] first proposed a recursion of the form of (10.2) for estimating the parameter of an HMP. They used

$$h(y_n;\,\phi) = E_\phi\{D_\phi \log p(y_n,\, S_n|S_{n-1};\,\phi)\,|\,y^n\} \quad (10.5)$$

and an empirical estimate of the incomplete data information matrix in the form of the adaptive matrix

$$F_n^{-1} = \frac{1}{n}\sum_{t=1}^{n} h\left(y_t;\,\hat{\phi}_{t-1}\right) h'\left(y_t;\,\hat{\phi}_{t-1}\right). \quad (10.6)$$

The conditional expectation in (10.5) is over $(S_{n-1},\, S_n)$ given y^n, and it can be efficiently calculated using a forward recursion form Section V-A. Note that $h(y_n;\,\phi)$ does not equal $D_\phi \log p(y_n|y^{n-1};\,\phi)$ and hence is not a score function. Evaluation of F_n is done recursively from F_{n-1} and $h(y_n;\,\hat{\phi}_{n-1})$ without matrix inversion [163, eq. 14]. Rydén [279] argued that the recursion of Holst and Lindgren aims at local minimization of the relative entropy rate $\overline{D}(P_{\phi^0}||P_\phi)$ defined in (4.41). Moreover, he showed that if $\hat{\phi}_n \to \phi^0$, then $n^{1/2}(\hat{\phi}_{n+1} - \phi^0)$ is asymptotically normal with zero mean and covariance matrix given by the inverse of $\lim_{n\to\infty} E_{\phi^0}\{h(Y_n;\,\phi^0)h'(Y_n;\,\phi^0)\}$. Lindgren and Holst [220] applied the recursion for estimating the parameter of a Markov modulated Poisson process. Holst, Lindgren, Holst, and Thuvesholmen [164] applied the recursion for estimating the parameter of a switching autoregressive process with Markov regime. Krishnamurthy and Moore [195, eq. 3.18] applied similar ideas to recursive estimation of a Markov chain observed in white Gaussian noise.

Rydén [279] proposed a recursion for estimating the parameter of an HMP which does not use the adaptive matrix F_n. The recursion uses vectors $\boldsymbol{y}_n = (y_{(n-1)m+1},\, \ldots,\, y_{nm})$ of m successive observations, and a projection P_G into a set G. Let $h(\boldsymbol{y};\,\phi) = D_\phi \log p(\boldsymbol{y};\,\phi)$ denote the score function where $p(\boldsymbol{y};\,\phi)$ is the m-dimensional density of the HMP given in (4.3). The recursion is given by

$$\hat{\phi}_{n+1} = P_G\left(\hat{\phi}_n + \gamma_n h\left(\boldsymbol{y}_{n+1};\,\hat{\phi}_n\right)\right) \quad (10.7)$$

where $\gamma_n = \gamma_0 n^{-\alpha}$ for some $\gamma_0 > 0$ and $\alpha \in (1/2,\, 1]$. The set G is assumed a compact convex subset of Φ which contains ϕ^0, it is the closure of its interior, G can be written as

$$G = \{\phi{:}\, g_j(\phi) \le 0,\, j = 1,\, \ldots,\, J\}$$

for some finite set $\{g_j,\, j = 1,\, \ldots,\, J\}$ of continuously differentiable functions, and at each $\phi \in \partial G$, the gradients of the active constraints (i.e., those g-functions with $g(\phi) = 0$) are linearly independent. The simplest G that satisfies these requirements is a simplex whereas all g-functions are linear.

Rydén [279] studied statistical properties of (10.7) assuming a stationary irreducible aperiodic Markov chain and some additional mild regularity conditions. These conditions are satisfied by many important parametric densities including normal densities with positive variances. The sequence $\{\hat{\phi}_n\}$ generated by (10.7) was shown to converge almost surely to the set of Kuhn–Tucker points for minimizing the relative entropy

$$K_m(\phi) = E_{\phi^0}\left\{\log \frac{p(Y^m;\,\phi^0)}{p(Y^m;\,\phi)}\right\} \quad (10.8)$$

over the set G [279, Corollary 1]. The relative entropy attains its global minimum at $\phi{\sim}\phi^0$ provided that the HMP is identifiable. Conditions for identifiability were given in Section VI-A where in particular $m \ge 2$ is required. The behavior of the relative entropy is otherwise not known and the set may contain other points. If the procedure is initialized sufficiently close to the true parameter ϕ^0 then $\hat{\phi}_n$ is expected to converge to ϕ^0 with high probability. Assuming that $\hat{\phi}_n \to \phi^0$, and some mild regularity conditions are satisfied, it was shown in [279, Lemma 2, Theorem 2] that the averaged estimator

$$\overline{\phi}_n = \frac{1}{n}\sum_{l=1}^{n} \hat{\phi}_l \quad (10.9)$$

converges at rate $n^{-1/2}$ and has similar asymptotic properties as the off-line MSDLE obtained from maximization of (6.7). The latter estimator is asymptotically normal and it performs similarly to the ML estimator [274].

A recursion for HMP parameter estimation using prediction error techniques was proposed by Collings, Krishnamurthy, and Moore [64] and demonstrated empirically to provide fast convergence.

XI. Signal Classification

In recent years, a series of papers on universal classification of Markov chains was published. Ziv [328] studied testing of a simple hypothesis from which a training sequence is available against a composite hypothesis in the set of all Markov chains up to a given order. He developed an asymptotically optimal test in the Neyman–Pearson sense. Gutman [155] characterized the tradeoffs between the best exponents of the two kinds of errors. He also extended the approach to multiple hypotheses from which training sequences are available and allowed rejection of all hypotheses. He developed a test with asymptotically vanishing error and reject probabilities. The generalized likelihood ratio test (GLRT), which relies on ML estimates of the unknown sources, was used in [155]. This test was implemented using empirical entropies. Merhav [237] developed a Bayesian approach for multiple hypotheses testing of first-order Markov chains using estimates of their transition matrices and studied its performance.

Optimality of the GLRT in testing a simple hypothesis, say $P = P_0$, against a composite hypothesis, say $P = P_1 \in \mathcal{P}$, where $P_0 \cup \mathcal{P}$ is a *subset* of all stationary ergodic mth-order Markov measures, was studied by Zeitouni, Ziv, and Merhav [325]. A version of the Neyman–Pearson criterion was used in which both error probabilities approach zero exponentially fast with the number of observations. It was shown that if $P_0 \cup \mathcal{P}$ is closed with respect to exponential combinations of P_0 and P_1, i.e., if for every $P_1 \in \mathcal{P}$, and every $\alpha \in [0, 1]$

$$Q_\alpha = C_\alpha P_0^\alpha P_1^{1-\alpha} \in P_0 \cup \mathcal{P}$$

where C_α is a normalization factor that makes Q_α a pmf, then the GLRT is asymptotically optimal in the above described sense [325, Theorem 2]. A closely related condition developed by Gutman (cited in [325]) is necessary and sufficient for asymptotic optimality of the GLRT. Whether the GLRT is optimal for classification of HMPs even with a finite alphabet is still an open problem.

Classification problems involving HMPs were studied by several authors. Merhav [235] studied a binary hypothesis testing problem for two statistically independent observation sequences to emerge from the same general HMP or from two different general HMPs. The observation conditional densities of the HMPs were assumed members of the exponential family (Koopman–Darmois). A modified GLRT was developed and was shown to be asymptotically optimal in a Neyman–Pearson sense. Kieffer [187] provided a strongly consistent code-based approach for identifying whether or not a given observation sequence y^n with unknown distribution was generated by a member of a finite class of constrained finite-state sources $\mathcal{P}_\phi$. Finite-alphabet HMPs are special cases of that class.

Nádas [244] studied a classification problem in which a test sequence $\boldsymbol{x} = x^n$ is generated by one out of J possible general HMPs whose parameters $\{\phi^{(1)}, \ldots, \phi^{(J)}\}$ are not explicitly known. A set of J training sequences $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_J\}, \boldsymbol{y}_j = y^{n_j}$, from the J HMPs is assumed available. The goal is to identify the HMP that generated $\boldsymbol{x}$ with minimum probability of error. Nádas developed a Bayesian approach assuming that the parameters $\{\phi^{(i)}\}$ are statistically independent random variables. In addition, $\{\boldsymbol{y}_j\}$ are statistically independent given $\{\phi^{(i)}\}$, and $\boldsymbol{x}$ and $\{\boldsymbol{y}_j\}$ are statistically independent given $\{\phi^{(j)}\}$ and the active source. All J hypotheses were assumed equally likely. He showed that the optimal decision rule is given by

$$\begin{aligned} j &= \arg\max_i p(\boldsymbol{x}|\boldsymbol{y}_i) \\ &= \arg\max_i \frac{\int p\left(\boldsymbol{x}, \boldsymbol{y}_i|\phi^{(i)}\right) p\left(\phi^{(i)}\right) d\phi^{(i)}}{\int p\left(\boldsymbol{y}_i|\phi^{(i)}\right) p\left(\phi^{(i)}\right) d\phi^{(i)}}. \end{aligned} \tag{11.1}$$

Merhav and Ephraim [238] proposed an approximation to this decision rule that does not require integration and explicit priors for the parameters. The approximate Bayesian decision rule is given by

$$j = \arg\max_i \frac{\max\limits_{\phi^{(i)}} p\left(\boldsymbol{x}, \boldsymbol{y}_i; \phi^{(i)}\right)}{\max\limits_{\phi^{(i)}} p\left(\boldsymbol{y}_i; \phi^{(i)}\right)}. \tag{11.2}$$

The ratio of the two maxima comprises a similarity measure between the test and training data. The ratio is likely to be larger for $\boldsymbol{x}$ and $\boldsymbol{y}_i$ emerging from the same HMP than for $\boldsymbol{x}$ and $\boldsymbol{y}_i$ originating from different HMPs. This decision rule is similar to universal decision rules developed by Ziv [328] and Gutman [155]. It was shown in [238, Theorem 1], under some regularity conditions, that the decision rules (11.1) and (11.2) have the same asymptotic behavior as the length of the test sequence $n \to \infty$. Furthermore, for HMPs with positive transition probabilities and a set of training sequences whose lengths $\{n_i\}$ grow at least linearly with the length n of the test sequence, the decision rule (11.1), and hence (11.2), provides exponentially decaying probability of error as $n \to \infty$. The error exponent in both cases is the same. When $n_i \gg n$

$$\arg\max_{\phi^{(i)}} p(\boldsymbol{x}, \boldsymbol{y}_i; \phi^{(i)}) \approx \arg\max_{\phi^{(i)}} p(\boldsymbol{y}_i; \phi^{(i)})$$

and (11.2) can be further approximated as

$$j = \arg\max_i p\left(\boldsymbol{x}; \hat{\phi}^{(i)}\right) \tag{11.3}$$

where $\hat{\phi}^{(i)}$ maximizes $p(\boldsymbol{y}_i; \phi^{(i)})$ over $\phi^{(i)} \in \Phi$. This is the standard plug-in decision rule used in HMP-based classification such as in automatic speech recognition applications, see, e.g., (14.6). The condition of $n_i \gg n$ is commonly satisfied in classification problems that are based on off-line training. Without this simplification, implementation of the decision rule (11.2) is hard since it requires on-line global maximization of the two likelihood functions.

Kehagias [183] studied a sequential classification problem. A set of HMPs is assumed given but the test sequence is a sample from a stationary ergodic process that is not necessarily an HMP. The goal is to recursively identify the HMP that is closest to the test sequence in the minimum relative entropy sense. A recursive algorithm was developed for associating a test sequence $\{x_t, t = 1, 2, \ldots\}$ with an HMP from a given set of finite or countably infinite HMPs. The algorithm was derived under the assumption that the test sequence was produced by one of the HMPs. The analysis of the algorithm, however, does not make this assumption. Let Z be a discrete random variable taking values in $\{1, 2, \ldots\}$. Let $q_t(j) = P(Z = j|x^t)$. Let $j(t)$ denote the jth HMP selected at time t according to

$$j(t) = \arg\max_i \, q_t(i). \tag{11.4}$$

The conditional probability $q_t(j)$ is recursively calculated using

$$q_{t+1}(j) = \frac{p\left(x_{t+1}|x^t;\, \phi^{(j)}\right) q_t(j)}{\sum\limits_i p\left(x_{t+1}|x^t;\, \phi^{(i)}\right) q_t(i)} \tag{11.5}$$

where $\phi^{(j)}$ is the parameter of the HMP associated with the jth hypothesis and $p(x_{t+1}|x^t;\, \phi^{(j)})$ can be recursively calculated using (4.4) and (4.30).

In analyzing the algorithm, the test sequence was assumed to be a sample from a finite-alphabet stationary ergodic process. The HMPs were assumed to have a finite alphabet and for each j the parameter $\phi^{(j)} \in \Phi_\delta$. Almost sure convergence of the recursive classification approach, as $t \to \infty$, to the hypothesis whose HMP is closest to the test sequence in the relative entropy rate sense was proved in [183, Theorem 2]. If there is more than one HMP that achieves the same minimum relative entropy rate with respect to the test sequence, then convergence is to the set of all such HMPs. This situation may occur when the HMPs are not identifiable.

Giudici, Rydén, and Vandekerkhove [139] applied standard χ^2 asymptotic theory to the GLRT for two composite hypotheses testing problems involving the parameter $\phi \in \Phi \subseteq \mathcal{R}^d$ of an HMP. They used the asymptotic results of Bickel, Ritov, and Rydén [36]. Let ϕ^0 denote the true parameter. In the first problem, a simple null hypothesis H_0: $\phi = \phi^0$ and an alternative hypothesis H_1: $\phi \neq \phi^0$ were tested. Next, let $\Phi_0 \subseteq \Phi$ and assume that Φ_0 is characterized by a set of constraints $R_i(\phi) = 0, i = 1, \ldots, J$, where $J \leq d$. In the second problem, a composite null hypothesis H_0: $\phi \in \Phi_0$ and an alternative hypothesis H_1: $\phi \in \Phi \backslash \Phi_0$ were tested. Let $L_n(\phi) = \log p(y^n;\, \phi)$ be the log likelihood of the HMP and let $\hat{\phi}_n \in \Phi$ denote the ML estimate of ϕ^0 as obtained from a sample of n observations. The likelihood ratio test used for the simple null hypothesis is given by

$$\lambda_n = 2\left\{L_n\left(\hat{\phi}_n\right) - L_n\left(\phi^0\right)\right\}. \tag{11.6}$$

Under H_0, and for large n, λ_n has approximately a χ^2 distribution with d degrees of freedom. Hence, a test with size approximately equal to α is obtained if H_0 is rejected when $\lambda_n > \chi^2_{d,\,1-\alpha}$, where $\chi^2_{d,\,1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2 distribution with d degrees of freedom. The likelihood ratio used for the composite null hypothesis problem is given by

$$\lambda_n = 2\left\{\sup_{\phi \in \Phi} L_n(\phi) - \sup_{\phi \in \Phi_0} L_n(\phi)\right\}. \tag{11.7}$$

Under H_0, and for large n, λ_n has approximately a χ^2 distribution with r degrees of freedom. Hence, a test with size approximately equal to α is obtained if H_0 is rejected when $\lambda_n > \chi^2_{r,\,1-\alpha}$.

