INVITED PAPER *Special Issue on IEICE Transactions*

What HMMs Can Do

Jeff A. BILMES†, ***Member***

SUMMARY

Since their inception almost fifty years ago, hidden Markov models (HMMs) have have become the predominant methodology for automatic speech recognition (ASR) systems — today, most state-of-the-art speech systems are HMM-based. There have been a number of ways to explain HMMs and to list their capabilities, each of these ways having both advantages and disadvantages. In an effort to better understand what HMMs can do, this tutorial article analyzes HMMs by exploring a definition of HMMs in terms of random variables and conditional independence assumptions. We prefer this definition as it allows us to reason more throughly about the capabilities of HMMs. In particular, it is possible to deduce that there are, in theory at least, no limitations to the class of probability distributions representable by HMMs. This paper concludes that, in search of a model to supersede the HMM (say for ASR), rather than trying to correct for HMM limitations in the general case, new models should be found based on their potential for better parsimony, computational requirements, and noise insensitivity.

key words: *Automatic Speech Recognition, Hidden Markov Models, HMMs, time-series processes, hand-writing recognition, graphical models, dynamic Bayesian networks, dynamic graphical models, stochastic processes, time-series densities, bio-informatics*

1. Introduction

More than any other statistical technique, the Hidden Markov model (HMM) has been most successfully applied to the automatic speech recognition (ASR) problem. Recent results have shown that HMMs are remarkably good even for difficult conversational speech-to-text [1] — the latest Switchboard word error rates are at around 13%. There have been many HMM tutorials [2]–[6]. In the widely read and now classic paper [6], an HMM is introduced as a collection of urns each containing a different proportion of colored balls. Sampling (generating data) from an HMM occurs by choosing a new urn based on only the previously chosen urn, and then choosing with replacement a ball from this new urn. The sequence of urn choices are not made public (and are said to be "hidden") but the ball choices are known (and are said to be "observed"). Along this line of reasoning, an HMM can be defined in such a generative way, where one first generates a sequence of hidden (urn) choices, and then generates a sequence of observed (ball) choices.

For statistical speech recognition, one is not only concerned about how HMMs generate data, but also, and more importantly, about an HMM's distributions over observations, and how those distributions for different utterances compare with each other. An alternative "non-generative" view of HMMs, therefore, can provide additional insight into what the capabilities of HMMs are, both in how they generate data and in how they might recognize and distinguish (or discriminate) between patterns.

To pursue this goal, this paper first provides a precise HMM definition, where an HMM is defined as a variable-size collection of random variables with an appropriate set of conditional independence properties. In an effort to better understand what HMMs can do, this paper also considers a list of properties, and discusses how they each might or might not apply to an HMM. In particular, it will be argued that, at least within the paradigm offered by statistical pattern classification [7]–[10], there is no general theoretical limit to HMMs given enough hidden states, rich enough observation distributions, sufficient training data, adequate computation, and appropriate training algorithms. Instead, only a particular instance of an HMM used, say, in a speech recognition system might be inadequate. This perhaps provides a reason for the continual speech-recognition accuracy improvements we have seen with HMM-based systems, and for the difficulty there has been in producing a model to supersede HMMs.

This paper does not argue, however, that HMMs should be the final technology for speech recognition. On the contrary, a main hope of this paper is to offer a better understanding of what HMMs can do, and consequently, a better understanding of their limitations so they may ultimately be improved upon in favor of a superior model. Indeed, HMMs are extremely flexible and might remain the preferred ASR method for quite some time. For speech recognition research, however, a main thrust should be searching for inherently more parsimonious models, ones that incorporate only the distinct properties of speech utterances relative to competing speech utterances. The rest of this paper is thus devoted to what HMMs can do.

Section 2 reviews random variables, conditional independence, and graphical models (Section 2.1), stochastic processes (Section 2.2), and discrete-time Markov chains (Section 2.3). Section 3 provides a definition of an HMM, that has both a generative and an "acceptive" point of view. Section 4 compiles a list of properties, and discusses how they might or might not apply to HMMs. Section 5 derives conditions for HMM accuracy in a Kullback-Leibler distance sense, proving a lower bound on the necessary number of hidden states. The section derives sufficient conditions as well. Section 6 reviews several alternatives to HMMs, and

Manuscript received October 3, 2005.

Manuscript revised December 20, 2005.

†J. Bilmes is with the University of Washington, Seattle, Dept. of Electrical Engineering. Box 352500, Seattle WA, 98195-2500, U.S.A. E-mail: <bilmes@ee.washington.edu>

concludes by presenting an intuitive criterion one might use when considering alternatives.

1.1 Notation

Measure theoretic principles are avoided in this paper, and discrete and continuous random variables are distinguished only where necessary. Capital letters (e.g., X, Q) will refer to random variables, lower case letters (e.g., x, q) will refer to values of those random variables, and script letters (e.g., $\mathcal{X}$, $\mathcal{Q}$) will refer to possible values so that $x \in \mathcal{X}$, $q \in \mathcal{Q}$. If X is distributed according to p, it will be written $X \sim p(X)$. Probabilities are denoted $p_X(X = x)$, $p(X = x)$, or $p(x)$ which are considered equivalent. For notational simplicity, $p(x)$ will at different times symbolize a continuous probability density or a discrete probability mass function. The distinction will be unambiguous when needed.

It will be necessary to refer to sets of integer indexed random variables. Let $A \triangleq \{a_1, a_2, \ldots, a_T\}$ be a set of T integers. Then $X_A \triangleq \{X_{a_1}, X_{a_2}, \ldots, X_{a_T}\}$. If $B \subset A$ then $X_B \subset X_A$. It will also be useful to define sets of integers using matlab-like ranges. As such, $X_{i:j}$ with $i < j$ will refer to the variables $X_i, X_{i+1}, \ldots, X_j$. $X_{<i} \triangleq \{X_1, X_2, \ldots, X_{i-1}\}$, and $X_{\neg t} \triangleq X_{1:T} \setminus X_t = \{X_1, X_2, \ldots, X_{t-1}, X_{t+1}, X_{t+2}, \ldots, X_T\}$ where T will be clear from the context, and $\setminus$ is the set difference operator. When referring to sets of T random variable, it will also be useful to define $X \triangleq X_{1:T}$ and $x \triangleq x_{1:T}$. Additional notation will be defined when needed.

2. Preliminaries

Because within an HMM lies a hidden Markov chain which in turn contains a sequence of random variables, it is useful to review a few prerequisite topics before beginning our analysis. Readers familiar with this material should skip directly to Section 3. Information theory, while necessary for a later section of this paper, is not reviewed and the reader is referred to the texts [11], [12].

2.1 Random Variables, Conditional Independence, and Graphical Models

A random variable takes on values (or in the continuous case, a range of values) with certain probabilities.† Different random variables might or might not have the ability to influence each other, a notion quantified by statistical independence. Two random variables X and Y are said to be (marginally) statistically independent if and only if $p(X = x, Y = y) = p(X = x)p(Y = y)$ for every value of x and y. This is written $X \perp\!\!\!\perp Y$. Independence implies that regardless of the outcome of one random variable, the probabilities of the outcomes of the other random variable stay the same.

Two random variables might or might not be independent of each other depending on knowledge of a third random variable, a concept captured by conditional independence. A random variable X is conditionally independent of a different random variable Y given a third random variable Z under a given probability distribution $p(\cdot)$, if the following relation holds:

$$p(X = x, Y = y|Z = z) = p(X = x|Z = z)p(Y = y|Z = z)$$

for all x, y, and z. This is written $X \perp\!\!\!\perp Y|Z$ and it is said that "X is independent of Y given Z under $p(\cdot)$". An equivalent definition is $p(X = x|Y = y, Z = z) = p(X = x|Z = z)$. The conditional independence of X and Y given Z has the following intuitive interpretation: if one has knowledge of Z, then knowledge of Y does not change one's knowledge of X and vice versa. Conditional independence is different from unconditional (or marginal) independence. Therefore, it might be true that $X \perp\!\!\!\perp Y$ but not true that $X \perp\!\!\!\perp Y|Z$. One valuable property of conditional independence follows: if $X_A \perp\!\!\!\perp Y_B|Z_C$, and subsets $A' \subset A$ and $B' \subset B$ are formed, then it follows that $X_{A'} \perp\!\!\!\perp Y_{B'}|Z_C$. Conditional independence is a powerful concept — when assumptions are made, a statistical model can undergo enormous simplifications. This is due to the fact that (conditional) independence implies that some factorization of the joint distribution exists. And given a factorization, it is possible to distribute sums inside of products, something that can yield huge computational savings [16]. Additional properties of conditional independence and factorization are presented in [17], [18].