XII. Signal Estimation

Let $\{Y_t\}$ and $\{W_t\}$ denote observation sequences from two statistically independent general HMPs. Assume that $\{Y_t\}$ is a desired signal and $\{W_t\}$ is a noise process. Let $Z_t = Y_t + W_t$ for $t = 1, 2, \ldots$. In this section, we review MMSE estimation of Y_t from Z^n, $n \geq t$. The problem arises in applications such as enhancement of noisy speech signals [105], channel decoding [252], and forecasting in econometrics [156, Ch. 22].

It is easy to check that the noisy signal $\{Z_t\}$ is an HMP [105], [313]. Let $\mathcal{S}$ and $\tilde{\mathcal{S}}$ denote the state spaces of $\{Y_t\}$ and $\{W_t\}$, respectively. The state space of $\{Z_t\}$ is given by $\overline{\mathcal{S}} = \mathcal{S} \times \tilde{\mathcal{S}}$. Let $\{S_t\}$ and $\{\tilde{S}_t\}$ denote the state sequences of $\{Y_t\}$ and $\{W_t\}$, respectively. Let $\{\overline{S}_t = (S_t,\, \tilde{S}_t)\}$ denote the state sequence of $\{Z_t\}$. We refer to $\overline{S}_t$ as a *composite* state of the noisy process at time t. The MMSE estimator of Y_t given a realization z^n of the noisy signal is given by [105]

$$\begin{aligned}\hat{Y}_t &= E\{Y_t|z^n\} \\ &= \sum_{\overline{s}_t} p(\overline{s}_t|z^n) E\{Y_t|\overline{s}_t,\, z_t\}.\end{aligned} \tag{12.1}$$

The conditional probabilities $p(\overline{s}_t|z^n)$ can be calculated using a forward–backward recursion from Section V-A. A similar estimator was developed by Magill [229] for a mixture of stationary ergodic processes where the state remains constant in its initially chosen value. Suppose that Y_t and W_t are k-dimensional vectors in $\mathcal{R}^k$, and that the observation conditional densities of $\{Y_t\}$ and $\{W_t\}$ are Gaussian with zero mean and covariance matrices $\{R_{s_t}\}$ and $\{\tilde{R}_{\tilde{s}_t}\}$, respectively. Then, the observation conditional densities of $\{Z_t\}$ are also Gaussian with zero mean and covariance matrices $\{\overline{R}_{\overline{s}_t} = R_{s_t} + \tilde{R}_{\tilde{s}_t}\}$. Furthermore

$$E\{Y_t|\overline{s}_t,\, z_t\} = R_{s_t}\left[R_{s_t} + \tilde{R}_{\tilde{s}_t}\right]^{-1} z_t \tag{12.2}$$

which is the Wiener estimator for Y_t given $\{\overline{s}_t,\, z_t\}$.

The causal MMSE estimator $E\{Y_t|z^t\}$ was analyzed by Ephraim and Merhav [104]. The MMSE given by

$$\overline{\varepsilon_t^2} = \frac{1}{k}\,\mathrm{tr}\, E\left\{\left(Y_t - \hat{Y}_t\right)\left(Y_t - \hat{Y}_t\right)'\right\} \tag{12.3}$$

was expressed as the sum of two terms denoted by $\overline{\xi_t^2}$ and $\overline{\eta_t^2}$. The first term $\overline{\xi_t^2}$ represents the average MMSE of the estimator that is informed of the exact composite state of the noisy signal Z_t and is given by

$$\overline{\xi_t^2} = \frac{1}{k}\,\mathrm{tr}\, E\left\{\mathrm{cov}\left(Y_t|\overline{S}_t,\, Z_t\right)\right\}. \tag{12.4}$$

The term $\overline{\eta_t^2}$ represents a sum of cross error terms for which no explicit expression is known. Tight lower and upper bounds on $\overline{\eta_t^2}$ were developed. For signal and noise HMPs with Gaussian observation conditional densities, these bounds were shown to approach zero at the same exponential rate as $k \to \infty$. The exponential rate is the same as that of the error probability for distinguishing between pairs of composite states.

Several other estimators for the signal Y_t from z^n were developed [105]. We note, in particular, the detector–estimator scheme proposed by Ephraim and Merhav [104] in which the composite state of the noisy signal is first estimated and then

MMSE signal estimation is performed. This estimator is given by $\hat{Y}_{t|\overline{s}_t^\star} = E\{Y_t|\overline{s}_t^\star, z_t\}$ where $\overline{s}_t^\star = \arg\max_{\overline{s}_t} p(\overline{s}_t|z^t)$. The MSE of this estimator approaches $\overline{\xi}_t^2$ when $k \to \infty$ and hence the estimator is asymptotically optimal in the MMSE sense. An estimator similar to (12.1) and (12.2) was used by Crouse, Nowak, and Baraniuk [69] for wavelet denoising of signals contaminated by white noise.

XIII. Hidden Markov Channels

FSCs were defined in Section IV-B4. An FSC with input $\{X_t\}$, output $\{Y_t\}$, and state sequence $\{C_t\}$ has a conditional transition density given by

$$p(y^n, c^n|x^n, c_0) = \prod_{t=1}^{n} p(y_t, c_t|c_{t-1}, x_t). \qquad (13.1)$$

The memory of the channel is captured by the Markov chain $\{C_t\}$. The states may represent fading levels as in wireless communications [205], previous channel inputs as in intersymbol interference channels, or a tendency of the channel to persist in a given mode as for bursty channels [133, Sec. 4.6]. FSCs are also encountered when a buffer exists at the input of a channel, in which case the states correspond to the buffer contents [89]. FSCs may be interpreted as hidden Markov channels since the state sequence is not known at the encoder and decoder. Posterior probabilities of the states can be calculated using recursions similar to those given in Section V-A [133, eq. 4.6.1], [141]. In this section, we focus on FSCs with finite input and output spaces, $\mathcal{X}$ and $\mathcal{Y}$, respectively, and review some of their properties and the Lapidoth–Ziv universal decoding algorithm [204]. A thorough discussion on reliable communication under channel uncertainties can be found in Lapidoth and Narayan [205].

The channel coding theorem for FSCs was derived by Gallager [133] and by Blackwell, Breiman, and Thomasian [43]. FSCs for which the effect of the initial state is rapidly forgotten are said to be *indecomposable*. A necessary and sufficient condition for an FSC to be indecomposable is that for some fixed n and each $x^n \in \mathcal{X}^n$ there exists a choice for the nth state, say c_n, such that $p(c_n|x^n, c_0) > 0$ for all $c_0 \in \mathcal{C}$ [133, Theorem 4.6.3]. If the FSC is indecomposable or if $\pi(c_0) > 0$ for every $c_0 \in \mathcal{C}$, the capacity of the channel is given by [133, Theorem 4.6.4], [205, Theorem 8]

$$C_{\text{FSC}} = \lim_{n\to\infty} \frac{1}{n} \max_{p(x^n)} \min_{c_0} I(X^n; Y^n|c_0) \qquad (13.2)$$

where $I(X^n; Y^n|c_0)$ denotes the conditional mutual information between the input and output of the channel for a given initial state c_0. Sequences of upper and lower bounds for C_{FSC}, which can be used to approximate the capacity to an arbitrary degree, were provided in [133, Theorem 5.9.2]. For any FSC, code rate $R < C_{\text{FSC}}$, and sufficiently large n, there exists a $(2^{nR}, n)$ code of 2^{nR} codewords of length n each that provides exponentially decaying probability of decoding error for any input message and initial state c_0 [133, Theorem 5.9.2]. If $R > C_{\text{FSC}}$, the probability of error cannot be made arbitrary small, independent of the initial state [133, Theorem 4.6.2].

The Gilbert–Elliott channel defined in Section IV-B5 is an example of an FSC. The capacity of this channel was calculated by Mushkin and Bar-David [243, Proposition 4]. Recall that the channel introduces an additive hidden Markov noise process, say $\{Z_t\}$. Let $\overline{H}(Z)$ denote the entropy rate of $\{Z_t\}$. Assume that the parameter characterizes the memory of the channel satisfies $|v| < 1$. The capacity of the channel is given by

$$C_{\text{GEC}} = 1 - \overline{H}(Z) = 1 - \lim_{n\to\infty} H(Z_n|Z^{n-1}). \qquad (13.3)$$

Convergence of $H(Z_n|Z^{n-1})$ occurs at an exponential rate as shown in [40], [161], see also Section IV-E. The capacity C_{GEC} increases monotonically with $v \in [0, 1)$. It ranges from the capacity of a memoryless channel ($v = 0$) to the capacity of a channel informed about its Markov state ($v = 1$). A decision-feedback decoder that achieves capacity was developed in [243].

A class of channels related to FSCs was studied by Ziv [327]. A channel in that class is described by the conditional transition pmf

$$p(y^n|x^n, c_0) = \prod_{t=1}^{n} p(y_t|c_t, x_t) \qquad (13.4)$$

and a deterministic next-state function

$$c_t = g(y_{t-1}, x_{t-1}, c_{t-1}).$$

Ziv developed an asymptotically optimal universal decoding approach for these channels. The same algorithm was shown by Lapidoth and Ziv [204] to be asymptotically optimal for FSCs described by (13.1). These results and the universal decoder are described next.

Let Θ denote the parameter space of all FSCs with common spaces $(\mathcal{X}, \mathcal{Y}, \mathcal{C})$. The parameter of each channel comprises an initial state $c_0 \in \mathcal{C}$ and all transition probabilities of the form $p(y_t, c_t|c_{t-1}, x_t)$. Consider an FSC with parameter $\theta \in \Theta$. Let $B_n \subset \mathcal{X}^n$ denote a permutation invariant subset of $\mathcal{X}^n$ in the sense that if $\boldsymbol{x} \in B_n$ then any permutation of the components of $\boldsymbol{x}$ results in a vector in B_n. Assume that a set of 2^{nR} n-length codewords $\{\boldsymbol{x}(i)\}$ are drawn uniformly and independently from B_n where R denotes the rate of the code. The collection of these codewords is referred to as a codebook. Let $\overline{P}_{\theta, ml}(\text{error})$ denote the probability of error of the ML decoder for the FSC θ averaged over all 2^{nR} messages and possible codebooks. Similarly, let $\overline{P}_{\theta, z}(\text{error})$ denote the average probability of error when Ziv's decoder is applied to the same channel without explicitly knowing its parameter θ. From [204, Theorem 1]

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{\overline{P}_{\theta, z}(\text{error})}{\overline{P}_{\theta, ml}(\text{error})} = 0. \qquad (13.5)$$

Let $\mathcal{D}_n$ be a deterministic code of 2^{nR} n-length codewords in B_n. Let $P_{\theta, z}(\text{error}|\mathcal{D}_n)$ and $P_{\theta, ml}(\text{error}|\mathcal{D}_n)$ denote, respectively, the probabilities of error for the particular code $\mathcal{D}_n$ using Ziv's decoder and the ML decoder. These error probabilities are averaged over the messages only. It was shown in [204, Theorem 1] that there exists such a deterministic code for which

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{P_{\theta, z}(\text{error}|\mathcal{D}_n)}{P_{\theta, ml}(\text{error}\mathcal{D}_n)} = 0. \qquad (13.6)$$

Admissibility of universal decoding for channels with memory, and in particular for FSCs, was studied by Feder and Lapidoth [112, Theorem 3].

Assume that a codebook $\{\boldsymbol{x}(i)\} \subset \mathcal{X}^n$ of 2^{nR} n-length codewords was drawn at random by the encoder and that a copy of the codebook is available at the decoder. The ML decoder for a given FSC $\theta \in \Theta$ decodes the received signal $\boldsymbol{y} \in \mathcal{Y}^n$ as coming from the ith message if

$$i = \arg\max_j p(\boldsymbol{y}|\boldsymbol{x}(j);\, \theta) \tag{13.7}$$

where $p(\boldsymbol{y}|\boldsymbol{x}(j);\, \theta)$ is the channel's pmf (4.17) specified for the given θ. If the maximum is not unique an error is declared. Ziv's decoder does not explicitly use θ in decoding the channel. Instead, a length function $u(\boldsymbol{x}(i),\, \boldsymbol{y})$ is calculated for each of the codewords $\{\boldsymbol{x}(i)\}$ and the received signal $\boldsymbol{y}$. The observed signal $\boldsymbol{y}$ is decoded as coming from the ith message if

$$i = \arg\min_j u(\boldsymbol{x}(j),\, \boldsymbol{y}). \tag{13.8}$$

If the minimum is not unique an error is declared. The length function $u(\boldsymbol{x},\, \boldsymbol{y})$ is calculated from joint parsing of $(\boldsymbol{x},\, \boldsymbol{y})$ much like the parsing in the Lempel–Ziv universal data compression algorithm described in Section VI-E. This length function is described next.

Let $c(\boldsymbol{x},\, \boldsymbol{y})$ denote the number of distinct phrases in $(\boldsymbol{x},\, \boldsymbol{y})$. The joint parsing of $(\boldsymbol{x},\, \boldsymbol{y})$ induces parsing of $\boldsymbol{y}$ into phrases that are not necessarily distinct. Let $c(\boldsymbol{y})$ denote the number of distinct phrases in the induced parsing of $\boldsymbol{y}$. Let $y(l)$, $1 \le l \le c(\boldsymbol{y})$, denote lth distinct phrase in the induced parsing of $\boldsymbol{y}$. Let $\boldsymbol{x}$ be parsed identically to $\boldsymbol{y}$ in the sense that if

$$\boldsymbol{y} = y_1^{l_1} y_{l_1+1}^{l_2} \cdots \boldsymbol{x}_{l_{k-1}+1}^{l_k}$$

then

$$\boldsymbol{x} = x_1^{l_1} x_{l_1+1}^{l_2} \cdots \boldsymbol{x}_{l_{k-1}+1}^{l_k}$$

where k is the total number of phrases in parsing $(\boldsymbol{x},\, \boldsymbol{y})$ of which at least $k-1$ phrases are distinct, i.e., $k-1 \le c(\boldsymbol{x},\, \boldsymbol{y}) \le k$. Let $c_l(\boldsymbol{x}|\boldsymbol{y})$ denote the number of distinct phrases in the parsing of $\boldsymbol{x}$ that appear jointly with $y(l)$. We have that

$$\sum_{l=1}^{c(\boldsymbol{y})} c_l(\boldsymbol{x}|\boldsymbol{y}) = c(\boldsymbol{x},\, \boldsymbol{y}). \tag{13.9}$$

The length function $u(\boldsymbol{x},\, \boldsymbol{y})$ required by the decision rule (13.8) is defined as

$$u(\boldsymbol{x},\, \boldsymbol{y}) = \sum_{l=1}^{c(\boldsymbol{y})} c_l(\boldsymbol{x}|\boldsymbol{y}) \log c_l(\boldsymbol{x}|\boldsymbol{y}). \tag{13.10}$$

These concepts are well demonstrated by the following example borrowed from [204]. Let $\mathcal{X} = \{0,\, 1\}$, $\mathcal{Y} = \{a,\, b\}$, and $n = 6$. Consider $\boldsymbol{x} = 010001$ and $\boldsymbol{y} = \mathrm{ababab}$. The joint parsing of $(\boldsymbol{x},\, \boldsymbol{y})$ yields $c(\boldsymbol{x},\, \boldsymbol{y}) = 4$ distinct phrases as shown below.

$$\begin{array}{lcl} \boldsymbol{x} = 0 & | & 1 & | & 0 \;\; 0 & | & 0 \;\; 1 \\ \boldsymbol{y} = \mathrm{a} & | & \mathrm{b} & | & \mathrm{a} \;\; \mathrm{b} & | & \mathrm{a} \;\; \mathrm{b} \end{array}. \tag{13.11}$$

The induced parsing of $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by

$$\begin{aligned} \boldsymbol{x} &= 0,\, 1,\, 00,\, 01 \\ \boldsymbol{y} &= \mathrm{a},\, \mathrm{b},\, \mathrm{ab},\, \mathrm{ab}. \end{aligned} \tag{13.12}$$

There are $c(\boldsymbol{y}) = 3$ distinct phrases for $\boldsymbol{y}$. These phrases and their joint occurrences are given by

$$\begin{aligned} &y(1) = \mathrm{a}, \quad y(2) = \mathrm{b}, \quad y(3) = \mathrm{ab} \\ &c_1(\boldsymbol{x}|\boldsymbol{y}) = 1, \quad c_2(\boldsymbol{x}|\boldsymbol{y}) = 1, \quad c_3(\boldsymbol{x}|\boldsymbol{y}) = 2. \end{aligned} \tag{13.13}$$

From (13.10), $u(\boldsymbol{x},\, \boldsymbol{y}) = 2$ when the logarithm's base is 2.