Graphical models [17]–[21] are quite useful to describe conditional independence and its consequences. Graphical models are an abstraction that encompasses an extremely large set of statistical ideas. Specifically, a graphical model is a graph $\mathcal{G} = (V, E)$ where V is a set of vertices and the set of edges E is a subset of the set $V \times V$. A particular graphical model is associated with a collection of random variables and a family of probability distributions over that collection. The vertex set V is in one-to-one correspondence with the set of random variables. The edge set E of the model in one way or another specifies a set of conditional independence (or factorization) properties of the random variables that are true for every the member of the associated family. There are different types of graphical models. The set of conditional independence assumptions specified by a graphical model, and therefore the family of probability distributions it constitutes, depends on its type.

One particularly famous type of directed graphical model (DGM) (where edges are directed) is called a Bayesian network (BN) [18], [21], [22], although one should not ascribe any Bayesian statistical interpretation to these models. In a DGM, if an edge is directed from node A towards node B, then A is a parent of B and B is a child

†In this paper, explanations often use discrete random variables to avoid measure theoretic notation needed in the continuous case. See [13]–[15] for a precise treatment of continuous random variables. Note also that random variables may be either scalar or vector valued.

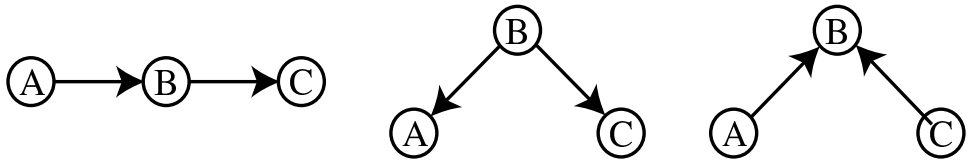

Fig. 1 Like any graphical model, the edges in a BN determine the conditional independence properties over the corresponding variables. For a BN, however, the arrow directions make an important difference. The figure shows three networks with different arrow directions over the same random variables, A, B, and C. On the left side, the variables form a three-variable first-order Markov chain $A \rightarrow B \rightarrow C$ (see Section 2.3). In the middle graph, the same conditional independence property is realized although one of the arrows is pointing in the opposite direction. Both these networks correspond the property $A \perp\!\!\!\perp C|B$. These two networks do not, however, insist that A and B are not independent. The right network corresponds to the property $A \perp\!\!\!\perp C$ but it does not imply that $A \perp\!\!\!\perp C|B$.

of A. One may also discuss ancestors, descendants, etc. of a node. A Dynamic Bayesian Network (DBN) [23]–[26] is one type of DGM containing edges pointing in the direction of time. There are several equivalent schemata that may serve to formally define the conditional independence relationships implied by a BN[17]. This includes d-separation [18], [21], the directed local Markov property [17] (which states that a variable is conditionally independent of its non-descendants given its parents), and the Bayes-ball procedure [27]. An undirected graphical model (or a Markov random field [28]) is one where conditional independence among the nodes corresponds simply to graph separation, and therefore has a easier semantics than Bayesian networks. The family of distributions associated with BNs is different from the family associated with undirected models, but the intersection of the two families is known as the decomposable models [17]. Other types of graphical models include causal models [29], chain graphs [17], factor graphs [30], and dependency networks [31].

Nodes in a graphical model can be either *hidden* (they have unknown value and signify a true random variable), or they can be *observed*, which means that the values are known. In fact, HMMs are so named because they possess a Markov chain that is hidden. A node may at different times be either hidden or observed, and for different reasons. For example, if one asks "what is the probability $p(C = c|A = a)$?" for the left graph in Figure 1, then B is hidden and A is observed. If instead one asks "what is the probability $p(C = c|B = b)$ or $p(A = a|B = b)$?" then B is observed. A node may be hidden because of missing values of certain random variables in samples from a database. Moreover, when the query "is $A \perp\!\!\!\perp B|C$?" is asked of a graphical model, it is implicitly assumed that A and B are hidden and C is observed. In general, if the value is known (i.e., if "evidence" has been supplied) for a node, then it is considered observed — otherwise, it is considered hidden.

A key problem with graphical models is that of computing the probability of one subset of nodes given values of some other subset, a procedure called probabilistic inference. Inference using a network containing hidden variables must "marginalize" them away. For example, given $p(A, B, C)$, the computation of $p(a|c)$ may be performed as:

$$p(a|c) = \frac{p(a,c)}{p(c)} = \frac{\sum_b p(a,b,c)}{\sum_{a,b} p(a,b,c)}$$

in which b has been marginalized (or integrated) away in the numerator. Inference is essential both to make predictions and to learn the network parameters with, say, the EM algorithm [32] or other iterative schemes [33].

In this paper, graphical models will help explicate the HMM conditional independence properties. An additional important property of graphical models, however, is that they supply more efficient inference procedures [21] than just, ignoring conditional independence, marginalizing away all unneeded and hidden variables. Inference can be either exact, as in the popular junction tree algorithm [21] (of which the Forward-Backward or Baum-Welch algorithm [4], [34] is an example [35]), or can be approximate [36]–[40] since in the general case inference is NP-Hard [41].

2.2 Stochastic Processes, Discrete-time Markov Chains, and Correlation

A discrete-time stochastic process is a collection $\{X_t\}$ for $t \in 1{:}T$ of random variables ordered by the discrete time index t. In general, the distribution for each of the variables X_t can be arbitrary and different for each t. There may also be arbitrary conditional independence relationships between different subsets of variables of the process — this corresponds to a graphical model with edges between all or most nodes.

Certain types of stochastic processes are common because of their analytical and computational simplicity. One example follows:

Definition 2.1. Independent and Identically Distributed (i.i.d.) *The stochastic process is said to be i.i.d.[11], [42], [43] if the following condition holds:*

$$p(X_t = x_t, X_{t+1} = x_{t+1}, \ldots, X_{t+h} = x_{t+h}) = \prod_{i=0}^{h} p(X = x_{t+i})$$

for all t, for all $h \geq 0$, for all $x_{t:t+h}$, and for some distribution $p(\cdot)$ that is independent of the index t.

An i.i.d. process therefore comprises an ordered collection of independent random variables each one having exactly the same distribution. A graphical model of an i.i.d process contains no edges at all.

If the statistical properties of variables within a time-window of a stochastic process do not evolve over time, the process is said to be stationary.

Definition 2.2. Stationary Stochastic Process *The stochastic process $\{X_t : t \geq 1\}$ is said to be (strongly) stationary [13] if the two collections of random variables*

$$\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}$$

and

$$\{X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h}\}$$

have the same joint probability distributions for all n and h.

In the continuous case, stationarity means that $F_{X_{t_{1:n}}}(a) = F_{X_{t_{1:n}+h}}(a)$ for all a where $F(\cdot)$ is the cumulative distribution and a is a valid vector-valued constant of length n. In the discrete case, stationarity is equivalent to the condition

$$\begin{aligned} &P(X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_n} = x_n) \\ &= P(X_{t_1+h} = x_1, X_{t_2+h} = x_2, \ldots, X_{t_n+h} = x_n) \end{aligned}$$

for all $t_1, t_2, \ldots, t_n$, for all $n > 0$, for all $h > 0$, and for all x_i. Every i.i.d. processes is stationary.

The covariance between two random vectors X and Y is defined as:

$$\begin{aligned} &\mathrm{cov}(X, Y) \\ &= E[(X - EX)(Y - EY)'] = E(XY') - E(X)E(Y)' \end{aligned}$$

It is said that X and Y are uncorrelated if $\mathrm{cov}(X, Y) = \vec{0}$ (equivalently, if $E(XY') = E(X)E(Y)'$) where $\vec{0}$ is the zero matrix. If X and Y are independent, then they are uncorrelated, but not vice versa unless they are jointly Gaussian [13].

2.3 Markov Chains

A collection of discrete-valued random variables $\{Q_t : t \geq 1\}$ forms an n^{th}-order Markov chain [13] if

$$\begin{aligned} &P(Q_t = q_t | Q_{t-1} = q_{t-1}, Q_{t-2} = q_{t-2}, \ldots, Q_1 = q_1) = \\ &P(Q_t = q_t | Q_{t-1} = q_{t-1}, Q_{t-2} = q_{t-2}, \ldots, Q_{t-n} = q_{t-n}) \end{aligned}$$

for all $t \geq 1$, and all $q_1, q_2, \ldots, q_t$. In other words, given the previous n random variables, the current variable is conditionally independent of every variable earlier than the previous n. A three-variable first order Markov chain is depicted in the left network in Figure 1.