An analogue of Ziv's inequality for FSCs can be inferred from Merhav [239, eqs. (7)–(9)]. Let $\theta \in \Theta$ denote the parameter of an FSC with finite spaces $(\mathcal{X},\, \mathcal{Y},\, \mathcal{C})$. It holds that

$$\max_{\theta \in \Theta} p(\boldsymbol{y}|\boldsymbol{x};\, \theta) \le 2^{-u(\boldsymbol{x},\, \boldsymbol{y}) + n(\epsilon_1(n,\, k) + \epsilon_2(k))} \tag{13.14}$$

where k is some integer that divides n, $\epsilon_1(n,\, k)$ and $\epsilon_2(k)$ are independent of $\boldsymbol{x}$ and $\boldsymbol{y}$, and $\lim_{k\to\infty} \lim_{n\to\infty} \epsilon_1(n,\, k) = 0$ and $\lim_{k\to\infty} \epsilon_2(k) = 0$. This result was used in [239] in a binary hypothesis testing problem for deciding whether a given channel output sequence was produced by a prescribe input sequence or by an alternative sequence. A decision rule similar to (8.4) was used.

A composite hypothesis testing approach applicable for decoding of unknown channels from a given family, in the relative minimax sense, was developed by Feder and Merhav [113]. FSCs are particular cases of that family. In this approach, the ratio of the probability of error of a decoder that is independent of the unknown channel parameter, and the minimum achievable probability of error for the channel, is optimized in the minimax sense. The optimal decision rule is obtained from minimization of the maximum of this ratio over all possible channel parameters. Asymptotically optimal decoders that are easier to implement were also derived in [113].

XIV. Selected Applications

One of the earliest applications of HMPs and their theory was in ecology. In 1967, Baum and Eagon [26] developed an iterative procedure for local maximization of the likelihood function of a finite-alphabet HMP. This procedure predated the EM approach developed in 1970 by Baum, Petrie, Soules, and Weiss [28]. Baum and Eagon observed that the likelihood function $p(y^n;\, \phi)$ is a homogeneous polynomial of degree $2n+1$ in the components $\{\pi_j,\, a_{ij},\, b_{ij}(l)\}$ of the parameter ϕ where

$$b_{ij}(l) = P(Y_t = l | S_{t-1} = i,\, S_t = j)$$

as in (4.9). In estimating $\{a_{ij}\}$, for example, they showed that the mapping $a_{ij} \to \overline{a}_{ij}$ from the domain

$$D = \left\{ a_{ij} \colon a_{ij} \ge 0, \sum_j a_{ij} = 1, i,\, j = 1,\, \ldots,\, M \right\}$$

into itself defined by

$$\overline{a}_{ij} = \frac{a_{ij} \frac{\partial p}{\partial a_{ij}}}{\sum_{j=1}^{M} a_{ij} \frac{\partial p}{\partial a_{ij}}} \tag{14.1}$$

increases the likelihood function unless a stationary point in D is reached [26], [29, Theorem 2]. The transformation (14.1) was named *growth transformation* by Baum and Sell [27] and its properties were studied. The recursion (14.1) turned out to be similar to a recursion developed in ecology for predicting the rate of population growth. Baum [29] showed that (14.1) and the re-estimation formula (6.15) for a_{ij} coincide. Similar conclusions hold for the re-estimation formulas for π_j and $b_{ij}(l)$.

Concurrently with the above application, a new application in the area of automatic character recognition emerged at IBM.

Raviv [265] studied this problem and developed the stable forward recursion (5.14) as well as a stable recursion similar to (5.9) for calculating $p(s_t|y^{t+1})$. Subsequently, a major application of HMPs to automatic speech recognition was undertaken at IBM. Jelinek, Bahl, Mercer, and Baker [171], [18], [172], [21], along with their colleagues, developed the first automatic speech recognition system based on hidden Markov modeling of speech signals. They also studied language modeling using Markov chains [173]. Language modeling using HMPs was studied by Cave and Neuwirth [55]. Numerous papers and several books were published on automatic speech recognition, see, e.g., [19], [262], [165], [263], [209], [173], [65], [230] and the references therein. Moreover, HMP-based automatic speech recognition software packages running on personal computers are now commercially available, see, e.g., *Via Voice* by IBM and *Naturally Speaking* by Dragon Systems. In the process of studying applications of HMPs to automatic speech recognition, several extensions of HMPs and new parameter estimation approaches were developed. These are briefly discussed in Sections XIV-A and XIV-B, respectively. In Section XIV-A, we also mention extensions of HMPs developed for other non-speech processing applications.

In recent years, numerous new applications of HMPs have emerged in many other areas, particularly in communications and information theory, econometrics, and biological signal processing. In some applications, the underlying processes are naturally HMPs. In others, HMPs were found reasonable statistical models for the underlying processes. In either cases, the readily available theory of HMPs provides elegant and often intuitive solutions. We briefly review these applications in Sections XIV-C–XIV-E. Additional applications can be found in [66] and [228]. For each application, we attempted to provide the original references as well as papers of tutorial nature. Unfortunately, it is impractical to provide an exhaustive list of references for each application due to the huge number of publications in each area.

A. Special HMPs

In some applications, the data associated with each state is overdispersed relative to any single density such as Gaussian or Poisson. Using an observation conditional density that is a mixture of J densities for each state may circumvent this problem [221], [256], [175], [262]. Such modeling results in two regime variables, S_t for the state at time t and H_t for the mixture component in state S_t. Using the standard conditional independence assumption (4.1) of observations given states we have

$$p(y^n|s^n) = \prod_{t=1}^{n} \sum_{h_t=1}^{J} p(h_t|s_t)p(y_t|s_t,\, h_t). \tag{14.2}$$

Let $c_{l|j} = P(H_t = l|S_t = j)$ denote the probability of choosing the lth mixture component in the jth state. Multiplying (14.2) by $p(s^n)$ and summing over s^n, and using (4.7), we obtain

$$p(y^n) = \sum_{s^n} \sum_{h^n} \prod_{t=1}^{n} a_{s_{t-1}s_t} c_{h_t|s_t} b(y_t|s_t,\, h_t). \tag{14.3}$$

Comparing (14.3) with (4.3) reveals that there is no principal difference between HMPs with a single or multiple mixture components per state. The use of multiple mixture components per state allows one to increase the number of observation conditional densities of the HMP without incurring a quadratic increase in the number of components of $\{\pi_i,\, a_{ij}\}$.

It has often been found useful to restrict the allowable transitions of the Markov chain. For example, left–right HMPs are commonly used in automatic speech recognition [262], [111]. In this case, the transition matrix is upper triangular or has nonzero elements only on the main diagonal and first off-diagonal. For a left–right HMP, all but the last state are transient states. The last state is absorbing. It constitutes a degenerate irreducible Markov chain. Left–right Markov chains are used for two reasons. First, this choice is natural in modeling speech signals, as states evolve in a manner that parallels the evolvement of the acoustic signal in time. Second, an HMP with a left–right Markov chain and the usual parametrization is characterized by a lower dimensional parameter compared to that of an HMP with positive transition probabilities. Such reduction in the parameter size helps preventing overfitting of the model to training data. Left–right HMPs are also mathematically tractable as was shown in Section VII-D.

Inherent to an HMP is a geometric distribution for the number d of consecutive time periods that the process spends in a given state j before leaving that state. This distribution is given by $p(d|j) = a_{jj}^{d-1}(1 - a_{jj})$. In some applications it was found useful to turn off self-state transitions $(a_{jj} = 0)$ and introduce explicit distribution for d that suits better the problem at hand, see Ferguson [115]. This approach was applied to automatic speech recognition [216], DNA sequencing [227], detection of ECG events [298], and seismic signal modeling [145]. Examples of possible distributions used for d include Poisson, binomial, and gamma [115], [216]. The resulting hidden component of the model is referred to as *semi-Markov* chain [145]. Let d_j denote a possible occupation time of state j and define $\overline{d}_0 = 0$ and $\overline{d}_m = \sum_{j=1}^{m} d_j$ for some integer m. Using standard conditional independence assumptions, and the simplifying assumption that an integer number of state transitions occurred in n time periods, we have

$$p(y^n) = \sum_{m} \sum_{d^m:\, \overline{d}_m = n} \sum_{s^m} \prod_{t=1}^{m} p(s_t|s_{t-1})p(d_t|s_t)p(\boldsymbol{y}_t|s_t,\, d_t) \tag{14.4}$$

where $\boldsymbol{y}_t = \{y_{\overline{d}_{t-1}+1},\, \ldots,\, y_{\overline{d}_t}\}$. This model can be seen as a standard HMP with an extended state space of $M \times D$ elements, where D is the largest possible duration. Extension of the Baum algorithm for estimating the parameter and state sequence of this model was proposed by Ferguson [115]. An alternative ML approach was provided by Goutsias and Mendel [145].

The next two extensions of HMPs were developed in biological signal processing and image processing, respectively. We have seen in Section IV-B3 that the observation conditional densities of HMPs may be dependent on past observations in addition to the current state of the Markov chain. A stronger assumption was necessary in a neurophysiology application where ion channel currents observed in colored noise were recorded from living cells [306], [307]. The model used for that application resulted in dependency of Y_t on Y_{t-m+1}^{t-1} as well as on S_{t-m+1}^{t}. The HMP is seen as a *vector* HMP and the sequence of states S_{t-m+1}^{t} is commonly referred to as a *metastate*. Note that a vector HMP is different from an HMP with vector observations as in Section IV-B1. In the latter case, vector observations are statistically independent given the state sequence.

Applications in coding of finite-alphabet images motivated the definition of a partially hidden Markov model by Forchhammer and Rissanen [122]. The hidden states of this process are supplemented by the so-called contexts which are subsequences of the observed signal. The partially hidden Markov model is defined by

$$p(y^n, s^n) = p(y_1, s_1) \prod_{t=2}^{n} p(y_t | s_t, r_{t-1}) p(s_t | s_{t-1}, x_{t-1}) \tag{14.5}$$

where $\{s_t\}$ is a state sequence and $\{r_t\}$ and $\{x_t\}$ are the contexts. The forward–backward and Baum algorithms extend to the processes mentioned above as was shown in the referenced papers.

B. Parameter Estimation in Speech Recognition

In this subsection, we briefly review three non-ML parameter estimation approaches that were tailored primarily to automatic speech recognition applications. We focus on the maximum mutual information (MMI) approach of Bahl, Brown, de Souza, and Mercer [20], the minimum discrimination information (MDI) approach of Ephraim, Dembo, and Rabiner [101], and the minimum empirical error rate (MEER) approach of Ephraim and Rabiner [102], Ljolje, Ephraim, and Rabiner [224], Juang and Katagiri [177], Chou, Juang, and Lee [58], [178], and Erlich [108]. See also Amari [8].

To motivate these approaches it is useful to review the role of HMPs in automatic speech recognition applications [173], [263], [165]. For simplicity, we discuss isolated word recognition only. Consider a vocabulary of J words. The density of the acoustic signal from each word is modeled as an HMP, and the parameter of the HMP is estimated from a training sequence of acoustic signals from that word. Let $p(\cdot; \phi)$ denote the density of an HMP with parameter $\phi \in \Phi$. Let $\boldsymbol{y}_i$ denote a training sequence of length n_i from the ith word. Let $\boldsymbol{\phi} = \{\phi^{(1)}, \ldots, \phi^{(J)}\}$ denote the J parameters of the HMPs for the J words. Let $\hat{\phi}^{(i)}$ denote an estimate of $\phi^{(i)}$ from $\boldsymbol{y}_i$. When ML estimation is used, $\hat{\phi}^{(i)} = \arg\max_{\phi \in \Phi} p(\boldsymbol{y}_i; \phi)$. All words are assumed *a priori* equally likely. A test acoustic signal y^n is associated with the jth word in the vocabulary if the signal is most likely to have been produced by the jth HMP, i.e.,

$$j = \arg \max_{1 \le i \le J} p\left(y^n; \hat{\phi}^{(i)}\right). \tag{14.6}$$

1) Maximum Mutual Information (MMI): MMI is a training approach in which the parameters of the J HMPs are *simultaneously* estimated, by minimizing the average empirical mutual information between the data and the hypotheses. The approach attempts to reduce the recognition error rate obtained when ML estimation is applied for individual estimation of each HMP. The MMI estimate of $\boldsymbol{\phi}$ is obtained from

$$\max_{\boldsymbol{\phi} \in \Phi^J} \sum_{j=1}^{J} \log \frac{p\left(\boldsymbol{y}_j; \phi^{(j)}\right)}{\sum\limits_{i=1}^{J} p\left(\boldsymbol{y}_j; \phi^{(i)}\right)}. \tag{14.7}$$

A re-estimation approach for MMI estimation was developed in [144]. It is based on a generalization of the growth transformation of Baum and Eagon [26] for maximization of homogeneous polynomials with nonnegative coefficients to maximization of rational functions. This approach requires specification of an exogenous constant whose practical value may result in slow convergent of the iterative approach [246]. Often this approach is implemented using general-purpose optimization procedures such as the steepest descent algorithm.