One often views the event $\{Q_t = i\}$ as if the chain is "in state i at time t" and the event $\{Q_t = i, Q_{t+1} = j\}$ as a transition from state i to state j starting at time t. This notion arises by viewing a Markov chain as a finite-state automata (FSA) [44] with probabilistic state transitions. In this case, the number of states corresponds to the cardinality of each random variable. In general, a Markov chain may have infinitely many states, but chain variables in this paper are assumed to have only finite cardinality.

An n^{th}-order Markov chain may always be converted into an equivalent first-order Markov chain [45] using the following procedure:

$$Q'_t \triangleq \{Q_t, Q_{t-1}, \ldots, Q_{t-n}\}$$

where Q_t is an n^{th}-order Markov chain. Then Q'_t is a first-order Markov chain because

$$\begin{aligned} &P(Q'_t = q'_t | Q'_{t-1} = q'_{t-1}, Q'_{t-2} = q'_{t-2}, \ldots, Q'_1 = q'_1) \\ &= P(Q_{t-n:t} = q_{t-n:t} | Q_{1:t-1} = q_{1:t-1}) \\ &= P(Q_{t-n:t} = q_{t-n:t} | Q_{t-n-1:t-1} = q_{t-n-1:t-1}) \\ &= P(Q'_t = q'_t | Q'_{t-1} = q'_{t-1}) \end{aligned}$$

This transformation implies that, given a large enough state space, a first-order Markov chain may represent any n^{th}-order Markov chain.

The statistical evolution of a Markov chain is determined by the state transition probabilities $a_{ij}(t) \triangleq P(Q_t = j | Q_{t-1} = i)$. In general, the transition probabilities can be a function both of the states at successive time steps and of the current time t. In many cases, it is assumed that there is no such dependence on t. Such a time-independent chain is called time-homogeneous (or just homogeneous) because $a_{ij}(t) = a_{ij}$ for all t.

The transition probabilities in a homogeneous Markov chain are determined by a transition matrix A where $a_{ij} \triangleq (A)_{ij}$. The rows of A form potentially different probability mass functions over the states of the chain. For this reason, A is also called a stochastic transition matrix (or just a transition matrix).

A state of a Markov chain may be categorized into one of three distinct categories [13]. A state i is said to be *transient* if, after visiting the state, it is possible for it never to be visited again, i.e.,:

$$p(Q_n = i \text{ for some } n > t | Q_t = i) < 1.$$

A state i is said to be *null-recurrent* if it is not transient but the expected return time is infinite (i.e., $E[\min\{n > t : Q_n = i\} | Q_t = i] = \infty$). Finally, a state is *positive-recurrent* if it is not transient and the expected return time to that state is finite. For a Markov chain with a finite number of states, a state can only be either transient or positive-recurrent.

Like any stochastic process, an individual Markov chain might or might not be a stationary process. The stationarity condition of a Markov chain, however, depends on 1) if the Markov chain transition matrix has (or "admits") a stationary distribution or not, and 2) if the current distribution over states is one of those stationary distributions.

If Q_t is a time-homogeneous stationary first-order Markov chain then:

$$P(Q_{t_1} = q_1, Q_{t_2} = q_2, \ldots, Q_{t_n} = q_n)$$
$$= P(Q_{t_1+h} = q_1, Q_{t_2+h} = q_2, \ldots, Q_{t_n+h} = q_n)$$

for all t_i, h, n, and q_i. Using the first order Markov property, the above can be written as:

$$\begin{aligned}
&P(Q_{t_n} = q_n | Q_{t_{n-1}} = q_{n-1}) \\
&\quad P(Q_{t_{n-1}} = q_{n-1} | Q_{t_{n-2}} = q_{n-2}) \\
&\quad\quad \ldots P(Q_{t_2} = q_2 | Q_{t_1} = q_1) P(Q_{t_1} = q_1) \\
&= P(Q_{t_n+h} = q_n | Q_{t_{n-1}+h} = q_{n-1}) \\
&\quad\quad P(Q_{t_{n-1}+h} = q_{n-1} | Q_{t_{n-2}+h} = q_{n-2}) \\
&\quad\quad\quad \ldots P(Q_{t_2+h} = q_2 | Q_{t_1+h} = q_1) P(Q_{t_1+h} = q_1)
\end{aligned}$$