2) Minimum Discrimination Information (MDI): Discrimination information is synonymous to relative entropy, cross entropy, divergence, and the Kullback–Leibler number. The MDI approach is suitable for modeling one random process such as a speech signal by another parametric process such as an HMP. The distribution of the first process is not explicitly known. The process is characterized by a partial set of moments. The MDI approach attempts to choose the HMP that provides MDI with respect to the set of all distributions of the first process that satisfy the given moments. The MDI approach is a generalization of the maximum entropy inference approach [68, Ch. 11]. Shore and Johnson [291] showed that MDI is a logically consistent axiomatic modeling approach. See also Csiszár [72] for further justification.

Let $\{Y_t, t = 1, \ldots, n\}$, $Y_t \in \mathcal{R}^k$, denote a set of n vectors from a source whose distribution is not explicitly known. Suppose that a set of moment constraints is available for these vectors. For example, let m_t and R_t denote the true mean and covariance of Y_t. Suppose that m_t and a band of R_t are available for each $t = 1, \ldots, n$. The band may comprise an upper left block of R_t or the main diagonal and some off-diagonals of R_t. Let $\mathcal{G}$ denote the set of all n-dimensional distributions $\{G^{(n)}\}$ that satisfy the given moment constraints. Let $P_\phi^{(n)}$ denote the n-dimensional distribution of an HMP with parameter $\phi \in \Phi$. Let $g(y^n)$ and $p(y^n; \phi)$ denote the pdfs corresponding to $G^{(n)}$ and $P_\phi^{(n)}$, respectively. Let

$$D\left(G^{(n)} \middle\| P_\phi^{(n)}\right) = \int g(y^n) \log \frac{g(y^n)}{p(y^n; \phi)}\, dy^n \tag{14.8}$$

denote the discrimination information between $G^{(n)}$ and $P_\phi^{(n)}$. The HMP is estimated from

$$\min_{\phi \in \Phi} \min_{G^{(n)} \in \mathcal{G}} D\left(G^{(n)} \middle\| P_\phi^{(n)}\right). \tag{14.9}$$

There is no closed-form solution for this optimization problem even for HMPs with Gaussian observation conditional densities and second-order moment constraints considered in [101]. An iterative approach for alternate minimization of $D(G^{(n)} \| P_\phi^{(n)})$ over $G^{(n)} \in \mathcal{G}$ and $\phi \in \Phi$ was developed in [101] following a similar approach due to Csiszár and Tusnady in [71]. Given an HMP with parameter $\phi_m \in \Phi$ at the end of the mth iteration, a new estimate of the process distribution

$G^{(n)} \in \mathcal{G}$ can, in principle, be obtained from the solution of a set of nonlinear equations for the Lagrange multipliers. Let the complete data density of the new estimate of $G^{(n)}$ be denoted by $g(s^n, y^n; \phi_m)$. Next, a new estimate ϕ_{m+1} of the HMP parameter can be obtained from maximization over $\phi \in \Phi$ of the auxiliary function

$$Q(\phi, \phi_m) = \sum_{s^n} \int g(s^n, y^n; \phi_m) \log p(s^n, y^n; \phi)\, dy^n. \tag{14.10}$$

The procedure is repeated until a fixed point is reached or some stopping criterion is met. Local convergence of $\{\phi_m\}$ to a stationary point of the MDI measure was demonstrated in [101]. The general convergence proof from [71] is not applicable to this problem since the set of HMP distributions of a given order is not a convex set of probability measures.

While maximization of the auxiliary function in (14.10) results in re-estimation formulas similar to those obtained in the Baum algorithm, estimation of the distribution $G^{(n)} \in \mathcal{G}$ is a hard problem. If a single state sequence dominates the MDI measure, then the MDI approach coincides with the Baum–Viterbi algorithm.

3) Minimum Empirical Error Rate (MEER): The MEER approach simultaneously estimates the parameters of the J HMPs by minimizing the empirical error rate of the recognizer for the given J training sequences. This criterion is directly related to the goal of automatic speech recognition. The theory of empirical risk minimization and the design of optimal separating hyperplanes using support vector machines has recently attracted much attention, see Vapnik [304], [305]. The extension of Vapnik's work to HMPs is still an open problem.

In the MEER approach, the nondifferentiable indicator functions of the error rate expression are approximated by smooth differentiable functions and minimization is performed using numerical procedures such as the steepest descent algorithm. Let $q_j(y^n)$ denote the pdf of an observation sequence y^n from the acoustic signal of the jth word. Assume that the decision rule is based on estimates of the HMPs. The jth word is recognized if the acoustic signal y^n is in the set

$$\Psi_j(\boldsymbol{\phi}) = \left\{ y^n : p\left(y^n; \phi^{(j)}\right) > \max_{i \neq j} p\left(y^n; \phi^{(i)}\right) \right\}. \tag{14.11}$$

The probability of correct decision is given by

$$\begin{aligned} P_c(\boldsymbol{\phi}) &= \frac{1}{J} \sum_{j=1}^{J} \int_{\Psi_j(\boldsymbol{\phi})} q_j(y^n)\, dy^n \\ &= \frac{1}{J} \sum_{j=1}^{J} \int 1_{\Psi_j(\boldsymbol{\phi})}(y^n) q_j(y^n)\, dy^n \end{aligned} \tag{14.12}$$

where $1_{\Psi_j(\boldsymbol{\phi})}(y^n)$ denotes an indicator function defined by

$$1_{\Psi_j(\boldsymbol{\phi})}(y^n) = \begin{cases} 1, & \text{if } y^n \in \Psi_j(\boldsymbol{\phi}) \\ 0, & \text{otherwise.} \end{cases} \tag{14.13}$$

Let

$$L_j(y^n; \boldsymbol{\phi}) = \log \frac{p\left(y^n; \phi^{(j)}\right)}{\left[\sum_{i \neq j} \left(p\left(y^n; \phi^{(i)}\right)\right)^{\eta} \right]^{1/\eta}}, \qquad \eta \geq 1. \tag{14.14}$$

For large η

$$L_j\left(y^n; \boldsymbol{\phi}\right) \approx \log \frac{p\left(y^n; \phi^{(j)}\right)}{\max_{i \neq j} p\left(y^n; \phi^{(i)}\right)} \tag{14.15}$$

and the decision rule can be approximated as

$$\Psi_j(\boldsymbol{\phi}) \approx \{y^n : L_j(y^n; \boldsymbol{\phi}) > 0\}. \tag{14.16}$$

The indicator function (14.13) can similarly be approximated as

$$1_{\Psi_j(\boldsymbol{\phi})}(y^n) \approx \begin{cases} 1, & \text{if } L_j(y^n; \boldsymbol{\phi}) > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{14.17}$$

This approximation makes the argument $L_j(y^n; \boldsymbol{\phi})$ of the indicator function differentiable in $\boldsymbol{\phi}$. Next, the indicator function itself is approximated by the differentiable sigmoid function as follows:

$$1_{\Psi_j(\boldsymbol{\phi})}(y^n) \approx \frac{1}{1 + \exp\{-\xi L_j(y^n; \boldsymbol{\phi})\}}, \qquad \xi > 0. \tag{14.18}$$

If $q_j(y^n)$ is assumed to be concentrated on the training sequence $\boldsymbol{y}_j$ from the jth word, i.e., $q_j(y^n) = \delta(y^n - \boldsymbol{y}_j)$ and $\delta(\cdot)$ denotes the Dirac function, we obtain from (14.12) and (14.18) the desired differentiable approximation for the probability of correct decision as

$$P_c(\boldsymbol{\phi}) \approx \frac{1}{J} \sum_{j=1}^{J} \frac{1}{1 + \exp\{-\xi L_j(\boldsymbol{y}_j; \boldsymbol{\phi})\}}. \tag{14.19}$$

This estimate approximates the empirical correct decision count of the HMP-based recognizer. The parameter $\boldsymbol{\phi}$ of the HMPs is estimated from

$$\max_{\boldsymbol{\phi} \in \Phi^J} P_c(\boldsymbol{\phi}). \tag{14.20}$$

C. Communications and Information Theory

In this subsection, we review applications of HMPs in communications and information theory that we have not discussed previously in this paper.

1) Source Coding: Ott [248] proposed in 1967 a uniquely decodable code for a sequence $\{Y_t\}$ from a finite-alphabet HMP. The coder assumes zero channel errors. At each time t, identical Huffman-type codes are produced at the transmitter and receiver for encoding Y_t. The codes are based on the conditional pmf $p(y_t|y^{t-1})$ which is calculated using (4.4) and the recursion (4.30). This recursion was originally developed for that purpose by Ott.

Merhav [236] studied in 1991 lossless block-to-variable length source coding for finite-alphabet HMPs. He investigated the probability of codeword length overflow and competitive optimality of the Lempel–Ziv data compression algorithm [326]. Consider an HMP $\{Y_t\}$ with observation space $\mathcal{Y}$ and entropy rate $\overline{H}(Y)$. He proved asymptotic optimality of the

Lempel–Ziv algorithm for any HMP, among all uniquely decodable codes, in the sense that the normalized length of its codeword $(1/n)u(Y^n)$ has the lowest probability of exceeding a constant α, as $n \rightarrow \infty$, for any $\overline{H}(Y) < \alpha < \log|\mathcal{Y}|$. For $\alpha \leq \overline{H}(Y)$, the problem is not feasible and it is trivial for $\alpha \geq \log|\mathcal{Y}|$. This probability was shown to vanish exponentially fast for unifilar sources. The Lempel–Ziv code for HMPs was demonstrated to be asymptotically optimal in the competitive sense of Cover and Thomas [68, Sec. 5.11]. In particular, the Lempel–Ziv algorithm provides most of the time a codeword shorter than that of any other competing algorithm within a normalized redundancy term of $1/n$. This result was first proved for Rissanen's MDL universal code [268] for unifilar sources using the method of types, and then inferred for the Lempel–Ziv code for HMPs using Ziv's inequality (6.29). It should also be noted that the Lempel–Ziv algorithm compresses any observation sequence from any finite-alphabet HMP with essentially the same efficiency as any arithmetic coder which explicitly uses the pmf of the HMP. This observation follows from the Ziv inequality.

Goblirsch and Farvardin [140] studied in 1992 the design of switched scalar quantizers for a stationary composite source with known transition matrix and densities. The encoder comprises a set of scalar quantizers and a next-quantizer distribution. This distribution is indexed by the quantizers and codewords. Upon quantization of each observation, a quantizer is selected for the next observation by sampling from the next-quantizer distribution using a pseudorandom generator. The decoder has exact copies of the code books and the next-quantizer distribution and is fully synchronized with the encoder. Quantization of sources which are not necessarily HMPs, using finite-state quantizers with deterministic next-state functions, was studied by Dunham and Gray [94]. See also [135, Ch. 14].

2) Channel Coding: Drake [92] studied in 1965 decoding of a binary-symmetric Markov chain $\{S_t\}$ observed through a binary-symmetric memoryless channel. A decoder is called *singlet* if it estimates the source symbol S_t as the received symbol y_t regardless of past observations $\{y_\tau, \tau < t\}$. Drake provided necessary and sufficient conditions for the singlet decoder to be optimal in the minimum probability of symbol error sense. The work was extended by Devore [82] to decoding from a sampled observation sequence, data-independent decoding, nonsequential decoding, and decoding through channels with binomial distributed noise. A singlet sequence decoder estimates the source symbol sequence S^n as the received symbol sequence y^n. Phamdo and Farvardin [252] provided necessary and sufficient conditions for the singlet sequence decoder to be optimal in the minimum probability of sequence error sense when a binary symmetric Markov chain source is observed through a binary symmetric memoryless channel. Alajaji, Phamdo, Farvardin, and Fuja [5] extended the results from [252] to decoding of binary asymmetric Markov chains observed through binary Markov channels.

Bahl, Cocke, Jelinek, and Raviv [17] used in 1974 the forward–backward recursions (5.7) and (5.8) for estimating the states of an HMP in the minimum symbol error rate sense. The HMP was observed through a memoryless channel and thus resulted in another HMP with the same Markov chain. The same approach was used for decoding of convolutional and linear block codes in the minimum symbol error rate sense. The decoding algorithm is commonly referred to as the BCJR decoder, and stabilized recursions have been used for decoding of turbo codes [32], [33]. Turbo decoding of a finite-alphabet HMP with unknown parameter transmitted over a Gaussian memoryless channel was developed by Garcia-Frias and Villasenor [131].

Kaleh and Vallet [179] studied blind deconvolution of an i.i.d. data sequence transmitted across a finite memory channel with unknown transfer function. We shall demonstrate the approach for linear channels. Nonlinear channels are treated similarly. Let $\{X_1, X_2, \ldots\}$ denote the input i.i.d. sequence where the random variable X_t takes values in a finite-alphabet set $\mathcal{X}$. Let h denote the $l \times 1$ vector of the finite impulse response of the channel. Let $S_t = (X_t, \ldots, X_{t-l+1})'$. Let $\{W_1, W_2, \ldots\}$ denote a sequence of i.i.d. Gaussian random variables with zero mean and σ^2 variance representing the white noise in the channel. The observed signal at the channel's output at time t is $Y_t = h'S_t + W_t$. Since $\{S_t\}$ is a first-order Markov chain with state space $\mathcal{X}^l$, $\{Y_t\}$ is an HMP. The parameter $\phi = \{h, \sigma^2\}$ of the channel is unknown but assumed constant during n observations, say Y^n. The memory length of the channel is assumed known. The parameter ϕ is estimated from n observations y^n using the Baum algorithm and then used to decode these observations. Let $\phi_m = \{h_m, \sigma_m^2\}$ denote the estimate of ϕ at the end of the mth iteration. A new estimate h_{m+1} is obtained from the solution of the set of linear normal equations

$$\left\{\sum_{t=1}^{n} \sum_{\xi \in \mathcal{X}^l} P(S_t = \xi | y^n; \phi_m) \xi \xi' \right\} h_{m+1} = \sum_{t=1}^{n} \sum_{\xi \in \mathcal{X}^l} P(S_t = \xi | y^n; \phi_m) y_t \xi. \quad (14.21)$$

The noise variance re-estimation formula is

$$\sigma_{m+1}^2 = \frac{1}{n} \sum_{t=1}^{n} \sum_{\xi \in \mathcal{X}^l} P(S_t = \xi | y^n; \phi_m) |y_t - h_{m+1}' \xi|^2. \quad (14.22)$$

Given an estimate $\hat{\phi}$ of ϕ, the symbol x_t is decoded in the minimum symbol error rate using the decision rule

$$\hat{x}_t = \arg\max_{\alpha \in \mathcal{X}} \sum_{\xi \in \mathcal{X}^l:\, x_t = \alpha} P\left(S_t = \xi | y^n; \hat{\phi}\right). \quad (14.23)$$

A problem similar to blind deconvolution arises in decoding pulse amplitude modulation (PAM) signals using a receiver that is not synchronized with the transmitter. Kaleh [180] formulated this problem as a decoding problem of an HMP and applied the above approach for estimating the parameter and for decoding the signal. The parameter comprises the clock offset between the receiver and transmitter and the white noise variance. Cirpan and Tsatsanis [62] used an approach similar to that of Kaleh and Vallet [179] for semiblind channel deconvolution. The finite impulse response of the channel is estimated from the received data as well as from an embedded training data. The presence of training data improves the channel estimation accuracy at the expanse of lowering the bit rate.

Krishnamurthy, Dey, and LeBlanc [196] studied blind equalization of linear channels with infinite impulse response all-pole transfer functions. Phamdo and Farvardin [252] and Miller and Park [241] studied decoding of vector quantized sources observed through finite-alphabet memoryless noisy channels using a causal approximate MMSE estimator similar to (12.1). Brushe and White [48] and Brushe, Krishnamurthy, and White [49] studied demodulation of a number of convolutional coded signals impinging on an antenna array assuming unknown channel and direction of arrival. Krishnamurthy and Logothetis [199] studied estimation of code-division multiple-access (CDMA) signals in the presence of a narrowband interference signal and white additive noise. The CDMA signal was assumed a Markov chain with states representing quantized signal levels. Chao and Yao [57] proposed hidden Markov modeling of the burst error sequence in Viterbi decoding of convolutional codes. Turin [302] studied MAP decoding for HMP observed through an FSC.

D. Signal Processing

In this subsection, we describe some applications of HMPs in processing audio, biomedical, radar, sonar, and image signals.

1) Audio: Mixture processes were found useful in modeling speech signals in speaker identification applications [138], [209]. HMPs were used in modeling speech signals and noise sources in noisy speech enhancement applications [105]. HMPs were also used in environmental sound recognition whereas a recorded acoustic signal is classified as being produced by a subset of noise sources that are simultaneously active at a given time [67], [134]. The noise sources were assumed statistically independent HMPs. The observed signal is a mixture of these HMPs [67].

2) Biomedical: Characterization of currents flowing through a single ion channel in living cell membranes has attracted significant research effort. An overview of stochastic models and statistical analysis applied to ion channels, and an extensive list of references, can be found in [22]. This is a rich and challenging area of current research. Ion channel currents are believed to be well represented by a finite-state continuous-time Markov process where the states represent conductance levels. Recordings are made using the patch clamp technique where substantial nonwhite noise and deterministic interferences may be added. In addition, several conductance levels may be aggregated into a single state representing a function of the Markov process. The sampled signal constitutes a noisy function of a finite-state discrete-time Markov chain or an HMP. The theory of HMPs was applied to ion channels in [59], [60], [129], [128], [306], [307]. The parameter of the HMP is estimated in the ML sense using the Baum as well as other optimization algorithms. Parameter estimation in the presence of deterministic interferences was studied in [60], [194], [197]. The states representing the conductance levels are estimated using the Viterbi algorithm or a forward–backward recursion. Of particular importance are estimations of the channel kinetics and mean dwell time within each state. Characterization of multichannel patch clamp recordings using HMPs with appropriate parametrization of the transition matrix was studied in [7], [190].

DNA sequencing based on hidden Markov modeling was studied in [61], [227]. The states represented different regions or segments of the DNA. Segmentation was inferred from MAP estimates of state sequences as obtained from the Viterbi algorithm or the forward–backward recursions. In another application [200], HMPs were applied to statistical modeling of protein families for database searching and multiple sequence alignment. In [53], neuron firing patterns were characterized by the most likely state sequence of an appropriately trained HMP. In [264], classification of neuronal responses to visual stimuli based on hidden Markov modeling was studied.

HMPs were also used in automated analysis and classification of ECG signals [63], [298], [193]. ECG wave patterns were associated with states and detected from the most likely state sequence of appropriately trained HMPs. In another application, HMPs were used to model epileptic seizure counts with varying Poisson rates [6], [207].

3) Spectral Estimation: A sinusoidal signal with a time-varying frequency observed in white noise comprises an HMP when the unknown frequency is assumed a Markov process. Algorithms for tracking quantized versions of the frequency using the Viterbi algorithm were developed in [296], [320], [313], [321], [322].

4) Radar and Sonar: A problem related to frequency tracking is that of maneuvering source tracking in sonar and radar systems [208]. The relative location and velocity of the source with respect to an observer comprised the state vector in a dynamical system. A quantized version of the state variables were tracked using the Viterbi algorithm. Due to the observer's motion, optimal control was designed using the theory of partially observed Markov decision processes. In [201], [12], ML target localization using over-the-horizon radar systems was studied. The uncertainties in the ionospheric propagation conditions were modeled as an HMP. The states represented ray mode types. The parameter of the HMP was estimated using smoothed bootstrap Monte Carlo resampling [96].

5) Image: Restoration from corrupted images modeled as hidden Markov random fields was studied by Besag [34]. The image was represented by a Markov field and its pixels were alternatively estimated in the ML sense. Classification of images represented by hidden Markov random fields or by one-dimensional HMPs was studied in [157], [257], [317], [217]. Partially hidden Markov processes were studied in [122] and applied to image compression.

E. Other Applications

In this subsection, we briefly review applications of HMPs in the area of fault detection, economics, and metrology.

1) Fault Detection: Fast failure detection and prediction in communication networks was studied in [16]. An HMP with two states representing good and bad conditions of the network, and a binary alphabet representing good and bad checksums in each state was assumed for the fault detection process. ML estimation of the network's condition (state) was performed using the Viterbi algorithm. In [293], [294], an HMP-based real-time fault detection system for NASA's deep space network antennas is described. Here multifaults are monitored by estimating their conditional probabilities at any given time using the forward recursion or the Viterbi algorithm. In [2], an HMP was constructed for inventory system with perishable items.

2) Economics: HMPs and switching autoregressive processes appear particularly suitable to model macroeconomic or financial time series over sufficiently long periods [156]. The regime of the process provides a convenient way to reflect on events that may affect the underlying statistics of the time series such as wars, changes in government policies, etc. A summary of many properties of HMPs and their application in economics is given by Hamilton [156, Ch. 22]. In [282], HMPs with Gaussian pdfs were used to model subseries of the S&P 500 return series as registered from 1928 to 1991. Explicit expressions for the second-order statistics of these HMPs were also given. Expressions for second-order statistics of HMPs with discrete observations such as HMPs with Poisson and binomial pmfs were derived in [228].

3) Metrology: HMPs were used in [331], [284] to model rainfall records assuming some "climate states" which were modeled as a Markov chain.

XV. Concluding Remarks

An overview of HMPs was presented in this paper. Clearly, the theory of HMPs is very rich with many results derived from statistics, probability theory, information theory, control theory, and optimization theory. While HMPs are fairly general processes, they are still amenable to mathematical analysis. Many of these results were developed only in the past few years. Many ingenious approaches have been invented to study and prove large-sample properties of HMPs. We have attempted to present the principles of the main theoretical results and to point out to differences in alternative proofs. The emphasis of the paper is on the new results even though some more classical material was included for completeness and proper perspective.

We have collected a large number of results primarily from the mathematical literature and described a range of selected applications. We have seen how results developed in one area are useful in another area. For example, the source-channel information-theoretic model for an HMP enables quick inference of their statistical properties using existing results, which otherwise are harder to prove directly. The forward–backward recursions are useful in decoding of turbo codes in data communications. Ergodic theorems for relative entropy densities of HMPs have significance in coding, estimation, and hypothesis testing of HMPs. The Ziv inequality which proved useful in order estimation and hypothesis testing can also be used in assessing the quality of a local ML estimator for finite-alphabet HMPs. The forward–backward recursions for HMPs become the Kalman filter and smoother under appropriate conditions. Otherwise, they provide optimal filters and smoothers for non-Gaussian nonlinear discrete-time signals.

Some aspects of HMPs were inevitably left out. Our primary focus was on discrete-time general HMPs. Some results concerning HMPs with separable compact state spaces were included. We did not cover continuous-time HMPs, nor did we treat hidden Markov fields which play an important role in image processing. Some references to these areas were provided in this paper. In addition, dynamical system approaches to these two areas can be found in [99].

HMPs have attracted significant research effort in recent years which has resulted in substantial gain in understanding their statistical properties and in designing asymptotically optimal algorithms for parameter estimation and for universal coding and classification. The intuitive appeal of HMPs in many applications combined with their solid theory and the availability of fast digital signal processors are expected to attract further significant research effort in years to come.

Acknowledgment

The authors are grateful to Amir Dembo, Robert J. Elliott, Robert M. Gray, Amos Lapidoth, Tobias Rydén, William Turin, Tsachy Weissman, and the anonymous referees for helpful comments and suggestions. They also thank Randal Douc, Catherine Matias, Éric Moulines, and Tobias Rydén for making available preprints of their most recent work. The authors thank Shlomo Shamai for coordinating the review process.

References

[1] R. L. Adler, "Ergodic and mixing properties of infinite memory channels," *Proc. Amer. Math. Soc.*, vol. 12, no. 6, pp. 924–930, 1961.

[2] L. Aggoun, L. Benkherouf, and L. Tadj, "A hidden Markov model for an inventory system with perishable items," *J. Appl. Math. Stochastic Anal.*, vol. 10, no. 4, pp. 423–430, 1997.

[3] S. M. Aji and R. J. McEliece, "The generalized distributive law," *IEEE Trans. Inform. Theory*, vol. 46, pp. 325–343, Mar. 2000.

[4] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.

[5] F. Alajaji, N. Phamdo, N. Farvardin, and T. E. Fuja, "Detection of binary Markov sources over channels with additive Markov noise," *IEEE Trans. Inform. Theory*, vol. 42, pp. 230–239, Jan. 1996.

[6] P. S. Albert, "A two-state Markov mixture model for a time series of epileptic seizure counts," *Biometrics*, vol. 47, pp. 1371–1381, Dec. 1991.

[7] A. Albertsen and U.-P. Hansen, "Estimation of kinetic rate constants from multi-channel recordings by a direct fit of the time series," *Biophys. J.*, vol. 67, pp. 1393–1403, Oct. 1994.

[8] S. Amari, "A theory of adaptive pattern classifiers," *IEEE Trans. Electron. Comput.*, vol. EC-16, pp. 299–307, June 1967.

[9] J. B. Anderson and J. B. Bodie, "Tree encoding of speech," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 379–387, July 1975.

[10] J. B. Anderson and C.-W. Law, "Real-number convolutional codes for speech-like quasistationary sources," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 778–782, Nov. 1977.

[11] B. D. O. Anderson, "From Wiener to hidden Markov models," *IEEE Contr. Syst. Mag.*, vol. 19, pp. 41–51, June 1999.

[12] R. H. Anderson and J. L. Krolik, "Over-the-horizon radar target localization using a hidden Markov model estimated from ionosonde data," *Radio Sci.*, vol. 33, no. 4, pp. 1199–1213, July–Aug. 1998.

[13] C. Andrieu and A. Doucet, "Simulated annealing for maximum a posteriori parameter estimation of hidden Markov models," *IEEE Trans. Inform. Theory*, vol. 46, pp. 994–1004, May 2000.

[14] R. B. Ash, *Information Theory*. New York: Dover, 1965.

[15] M. Askar and H. Derin, "A recursive algorithm for the Bayes solution of the smoothing problem," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 558–561, Apr. 1981.

[16] E. Ayanoglu, "Robust and fast failure detection and prediction for fault-tolerant communication network," *Electron. Lett.*, vol. 28, no. 10, pp. 940–941, May 1992.

[17] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, “Optimal decoding of linear codes for minimizing symbol error rate,” *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 284–287, Mar. 1974.

[18] L. R. Bahl and F. Jelinek, “Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition,” *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 404–411, July 1975.

[19] L. R. Bahl, F. Jelinek, and R. L. Mercer, “A maximum likelihood approach to continuous speech recognition,” *IEEE Trans. Pattern Anal. Machine Intel.*, vol. PAMI-5, pp. 179–190, Mar. 1983.

[20] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, “Maximum mutual information estimation of hidden Markov model parameters for speech recognition,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Apr. 1986, pp. 49–52.

[21] J. K. Baker, “The DRAGON system—An overview,” *IEEE Trans. Acoust., Speech Signal Processing*, vol. ASSP-23, pp. 24–29, Feb. 1975.

[22] F. G. Ball and J. A. Rice, “Stochastic models for ion channels: Introduction and bibliography,” *Math. Biosci.*, vol. 112, pp. 189–206, 1992.

[23] A. R. Barron, “The strong ergodic theorem for densities: Generalized Shannon–McMillan-Breiman theorem,” *Ann. Probab.*, vol. 13, no. 4, pp. 1292–1303, 1985.

[24] A. Barron, J. Rissanen, and B. Yu, “The minimum description length principle in coding and modeling,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 2743–2760, Oct. 1998.

[25] L. E. Baum and T. Petrie, “Statistical inference for probabilistic functions of finite state Markov chains,” *Ann. Math. Statist.*, vol. 37, pp. 1554–1563, 1966.

[26] L. E. Baum and J. A. Eagon, “An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology,” *Bull. Amer. Math. Soc.*, vol. 73, pp. 360–363, 1967.

[27] L. E. Baum and G. R. Sell, “Growth transformations for functions on manifolds,” *Pacific J. Math.*, vol. 27, no. 2, pp. 211–227, 1968.

[28] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, “A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains,” *Ann. Math. Statist.*, vol. 41, pp. 164–171, 1970.

[29] L. E. Baum, “An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes,” in *Inequalities, III (Proc. 3rd Symp., Univ. Calif., Los Angeles, Calif., 1969; dedicated to the memory of Theodore S. Motzkin)*. New York: Academic, 1972, pp. 1–8.

[30] R. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press, 1957.

[31] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[32] C. Berrou, A. Glavieux, and P. Thitimajshima, “Near Shannon limit error-correcting coding and decoding: Turbo-codes,” in *Proc. ICC’93*, May 1993, pp. 1064–1070.

[33] C. Berrou and A. Glavieux, “Near optimum error correcting coding and decoding: Turbo-codes,” *IEEE Trans. Commun.*, vol. 44, pp. 1261–1271, Oct. 1996.

[34] J. Besag, “On the statistical analysis of dirty pictures,” *J. Roy. Statist. Soc. B*, vol. 48, no. 3, pp. 259–302, 1986.

[35] P. J. Bickel and Y. Ritov, “Inference in hidden Markov models I: Local asymptotic normality in the stationary case,” *Bernoulli*, vol. 2, no. 3, pp. 199–228, 1996.

[36] P. J. Bickel, Y. Ritov, and T. Rydén, “Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models,” *Ann. Statist.*, vol. 26, no. 4, pp. 1614–1635, 1998.

[37] P. Billingsley, *Convergence of Probability Measures*. New York: Wiley, 1968.

[38] ——, *Probability and Measure*. New York: Wiley, 1995.

[39] M. Billio, A. Monfort, and C. P. Robert, “Bayesian estimation of switching ARMA models,” *J. Econometrics*, vol. 93, pp. 229–255, 1999.

[40] J. J. Birch, “Approximations for the entropy for functions of Markov chains,” *Ann. Math. Statist.*, vol. 33, pp. 930–938, 1962.