Therefore, a time-homogeneous Markov chain is stationary only when $P(Q_{t_1} = q) = P(Q_{t_1+h} = q) = P(Q_t = q)$ for all $q \in \mathcal{Q}$ (clearly, this is an if and only if condition). This is called a stationary distribution of the Markov chain and will be designated by ξ with $\xi_i = P(Q_t = i)$.[†]

According to the definition of the transition matrix, a stationary distribution has the property that $\xi A = \xi$ implying that ξ must be a left eigenvector of the transition matrix A. For example, let $p_1 = [.5, .5]$ be the current distribution over a 2-state Markov chain (using matlab notation). Let $A_1 = [.3, .7; .7, .3]$ be the transition matrix. The Markov chain is stationary since $p_1 A_1 = p_1$. If the current distribution is $p_2 = [.4, .6]$, however, then $p_2 A_1 \neq p_2$, so the chain is no longer stationary.

In general, there can be more than one stationary distribution for a given Markov chain (as there can be more than one eigenvector of a matrix). The condition of stationarity for the chain, however, depends on if the chain "admits" a stationary distribution, and if it does, whether the current marginal distribution over the states is one of the stationary distributions. If a chain does admit a stationary distribution ξ, then $\xi_j = 0$ for all j that are transient and null-recurrent [13]; i.e., a stationary distribution has positive probability only for positive-recurrent states (states that are assuredly re-visited).

The time-homogeneous property of a Markov chain is distinct from the stationarity property. Stationarity, however, does implies time-homogeneity. To see this, note that if the process is stationary then $P(Q_t = i, Q_{t-1} = j) = P(Q_{t-1} = i, Q_{t-2} = j)$ and $P(Q_t = i) = P(Q_{t-1} = i)$. Therefore, $a_{ij}(t) = P(Q_t = i, Q_{t-1} = j)/P(Q_{t-1} = j) = P(Q_{t-1} = i, Q_{t-2} = j)/P(Q_{t-2} = j) = a_{ij}(t-1)$, so by induction $a_{ij}(t) = a_{ij}(t+\tau)$ for all τ, and the chain is time-homogeneous. On the other hand, a time-homogeneous Markov chain might not admit a stationary distribution and therefore never correspond to a stationary random process.

Note that an inhomogeneous Markov chain might seem like it has a stationary distribution, and in fact the marginal distribution of states at time t (or $p(Q_t = q)$) might not change over time for inhomogeneous Markov chains. For example, let $A_t = [.3, .7; .7, .3]$ when t is even and $A_t = [.4, .6; .6, .4]$ when t is odd. Then the Markov chain is inhomogeneous but if the current state distribution is $p = [.5, .5]$, then $pA_t = p$ for t both even and odd. Note that this is not a stationary distribution. When t is even, we have that $p(Q_t = 0, Q_{t+1} = 1) = 0.5 \times 0.3$ but when t is odd, $p(Q_t = 0, Q_{t+1} = 1) = 0.5 \times 0.4$, so the chain does not exhibit a stationary distribution according to the definition.

The idea of "probability flow" may help to determine if a first-order Markov chain admits a stationary distribution. Stationary, or $\xi A = \xi$, implies that for all i

$$\xi_i = \sum_j \xi_j a_{ji}$$

or equivalently,

$$\xi_i (1 - a_{ii}) = \sum_{j \neq i} \xi_j a_{ji}$$

which is the same as

$$\sum_{j \neq i} \xi_i a_{ij} = \sum_{j \neq i} \xi_j a_{ji}$$

The left side of this equation can be interpreted as the probability flow out of state i and the right side can be interpreted as the flow into state i. A stationary distribution requires that the inflow and outflow cancel each other out for every state.