[41] D. Blackwell, “The entropy of functions of finite-state Markov chains,” in *Trans. 1st Prague Conf. Information Theory, Statistical Decision Functions, Random Processes*. Prague, Czechoslovakia: Pub. House Czechoslovak Acad. Sci., 1957, pp. 13–20.

[42] D. Blackwell and L. Koopmans, “On the identifiability problem for functions of finite Markov chains,” *Ann. Math. Statist.*, vol. 28, pp. 1011–1015, 1957.

[43] D. Blackwell, L. Breiman, and A. J. Thomasian, “Proof of Shannon’s transmission theorem for finite-state indecomposable channels,” *Ann. Math. Stat.*, vol. 29, pp. 1209–1220, 1958.

[44] I. A. Boguslavskii and M. Y. Borodovskii, “On identification of states of a sequence generated by a hidden Markov model,” *J. Comput. Syst. Sci. Int.*, vol. 37, no. 4, pp. 551–556, 1998.

[45] P. Bougerol and N. Picard, “Strict stationarity of generalized autoregressive processes,” *Ann. Probab.*, vol. 20, no. 4, pp. 1714–1730, 1992.

[46] A. Brandt, “The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients,” *Adv. Appl. Probab.*, vol. 18, pp. 211–220, 1986.

[47] D. R. Brillinger, *Time Series-Data Analysis and Theory*. New York: Holt, Rinehart &Winston, 1975.

[48] G. D. Brushe and L. B. White, “Spatial filtering of superimposed convolutional coded signals,” *IEEE Trans. Commun.*, vol. 45, pp. 1144–1153, Sept. 1997.

[49] G. D. Brushe, V. Krishnamurthy, and L. B. White, “A reduced-complexity online state sequence and parameter estimator for superimposed convolutional coded signals,” *IEEE Trans. Commun.*, vol. 45, pp. 1565–1574, Dec. 1997.

[50] G. D. Brushe, R. E. Mahony, and J. B. Moore, “A soft output hybrid algorithm for ML/MAP sequence estimation,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 3129–3134, Nov. 1998.

[51] P. Bryant and J. A. Williamson, “Asymptotic behavior of classification maximum likelihood estimates,” *Biometrika*, vol. 65, no. 2, pp. 273–281, 1978.

[52] C. J. Burke and M. Rosenblatt, “A Markovian function of a Markov chain,” *Ann. Math. Statist.*, vol. 29, pp. 1112–1122, 1958.

[53] A.-C. Camproux, F. Saunier, G. Chouvet, J.-C. Thalabard, and G. Thomas, “A hidden Markov model approach to neuron firing patters,” *Biophys. J.*, vol. 71, pp. 2404–2412, Nov. 1996.

[54] J. W. Carlyle, “Identification of state-calculable functions of finite Markov chains,” *Ann. Math. Statist.*, vol. 38, pp. 201–205, 1967.

[55] R. L. Cave and L. P. Neuwirth, “Hidden Markov models for English,” in *Proc. Symp. Application of Hidden Markov Models to Text and Speech*, J. D. Ferguson, Ed. Princeton, NJ: IDA-CRD, 1980, pp. 16–56.

[56] R. W. Chang and J. C. Hancock, “On receiver structures for channels having memory,” *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 463–468, Oct. 1966.

[57] C.-C. Chao and Y.-L. Yao, “Hidden Markov models for the burst error statistics of Viterbi decoding,” *IEEE Trans. Commun.*, vol. 44, pp. 1620–1622, Dec. 1996.

[58] W. Chou, B.-H. Juang, and C. H. Lee, “Segmental GPD training of HMM based speech recognizer,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1992, pp. I-473–I-476.

[59] S. H. Chung, J. B. Moore, L. Xia, L. S. Premkumar, and P. W. Gage, “Characterization of single channel currents using digital signal processing techniques based on hidden Markov models,” *Phil. Trans. Roy. Soc. London B*, vol. 329, pp. 265–285, 1990.

[60] S. H. Chung, V. Krishnamurthy, and J. B. Moore, “Adaptive processing techniques based on hidden Markov models for characterizing very small channel currents buried in noise and deterministic interferences,” *Phil. Trans. Roy. Soc. London B*, vol. 334, pp. 357–384, 1991.

[61] G. A. Churchill, “Stochastic models for Heterogeneous DNA sequences,” *Bull. Math. Biology*, vol. 51, no. 1, pp. 79–94, 1989.

[62] H. A. Cirpan and M. K. Tsatsanis, “Stochastic maximum likelihood methods for semi-blind channel estimation,” *IEEE Signal Processing Lett.*, vol. 5, pp. 21–24, Jan. 1998.

[63] D. A. Coast, G. G. Cano, and S. A. Briller, “Use of hidden Markov models for electrocardiographic signal analysis,” *J. Electrocardiol.*, vol. 23 Suppl., pp. 184–191, 1990.

[64] I. B. Collings, V. Krishnamurthy, and J. B. Moore, “On-line identification of hidden Markov models via recursive prediction error techniques,” *IEEE Trans. Signal Processing*, vol. 42, pp. 3535–3539, Dec. 1994.

[65] R. Comerford, J. Makhoul, and R. Schwartz, “The voice of the computer is heard in the land and it listens too!,” *IEEE Spectrum*, vol. 34, pp. 39–43, Dec. 1997.

[66] C. Couvreur, “Hidden Markov models and their mixtures,” Dept. Math., Université Catholique de Louvain, Louvain, Belgium, 1996.

[67] ——, “Enviromental sound recognition: A statistical approach,” D.Sc. dissertation, Faculté Polytechnique De Mons, June 1997.

[68] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[69] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, “Wavelet-based statistical signal processing using hidden Markov models,” *IEEE Trans. Signal Processing*, vol. 46, pp. 886–902, Apr. 1998.

[70] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[71] I. Csiszár and G. Tusnady, “Information geometry and alternating maximization procedures,” *Statist. Decisions Suppl.*, vol. 1, pp. 205–237, 1984.

[72] I. Csiszár, “Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems,” *Ann. Statist.*, vol. 19, no. 4, pp. 2032–2066, 1991.

[73] ——, "The method of types," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2505–2523, Oct. 1998.

[74] I. Csiszár and P. C. Shields, "The consistency of the BIC Markov order estimator," *Ann. Statist.*, vol. 28, no. 6, pp. 1601–1619, 2000.

[75] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.

[76] B. Delyon, "Remarks on linear and nonlinear filtering," *IEEE Trans. Inform. Theory*, vol. 41, pp. 317–322, Jan. 1995.

[77] A. Dembo and O. Zeitouni, "Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm," *Stochastic Processes Their Applic.*, vol. 23, no. 1, pp. 91–113, 1986.

[78] ——, "On the parameters estimation of continuous-time ARMA processes from noisy observations," *IEEE Trans. Automat. Contr.*, vol. AC-32, pp. 361–364, Apr. 1987.

[79] ——, "Maximum *a posteriori* estimation of time-varying ARMA processes from noisy observations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 471–476, Apr. 1988.

[80] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[81] P. A. Devijver, "Baum's forward–backward algorithm revisited," *Pattern Recogn. Lett.*, vol. 3, pp. 369–373, 1985.

[82] J. L. Devore, "A note on the observation of a Markov source through a noisy channel," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 762–764, Nov. 1974.

[83] S. W. Dharmadhikari, "Functions of finite Markov chains," *Ann. Math. Statist.*, vol. 34, pp. 1022–1032, 1963.

[84] ——, "Sufficient conditions for a stationary process to be a function of a finite Markov chain," *Ann. Math. Statist.*, vol. 34, pp. 1033–1041, 1963.

[85] ——, "Exchangeable processes which are functions of stationary Markov chain," *Ann. Math. Statist.*, vol. 35, pp. 429–430, 1964.

[86] ——, "A characterization of a class of functions of finite Markov chains," *Ann. Math. Statist.*, vol. 36, pp. 524–528, 1965.

[87] J. Diebolt and C. P. Robert, "Estimation of finite mixture distributions through Bayesian sampling," *J. Roy. Statist. Soc. B*, vol. 56, no. 2, pp. 363–373, 1994.

[88] J. Diebolt and E. H. S. Ip, "Stochastic EM: Method and application," in *Markov Chain Monte Carlo In Practice*, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds. London, U.K.: Chapman & Hall, 1996, pp. 259–273.

[89] S. N. Diggavi and M. Grossglauser, "Information transmission over a finite buffer channel," in *Proc. IEEE Int. Symp. Information Theory (ISIT 2000)*, Sorrento, Italy, June 2000, p. 52.

[90] R. Douc and C. Matias, "Asymptotics of the maximum likelihood estimator for general hidden Markov models," *Bernoulli*, vol. 7, no. 3, pp. 381–420, 2001.

[91] R. Douc, É. Moulines, and T. Rydén, "Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime," *Ann. Statist.*, submitted for publication.

[92] A. W. Drake, "Observation of a Markov source through a noisy channel," in *Proc. IEEE Symp. Signal Transmission and Processing*, Columbia Univ., New York, 1965, pp. 12–18.

[93] H. Drucker, "Speech processing in a high ambient noise environment," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 165–168, June 1968.

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

[95] A. P. Dunmur and D. M. Titterington, "The influence of initial conditions on maximum likelihood estimation of the parameters of a binary hidden Markov model," *Statist. Probab. Lett.*, vol. 40, no. 1, pp. 67–73, 1998.

[96] B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia, PA: SIAM, 1982.

[97] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell Syst. Tech. J.*, vol. 42, pp. 1977–1997, Sept. 1963.

[98] R. J. Elliott, "New finite-dimensional filters and smoothers for noisily observed Markov chains," *IEEE Trans. Inform. Theory*, vol. 39, pp. 265–271, Jan. 1993.

[99] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*. New York: Springer-Verlag, 1994.

[100] R. J. Elliott and V. Krishnamurthy, "New finite-dimensional filters for parameter estimation of discrete-time linear Gaussian models," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 938–951, May 1999.

[101] Y. Ephraim, A. Dembo, and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 1001–1013, Sept. 1989.

[102] Y. Ephraim and L. R. Rabiner, "On the relations between modeling approaches for speech recognition," *IEEE Trans. Inform. Theory*, vol. IT-36, pp. 372–380, Mar. 1990.

[103] Y. Ephraim, "Speech enhancement using state dependent dynamical system model," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Mar. 1992, pp. 289–292.

[104] Y. Ephraim and N. Merhav, "Lower and upper bounds on the minimum mean square error in composite source signal estimation," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1709–1724, Nov. 1992.

[105] Y. Ephraim, "Statistical model-based speech enhancement systems," *Proc. IEEE*, vol. 80, pp. 1526–1555, Oct. 1992.

[106] Y. Ephraim and M. Rahim, "On second-order statistics and linear estimation of cepstral coefficients," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 162–176, Mar. 1999.

[107] R. V. Erickson, "Functions of Markov chains," *Ann. Math. Statist.*, vol. 41, no. 3, pp. 843–850, 1970.

[108] Y. Erlich, "On HMM based speech recognition using MCE approach," M.Sc. thesis, Dept. Elec. Eng., Technion-Israel Inst. Technol., Haifa, Israel, 1996.

[109] B. S. Everitt and D. J. Hand, *Finite Mixture Distributions*. London-New York: Chapman and Hall, 1981.

[110] V. Fabian, "On asymptotically efficient recursive estimation," *Ann. Statist.*, vol. 4, pp. 854–866, 1978.

[111] A. Faragó and G. Lugosi, "An algorithm to find the global optimum of left-to-right hidden Markov model parameters," *Probl. Contr. Inform. Theory*, vol. 18, no. 6, pp. 435–444, 1989.

[112] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1726–1745, Sept. 1998.

[113] M. Feder and N. Merhav, "Universal composite hypothesis testing: A competitive minimax approach," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1504–1517, June 2002.

[114] P. D. Feigin and R. L. Tweedie, "Random coefficient autoregressive processes: A Markov chain analysis of stationarity and finiteness of moments," *J. Time Ser. Anal.*, vol. 6, no. 1, pp. 1–14, 1985.

[115] J. D. Ferguson, Ed., *Proc. Symp. Application of Hidden Markov Models to Text and Speech*. Princeton, NJ: IDA-CRD, 1980.

[116] L. Finesso, "Consistent estimation of the order for Markov and hidden Markov chains," Ph.D. dissertation, Univ. Maryland, College Park, 1990.

[117] W. Fischer and K. Meier-Hellstern, "The Markov-modulated Poisson process (MMPP) cookbook," *Perf. Eval.*, vol. 18, pp. 149–171, 1992.

[118] R. J. Fontana, "Universal codes for a class of composite sources," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 480–482, July 1980.

[119] ——, "Limit theorems for slowly varying composite sources," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 702–709, Nov. 1980.

[120] R. J. Fontana, R. M. Gray, and J. C. Kieffer, "Asymptotically mean stationary channels," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 308–316, May 1981.

[121] R. J. Fontana, "On universal coding for classes of composite and remote sources with memory," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 784–786, Nov. 1981.

[122] S. Forchhammer and J. Rissanen, "Partially hidden Markov models," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1253–1256, July 1996.

[123] L. A. Foreman, "Generalization of the Viterbi algorithm," *IMA J. Math. Applied in Business & Industry*, vol. 4, pp. 351–367, 1993.

[124] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE*, vol. 61, pp. 268–278, Mar. 1973.

[125] M. Fox, "Conditions under which a given process is a function of a Markov chain (Abstract)," *Ann. Math. Statist.*, vol. 33, p. 1206, 1962.

[126] C. Francq and M. Roussignol, "On white noises driven by hidden Markov chains," *J. Time Ser. Anal.*, vol. 18, no. 6, pp. 553–578, 1997.

[127] ——, "Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum-likelihood estimator," *Statistics*, vol. 32, pp. 151–173, 1998.

[128] D. R. Fredkin and J. A. Rice, "Maximum likelihood estimation and identification directly from single-channel recordings," *Proc. Roy. Soc. London B*, vol. 249, pp. 125–132, 1992.

[129] ——, "Bayesian restoration of single channel patch clamp recordings," *Biometrics*, vol. 48, pp. 427–448, 1992.

[130] D. S. Freed and L. A. Shepp, "A Poisson process whose rate is a hidden Markov process," *Adv. Appl. Probab.*, vol. 14, no. 1, pp. 21–36, 1982.

[131] J. Garcia-Frias and J. Villasenor, "Turbo decoding of hidden Markov sources with unknown parameters," in *Proc. IEEE Data Compression Conf.*, Utah, Mar. 1998, pp. 159–168.

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

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

[134] P. Gaunard, C. G. Mubikangiey, C. Couvreur, and V. Fontaine, "Automatic classification of environmental noise events by hidden Markov models," in *IEEE Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 1998, pp. 3609–3612.

[135] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer, 1991.

[136] E. J. Gilbert, "On the identifiability problem for functions of finite Markov chains," *Ann. Math. Statist.*, vol. 30, pp. 688–697, 1959.

[137] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Syst. Tech. J.*, vol. 39, pp. 1253–1265, Sept. 1960.

[138] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Mag.*, vol. 11, pp. 18–32, Oct. 1994.

[139] P. Giudici, T. Rydén, and P. Vandekerkhove, "Likelihood-ratio tests for hidden Markov models," *Biometrics*, vol. 56, pp. 742–747, Sept. 2000.

[140] D. M. Goblirsch and N. Farvardin, "Switched scalar quantizers for hidden Markov sources," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1455–1473, Sept. 1992.

[141] A. J. Goldsmith and P. P. Varaiya, "Capacity, mutual information, and coding for finite-state Markov channels," *IEEE Trans. Inform. Theory*, vol. 42, pp. 868–886, May 1996.

[142] G. Golubev and R. Khasminskii, "Asymptotically optimal filtering for a hidden Markov model," *Math. Methods Statist.*, vol. 7, no. 2, pp. 192–209, 1998.

[143] G. K. Golubev, "On filtering for a hidden Markov chain under square performance criterion," *Probl. Inform. Transm.*, vol. 36, no. 3, pp. 213–219, 2000.

[144] P. S. Gopalakrishnan, D. Kanevsky, A. Nádas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Trans. Inform. Theory*, vol. 37, pp. 107–113, Jan. 1991.

[145] J. Goutsias and J. M. Mendel, "Optimal simultaneous detection and estimation of filtered discrete semi-Markov chains," *IEEE Trans. Inform. Theory*, vol. 34, pp. 551–568, May 1988.

[146] R. M. Gray, "Rate distortion functions for finite-state finite-alphabet Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 127–134, Mar. 1971.

[147] R. M. Gray and L. D. Davisson, "The ergodic decomposition of stationary discrete random processes," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 625–636, Sept. 1974.

[148] R. M. Gray and J. C. Kieffer, "Asymptotically mean stationary measures," *Ann. Probab.*, vol. 8, no. 5, pp. 962–973, 1980.

[149] R. M. Gray, A. H. Gray, Jr., G. Rebolledo, and J. E. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 708–721, Nov. 1981.

[150] R. M. Gray, M. O. Dunham, and R. L. Gobbi, "Ergodicity of Markov channels," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 656–664, Sept. 1987.

[151] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*. New York: Springer-Verlag, 1988.

[152] ——, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.

[153] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2325–2383, Oct. 1998.

[154] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*. Oxford, U.K.: Oxford Univ. Press, 2001.

[155] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inform. Theory*, vol. 35, pp. 401–408, Mar. 1989.

[156] J. D. Hamilton, *Time Series Analysis*. Princeton, NJ: Princeton Univ. Press, 1994.

[157] Y. He and A. Kundu, "2-D shape classification using hidden Markov models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 1172–1184, Nov. 1991.

[158] H. Heffes, "A class of data traffic processes—covariance function characterization and related queuing results," *Bell Syst. Tech. J.*, vol. 59, no. 6, pp. 897–929, July–Aug. 1980.

[159] H. Heffes and D. M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. Select. Areas Commun.*, vol. 4, pp. 856–868, Sept. 1986.

[160] A. Heller, "On stochastic processes derived from Markov chains," *Ann. Math. Statist.*, vol. 36, pp. 1286–1291, 1965.

[161] B. M. Hochwald and P. R. Jelenković, "State learning and mixing in entropy of hidden Markov processes and the Gilbert–Elliott channel," *IEEE Trans. Inform. Theory*, vol. 45, pp. 128–138, Jan. 1999.

[162] P. G. Hoel, S. C. Port, and C. J. Stone, *Introduction to Stochastic Processes*. Boston, MA: Houghton Mifflin, 1972.

[163] U. Holst and G. Lindgren, "Recursive estimation in mixture models with Markov regime," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1683–1690, Nov. 1991.

[164] U. Holst, G. Lindgren, J. Holst, and M. Thuvesholmen, "Recursive estimation in switching autoregressions with a Markov regime," *J. Time Ser. Anal.*, vol. 15, no. 5, pp. 489–506, 1994.

[165] X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov Models for Speech Recognition*. Edinburgh, Scotland: Edinburgh Univ. Press, 1990.

[166] J. P. Hughes, "Computing the observed information in the hidden Markov model using the EM algorithm," *Statist. Probab. Lett.*, vol. 32, pp. 107–114, 1997.

[167] H. Ito, S.-I. Amari, and K. Kobayashi, "Identifiability of Hidden Markov information sources and their minimum degrees of freedom," *IEEE Trans. Inform. Theory*, vol. 38, pp. 324–333, Mar. 1992.

[168] M. R. James, V. Krishnamurthy, and F. Le Gland, "Time discretization of continuous-time filters and smoothers for HMM parameter estimation," *IEEE Trans. Inform. Theory*, vol. 42, pp. 593–605, Mar. 1996.

[169] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York: Academic, 1970.

[170] F. Jelinek, "A fast sequential decoding algorithm using a stack," *IBM J. Res. Develop.*, vol. 13, pp. 675–685, Nov. 1969.

[171] F. Jelinek, L. R. Bahl, and R. L. Mercer, "Design of a linguistic statistical decoder for recognition of continuous speech," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 3, pp. 250–256, May 1975.

[172] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532–556, Apr. 1976.

[173] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press, 1998.

[174] J. L. Jensen and N. V. Petersen, "Asymptotic normality of the maximum likelihood estimator in state space models," *Ann. Statist.*, vol. 27, no. 2, pp. 514–535, 1999.

[175] B.-H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 1404–1413, Dec. 1985.

[176] ——, "The segmental k-means algorithm for estimating parameters of hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1639–1641, Sept. 1990.

[177] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol. 40, pp. 3043–3054, Dec. 1992.

[178] B.-H. Juang, W. Chou, and C.-H. Lee, "Statistical and discriminative methods for speech recognition," in *Automatic Speech and Speaker Recognition*, C.-H. Lee, F. K. Soong, and K. K. Paliwal, Eds. Boston, MA: Kluwer , 1996, pp. 109–132.

[179] G. K. Kaleh and R. Vallet, "Joint parameter estimation and symbol detection for linear or nonlinear unknown channels," *IEEE Trans. Commun.*, vol. 42, pp. 2406–2413, July 1994.

[180] G. K. Kaleh, "The Baum–Welch algorithm for the detection of time-unsynchronized rectangular PAM signals," *IEEE Trans. Commun.*, vol. 42, pp. 260–262, Feb./Mar./Apr. 1994.

[181] H. A. Karlsen, "Existence of moments in a stationary stochastic difference equation," *Adv. Appl. Probab.*, vol. 22, pp. 129–146, 1990.

[182] R. L. Kashyap, "Identification of a transition matrix of a Markov chain from noisy measurements of state," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 161–166, Mar. 1970.

[183] A. Kehagias, "Bayesian classification of hidden Markov models," *Mathl. Comput. Modelling*, vol. 23, no. 5, pp. 25–43, 1996.

[184] R. Khasminskii, B. Lazareva, and J. Stapleton, "Some procedures for state estimation of a hidden Markov chain with two states," in *Statistical Decision Theory and Related Topics*. New York: Springer, 1994, vol. V (West Lafayette, IN, 1992), pp. 477–487.

[185] R. Khasminskii and O. Zeitouni, "Asymptotic filtering for finite state Markov chains," *Stochastic Processes Their Applic.*, vol. 63, pp. 1–10, 1996.

[186] J. C. Kieffer and M. Rahe, "Markov channels are asymptotically mean stationary," *SIAM J. Math. Anal.*, vol. 12, no. 3, pp. 293–305, May 1981.

[187] J. C. Kieffer, "Strongly consistent code-based identification and order estimation for constrained finite-state model classes," *IEEE Trans. Inform. Theory*, vol. 39, pp. 893–902, May 1993.

[188] J. F. C. Kingman, "Subadditive ergodic theory," *Ann. Probab.*, vol. 1, no. 6, pp. 883–909, 1973.

[189] G. Kitagawa, “Non-Gaussian state-space modeling of nonstationary time series,” *J. Amer. Statist. Assoc.*, vol. 82, no. 400, pp. 1032–1041, Dec. 1987.

[190] S. Klein, J. Timmer, and J. Honerkamp, “Analysis of multichannel patch clamp recordings by hidden Markov models,” *Biometrics*, vol. 53, pp. 870–884, Sept. 1997.

[191] J. A. Kogan, “Hidden Markov models estimation via the most informative stopping times for the Viterbi algorithm,” in *Image Models (and Their Speech Model Cousins)*. New York: Springer, 1996. Minneapolis, MN, 1993/1994, IMA Vol. Math. Appl., 80.

[192] R. Kohn and C. F. Ansley, “Comments on Kitagawa [189],” *J. Amer. Statist. Assoc.*, vol. 82, no. 400, pp. 1041–1044, Dec. 1987.

[193] A. Koski, “Modeling ECG signals with hidden Markov models,” *Artificial Intell. in Medicine*, vol. 8, pp. 453–471, 1996.

[194] V. Krishnamurthy, J. B. Moore, and S.-H. Chung, “Hidden Markov model signal processing in presence of unknown deterministic interferences,” *IEEE Trans. Automat. Contr.*, vol. 38, pp. 146–152, Jan. 1993.

[195] V. Krishnamurthy and J. B. Moore, “On-line estimation of hidden Markov model parameters based on the Kullback–Leibler information measure,” *IEEE Trans. Signal Processing*, vol. 41, pp. 2557–2573, Aug. 1993.

[196] V. Krishnamurthy, S. Dey, and J. P. LeBlanc, “Blind equalization of IIR channels using hidden Markov models and extended least squares,” *IEEE Trans. Signal Processing*, vol. 43, pp. 2994–3006, Dec. 1995.

[197] V. Krishnamurthy and R. J. Elliott, “A filtered EM algorithm for joint hidden Markov model and sinusoidal parameter estimation,” *IEEE Trans. Signal Processing*, vol. 43, pp. 353–358, Jan. 1995.

[198] V. Krishnamurthy and T. Rydén, “Consistent estimation of linear and nonlinear autoregressive models with Markov regime,” *J. Time Ser. Anal*, vol. 19, no. 3, pp. 291–307, 1998.

[199] V. Krishnamurthy and A. Logothetis, “Adaptive nonlinear filters for narrow-band interference suppression in spread-spectrum CDMA systems,” *IEEE Trans. Commun.*, vol. 47, pp. 742–753, May 1999.

[200] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler, “Hidden Markov models in computational biology applications to protein modeling,” *J. Molec. Biol.*, vol. 235, pp. 1501–1531, 1994.

[201] J. L. Krolik and R. H. Anderson, “Maximum likelihood coordinate registration for over-the-horizon radar,” *IEEE Trans. Signal Processing*, vol. 45, pp. 945–959, Apr. 1997.

[202] H.-M. Krolzig, *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Applications to Business Cycle Analysis (Lecture Notes in Economics and Mathematical Systems)*. Berlin, Germany: Springer Verlag, 1997, vol. 454.

[203] H. R. Künsch, “State space and hidden Markov models,” in *Complex Stochastic Systems*, O. E. Barndorff-Nielsen, D. R. Cox, and C. Kluppelberg, Eds. Boca Raston, FL: Chapman & Hall/CRC Press, 2001, pp. 109–173.

[204] A. Lapidoth and J. Ziv, “On the universality of the LZ-based decoding algorithm,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 1746–1755, Sept. 1998.

[205] A. Lapidoth and P. Narayan, “Reliable communication under channel uncertainty,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 2148–2177, Oct. 1998.

[206] B. Larget, “A canonical representation for aggregated Markov processes,” *J. Appl. Probab.*, vol. 35, pp. 313–324, 1998.

[207] N. D. Le, B. G. Leroux, and M. L. Puterman, “Exact likelihood evaluation in a Markov mixture model for time series of seizure counts,” *Biometrics*, vol. 48, pp. 317–323, Mar. 1992.

[208] J.-P. Le Cadre and O. Tremois, “Bearing-only tracking for maneuvering sources,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 34, pp. 179–193, Jan. 1998.

[209] C.-H. Lee, F. K. Soong, and K. K. Paliwal, Eds., *Automatic Speech and Speaker Recognition*. Boston, MA: Kluwer, 1996.

[210] F. Le Gland and L. Mevel, “Exponential forgetting and geometric ergodicity in hidden Markov models,” *Math. Contr. Signals Syst.*, vol. 13, pp. 63–93, 2000.

[211] E. L. Lehmann, *Theory of Point Estimation*. Pacific Grove, CA: Wadsworth & Brooks/Cole, 1991.

[212] B. G. Leroux and M. L. Puterman, “Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models,” *Biometrics*, vol. 48, pp. 545–558, June 1992.

[213] B. G. Leroux, “Consistent estimation of a mixing distribution,” *Ann. Statist.*, vol. 20, no. 3, pp. 1350–1360, 1992.

[214] ——, “Maximum-likelihood estimation for hidden Markov models,” *Stochastic Processes Their Applic.*, vol. 40, pp. 127–143, 1992.

[215] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, “An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition,” *Bell Syst. Tech. J.*, vol. 62, no. 4, pp. 1035–1074, Apr. 1983.

[216] S. E. Levinson, “Continuously variable duration hidden Markov models for automatic speech recognition,” *Comput., Speech Language*, vol. 1, pp. 29–45, 1986.

[217] J. Li, A. Najmi, and R. M. Gray, “Image classification by two-dimensional hidden Markov model,” *IEEE Trans. Signal Processing*, vol. 48, pp. 517–532, Feb. 2000.

[218] Y. Linde, A. Buzo, and R. M. Gray, “An algorithm for vector quantizer design,” *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.

[219] G. Lindgren, “Markov regime models for mixed distributions and switching regressions,” *Scan. J. Statist.*, vol. 5, pp. 81–91, 1978.

[220] G. Lindgren and U. Holst, “Recursive estimation of parameters in Markov-modulated Poisson processes,” *IEEE Trans. Commun.*, vol. 43, pp. 2812–2820, Nov. 1995.

[221] L. A. Liporace, “Maximum Likelihood estimation for multivariate observations of Markov sources,” *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 729–734, Sept. 1982.

[222] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes, Part I*. New York: Springer-Verlag, 1977.

[223] C.-C. Liu and P. Narayan, “Order estimation and sequential universal data compression of a hidden Markov source by the method of mixtures,” *IEEE Trans. Inform. Theory*, vol. 40, pp. 1167–1180, July 1994.

[224] A. Ljolje, Y. Ephraim, and L. R. Rabiner, “Estimation of hidden Markov model parameters by minimizing empirical error rate,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Albuquerque, NM, Apr. 1990, pp. 709–712.

[225] T. A. Louis, “Finding the observed information matrix when using the EM algorithm,” *J. Roy. Statist. Soc. B*, vol. 44, no. 2, pp. 226–233, 1982.

[226] D. G. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1984.

[227] A. V. Lukashin and M. Borodovsky, “GeneMark.hmm: New solutions for gene finding,” *Nucleic Acids Res.*, vol. 26, no. 4, pp. 1107–1115, 1998.

[228] I. L. MacDonald and W. Zucchini, *Hidden Markov and Other Models for Discrete-Valued Time Series*. London, U.K.: Chapman & Hall, 1997.

[229] D. T. Magill, “Optimal adaptive estimation of sampled stochastic processes,” *IEEE Trans. Automat. Contr.*, vol. AC-10, no. 4, pp. 434–439, Oct. 1965. Cf. Author’s reply, *IEEE Trans. Automat. Contr.*, vol. AC-14, pp. 216–218, Apr. 1969.

[230] J. Makhoul and R. Schwartz, “What is a hidden Markov model?,” *IEEE Spectrum*, vol. 34, pp. 44–47, Dec. 1997.

[231] M. Maxwell and M. Woodroofe, “A local limit theorem for hidden Markov chains,” *Statist. Probab. Lett.*, vol. 32, pp. 125–131, 1997.

[232] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.

[233] K. S. Meier-Hellstern, “A fitting algorithm for Markov-modulated Poisson processes having two arrival rates,” *Europ. J. Opt. Res.*, vol. 29, pp. 370–377, 1987.

[234] N. Merhav and Y. Ephraim, “Hidden Markov modeling using a dominant state sequence with application to speech recognition,” *Computer, Speech, and Language*, vol. 5, pp. 327–339, Oct. 1991.

[235] N. Merhav, “Universal classification for hidden Markov models,” *IEEE Trans. Inform. Theory*, vol. 37, pp. 1586–1594, Nov. 1991.

[236] ——, “Universal coding with minimum probability of codeword length overflow,” *IEEE Trans. Inform. Theory*, vol. 37, pp. 556–563, May 1991.

[237] N. Merhav and J. Ziv, “A Bayesian approach for classification of Markov sources,” *IEEE Trans. Inform. Theory*, vol. 37, pp. 1067–1071, July 1991.

[238] N. Merhav and Y. Ephraim, “A Bayesian classification approach with application to speech recognition,” *IEEE Trans. Signal Processing*, vol. 39, pp. 2157–2166, Oct. 1991.

[239] N. Merhav, “Universal detection of messages via finite-state channels,” *IEEE Trans. Inform. Theory*, vol. 46, pp. 2242–2246, Sept. 2000.

[240] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. New York: Springer-Verlag, 1996.

[241] D. J. Miller and M. Park, “A sequence-based approximate MMSE decoder for source coding over noisy channels using discrete hidden Markov models,” *IEEE Trans. Commun.*, vol. 46, pp. 222–231, Feb. 1998.

[242] G. E. Monahan, “A survey of partially observable Markov decision processes: Theory, models, and algorithm,” *Manag. Sci.*, vol. 28, no. 1, pp. 1–16, 1982.

[243] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert–Elliott channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1277–1290, Nov. 1989.

[244] A. Nádas, "Optimal solution of a training problem in speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 326–329, Feb. 1985.

[245] D. F. Nicholls and B. G. Quinn, *Random Coefficients Autoregressive Models: An Introduction. Lecture Notes in Statist.*. Berlin, Germany: Springer-Ferlag, 1982, vol. 11.

[246] Y. Normandin, R. Cardin, and R. De Mori, "High-performance connected digit recognition using maximum mutual information estimation," *IEEE Trans. Speech Audio Processing*, vol. SAP-2, pp. 299–311, Apr. 1994.

[247] B. Øksendal, *Stochastic Differential Equations*, 5th ed. Berlin, Germany: Springer-Verlag, 1998.

[248] G. Ott, "Compact encoding of stationary Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 82–86, Jan. 1967.

[249] M. Park and D. J. Miller, "Low-delay optimal MAP state estimation in HMM's with application to symbol decoding," *IEEE Signal Processing Lett.*, vol. 4, pp. 289–292, Oct. 1997.

[250] A. Paz, *Introduction to Probabilistic Automata*. New York: Academic, 1971.

[251] T. Petrie, "Probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, vol. 40, no. 1, pp. 97–115, 1969.

[252] N. Phamdo and N. Farvardin, "Optimal detection of discrete Markov sources over discrete memoryless channels—Applications to combined source-channel coding," *IEEE Trans. Inform. Theory*, vol. 40, pp. 186–193, Jan. 1994.

[253] E. Plotnik, M. J. Weinberger, and J. Ziv, "Upper bounds on the probability of sequences emitted by finite-state sources and on the redundency of the Lempel–Ziv algorithm," *IEEE Trans. Inform. Theory*, vol. 38, pp. 66–72, Jan. 1992.

[254] A. B. Poritz, "Linear predictive hidden Markov models," in *Proc. Symp. Application of Hidden Markov Models to Text and Speech*, J. D. Ferguson, Ed. Princeton, NJ: IDA-CRD, 1980, pp. 88–142.

[255] A. B. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1982, pp. 1291–1294.

[256] ——, "Hidden Markov models: A guided tour," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1988, pp. 7–13.

[257] B. R. Povlow and S. M. Dunn, "Texture classification using noncausal hidden Markov models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, pp. 1010–1014, Oct. 1995.

[258] M. B. Priestley, *Spectral Analysis and Time Series*. New York: Academic, 1994.

[259] J. G. Proakis and M. Salehi, *Communication Systems Engineering*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 2002.

[260] W. Qian and D. M. Titterington, "Estimation of parameters in hidden Markov models," *Phil. Trans. Roy. Soc.*, vol. 337, pp. 407–428, 1991. London Series A.

[261] L. R. Rabiner, J. G. Wilpon, and B.-H. Juang, "A segmental k-means training procedure for connected word recognition," *AT&T Tech. J.*, vol. 65, pp. 21–40, May–June 1986.

[262] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[263] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[264] G. Radons, J. D. Becker, B. Dülfer, and J. Krüger, "Analysis, classification, and coding of multielectrode spike trains with hidden Markov models," *Biol. Cybern.*, vol. 71, pp. 359–373, 1994.

[265] J. Raviv, "Decision making in Markov chains applied to the problem of pattern recognition," *IEEE Trans. Inform. Theory*, vol. IT-3, pp. 536–551, Oct. 1967.

[266] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.*, vol. 26, no. 2, pp. 195–239, Apr. 1984.

[267] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[268] ——, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July 1984.

[269] C. P. Robert, G. Celeux, and J. Diebolt, "Bayesian estimation of hidden Markov chains: A stochastic implementation," *Statist. Probab. Lett.*, vol. 16, pp. 77–83, 1993.

[270] C. P. Robert, "Mixtures of distributions: Inference and estimation," in *Markov Chain Monte Carlo In Practice*, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds. London, U.K.: Chapman & Hall, 1996.

[271] C. P. Robert and D. M. Titterington, "Reparameterization strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation," *Statist. Comput.*, vol. 8, pp. 145–158, 1998.

[272] M. Rosenblatt, *Markov Processes, Structure and Asymptotic Behavior*. New York: Springer-Verlag, 1971.

[273] T. Rydén, "Parameter estimation for Markov modulated Poisson processes," *Commun. Statist. Stochastic Models*, vol. 10, no. 4, pp. 795–829, 1994.

[274] ——, "Consistency and asymptotically normal parameter estimates for hidden Markov models," *Ann. Statisti.*, vol. 22, no. 4, pp. 1884–1895, 1994.

[275] ——, "An EM algorithm for estimation in Markov-modulated Poisson processes," *Comput. Statist. Data Anal.*, vol. 21, pp. 431–447, 1996.

[276] ——, "Consistent and asymptotically normal parameter estimates for Markov modulated Poisson processes," *Scand. J. Statist.*, vol. 22, pp. 295–303, 1995.

[277] ——, "Estimating the order of hidden Markov models," *Statistics*, vol. 26, pp. 345–354, 1995.

[278] ——, "On identifiability and order of continuous-time aggregated Markov chains, Markov-modulated Poisson processes, and phase-type distributions," *J. Appl. Probab.*, vol. 33, pp. 640–653, 1996.

[279] ——, "On recursive estimation for hidden Markov models," *Stochastic Processes Their Applic.*, vol. 66, pp. 79–96, 1997.

[280] T. Rydén and D. M. Titterington, "Computational Bayesian analysis of hidden Markov models," *J. Comput. Graph. Statist.*, vol. 7, no. 2, pp. 194–211, 1998.

[281] T. Rydén, "Asymptotically efficient recursive estimation for incomplete data models using the observed information," *Metrika*, vol. 47, pp. 119–145, 1998.

[282] T. Rydén, T. Teräsvirta, and S. Åsbrink, "Stylized facts of daily returns series and the hidden Markov model," *J. Appl. Econ.*, vol. 13, pp. 217–244, 1998.

[283] M. J. Sabin and R. M. Gray, "Global convergence and empirical consistency of the generalized Lloyd algorithm," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 148–155, Mar. 1986.

[284] J. Sansom, "A hidden Markov model for rainfall using breakpoint data," *J. Climate*, vol. 11, pp. 42–53, Jan. 1998.

[285] L. L. Scharf, D. D. Cox, and C. J. Masreliez, "Modulo-2π phase sequence estimation," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 615–620, Sept. 1980.

[286] L. L. Scharf, *Statistical Signal Processing*. New York: Addison-Wesley, 1991.

[287] G. Schwarz, "Estimating the dimension of a model," *Ann. Statisti.*, vol. 6, no. 2, pp. 461–464, 1978.

[288] A. Segall, "Stochastic processes in estimation theory," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 275–286, May 1976.

[289] ——, "Recursive estimation from discrete-time point processes," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 422–431, July 1976.

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

[291] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 26–37, Jan. 1980. Cf. comments and corrections, *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 942–943, Nov. 1983.

[292] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Ser. Anal.*, vol. 3, no. 4, pp. 253–264, 1982.

[293] P. Smyth, "Hidden Markov models for fault detection in dynamic systems," *Pattern Recogn.*, vol. 27, no. 1, pp. 149–164, 1994.

[294] ——, "Markov monitoring with unknown states," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 1600–1612, Sept. 1994.

[295] D. L. Snyder and M. I. Miller, *Random Point Processes in Time and Space*. New York: Springer, 1991.

[296] R. L. Streit and R. F. Barrett, "Frequency line tracking using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 586–598, Apr, 1990.

[297] H. Teicher, "Identifiability of mixtures of product measures," *Ann. Math. Statist.*, vol. 38, no. 4, pp. 1300–1302, 1967.

[298] L. Thoraval, G. Carrault, and J. J. Bellanger, "Heart signal recognition by hidden Markov models—The ECG case," *Methods Inform. Medicine*, vol. 33, pp. 10–14, 1994.

[299] D. M. Titterington, "Comments on 'Application of the conditional population-mixture model to image segmentation'," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 656–657, Sept. 1984.

[300] ——, "Recursive parameter estimation using incomplete data," *J. Roy. Statist. Soc. B*, vol. 46, no. 2, pp. 257–267, 1984.

[301] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixtures Distributions*. New York: Wiley, 1985.

[302] W. Turin, "MAP symbol decoding in channels with error bursts," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1832–1838, July 2001.

[303] T. R. Turner, M. A. Cameron, and P. J. Thomson, "Hidden Markov chains in generalized linear models," *Canad. J. Statist.*, vol. 26, no. 1, pp. 107–125, 1998.

[304] V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data*. New York: Springer-Verlag, 1982.

[305] ——, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[306] L. Venkataramanan, J. L. Walsh, R. Kuc, and F. J. Sigworth, "Identification of hidden Markov models for ion channel currents—Part I: Colored background noise," *IEEE Trans. Signal Processing*, vol. 46, pp. 1901–1915, July 1998.

[307] L. Venkataramanan, R. Kuc, and F. J. Sigworth, "Identification of hidden Markov models for ion channel currents—Part II: State-dependent excess noise," *IEEE Trans. Signal Processing*, vol. 46, pp. 1916–1929, July 1998.

[308] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 260–269, Apr. 1967.

[309] M. J. Weinberger, N. Merhav, and M. Feder, "Optimal sequential probability assignment for individual sequences," *IEEE Trans. Inform. Theory*, vol. 40, pp. 384–396, Mar. 1994.

[310] E. Weinstein, M. Feder, and A. V. Oppenheim, "Sequential algorithms for parameter estimation based on the Kullback–Leibler information measure," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1652–1654, Sept. 1990.

[311] L. R. Welch, unpublished work.

[312] C. J. Wellekens, "Explicit time correlation in hidden Markov models for speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Apr. 1987, pp. 384–386.

[313] L. B. White, "Cartesian hidden Markov models with applications," *IEEE Trans. Signal Processing*, vol. 40, pp. 1601–1604, June 1992.

[314] L. B. White, R. Mahony, and G. D. Brushe, "Lumpable hidden Markov models—Model reduction and reduced complexity filtering," *IEEE Trans. Automat. Contr.*, vol. 45, pp. 2297–2306, Dec. 2000.

[315] W. M. Wonham, "Some applications of stochastic differential equations to optimal nonlinear filtering," *SIAM J. Contr.*, ser. A, vol. 2, no. 3, pp. 347–369, 1965.

[316] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 1983.

[317] W. -R Wu and S.-C. Wei, "Rotation and Gray-scale transform-invariant texture classification using spiral resampling, subband decomposition, and hidden Markov models," *IEEE Trans. Image Processing*, vol. 5, pp. 1423–1434, Oct. 1996.

[318] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 1–10, Jan. 1976.

[319] ——, "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1250–1258, Nov. 1989.

[320] X. Xie and R. J. Evans, "Multiple target tracking and multiple frequency line tracking using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 39, pp. 2659–2676, Dec. 1991.

[321] ——, "Multiple frequency line tracking with hidden Markov models—Further results," *IEEE Trans. Signal Processing*, vol. 41, pp. 334–343, Jan. 1993.

[322] ——, "Frequency-wavenumber tracking using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 41, pp. 1391–1394, Mar. 1993.

[323] Y.-C. Yao, "Estimation of noisy telegraph processes: Nonlinear filtering versus nonlinear smoothing," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 444–446, May 1985.

[324] O. Zeitouni and A. Dembo, "Exact filters for the estimation of the number of transitions of finite-state continuous-time Markov processes," *IEEE Trans. Inform. Theory*, vol. 34, pp. 890–893, July 1988.

[325] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?," *IEEE Trans. Inform. Theory*, vol. IT-38, pp. 1597–1602, Sept. 1992.

[326] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, Sept. 1978.

[327] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 453–460, July 1985.

[328] ——, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inform. Theory*, vol. 34, pp. 278–286, Mar. 1988.

[329] ——, "Compression, tests for randomness, and estimating the statistical model of an individual sequence," in *Proc. Sequences*, R. M. Capocelli, Ed. New York: Springer-Verlag, 1990, pp. 366–373.

[330] J. Ziv and N. Merhav, "Estimating the number of states of a finite-state source," *IEEE Trans. Inform. Theory*, vol. 38, pp. 61–65, Jan. 1992.

[331] W. Zucchini and P. Guttorp, "A hidden Markov model for space–time precipitation," *Water Resources Res.*, vol. 27, no. 8, pp. 1917–1923, Aug. 1991.