3. Hidden Markov Models

Perhaps the earliest discussion of what is now called a hidden Markov model is given in by Shannon in [46]. This was followed by papers such as [46]–[49] where they were called "functions of finite Markov chains" (where the observation distributions were deterministic), and where learning the transition matrix A was termed the "identifiability problem." In [50], [51], the more general fully-stochastic HMMs were defined and termed "stochastic functions of finite Markov chains." A thorough account of the history of HMMs is given in [5]. There are also a number of accounts of HMMs and their use in speech recognition and beyond [2], [3], [5], [6], [52]–[58]. Inference in hidden Markov models was shown to be a special case of Bayesian network inference in [35].

Our definition of an HMM, presented next, is inspired by their generalizations to Bayesian networks:

Definition 3.1. Hidden Markov Model *A hidden Markov model (HMM) is collection of random variables consisting of a set of T discrete scalar variables $Q_{1:T}$ and a set of T other variables $X_{1:T}$ which may be either discrete or continuous (and either scalar- or vector-valued). These variables, collectively, possess the following conditional independence properties:*

[†]Note that in the speech recognition literature, the symbol π is often used to indicate the initial (time $t = 1$) state distribution which might or might not be stationary.

$$\{Q_{t:T}, X_{t:T}\} \perp\!\!\!\perp \{Q_{1:t-2}, X_{1:t-1}\} | Q_{t-1} \quad (1)$$

and

$$X_t \perp\!\!\!\perp \{Q_{\neg t}, X_{\neg t}\} | Q_t \quad (2)$$

for each $t \in 1 : T$. No other conditional independence properties are true in general, unless they follow from Equations 1 and 2. The length T of these sequences is itself an integer-valued random variable having a complex distribution (see Section 4.7).

Let us suppose that each Q_t may take values in a finite set, so $Q_t \in \mathcal{Q}$ where $\mathcal{Q}$ is called the state space which has cardinality $|\mathcal{Q}|$.

Equations (1) and (2) imply a large assortment of conditional independence statements. Equation 1 states that the future is conditionally independent of the past given the present. One implication† is that $Q_t \perp\!\!\!\perp Q_{1:t-2} | Q_{t-1}$ which means the variables $Q_{1:T}$ form a discrete-time, discrete-valued, first-order Markov chain. Another implication of Equation 1 is $Q_t \perp\!\!\!\perp \{Q_{1:t-2}, X_{1:t-1}\} | Q_{t-1}$ which means that X_τ is unable, given Q_{t-1}, to affect Q_t for $\tau < t$. This does not imply, given Q_{t-1}, that Q_t is unaffected by future variables. In fact, the distribution of Q_t could dramatically change, even given Q_{t-1}, when the variables X_τ or $Q_{\tau+1}$ change, for $\tau > t$.

The other variables $X_{1:T}$ form a general discrete time stochastic process with, as we will see, great flexibility. Equation 2 states that given an assignment to Q_t, the distribution of X_t is independent of every other variable (both in the future *and* in the past) in the HMM. One implication is that $X_t \perp\!\!\!\perp X_{t+1} | \{Q_t, Q_{t+1}\}$ which follows since $X_t \perp\!\!\!\perp \{X_{t+1}, Q_{t+1}\} | Q_t$ and $X_t \perp\!\!\!\perp X_{t+1} | Q_{t+1}$.

Definition 3.1 does not limit the number of states $|\mathcal{Q}|$ in the Markov chain, does not require the observations $X_{1:T}$ to be either discrete, continuous, scalar-, or vector- valued, does not designate the implementation of the dependencies (e.g., general regression, probability table, neural network, etc.), does not determine the model families for each of the variables (e.g., Gaussian, Laplace, etc.), does not force the underlying Markov chain to be time-homogeneous, and does not fix the parameters or any tying mechanism. Note that the HMM makes *no* marginal independence assumptions, meaning nothing in an HMM is independent of anything else. That is, no statements are made of the form $A \perp\!\!\!\perp B$, where A and B are sets of variables — only assumptions of conditional independence exist in an HMM, which are of the form $A \perp\!\!\!\perp B | C$.

Any joint probability distribution over an appropriately typed set of random variables that obeys the above set of conditional independence rules is then an HMM. The two above conditional independence properties imply that, for a given T, the joint distribution over all the variables may be expanded as shown in Figure 2:

To parameterize an HMM, one therefore needs the following quantities: 1) the distribution over the initial chain variable $p(q_1)$, 2) the conditional "transition" distributions for the first-order Markov chain $p(q_t|q_{t-1})$, and 3) the conditional distribution for the other variables $p(x_t|q_t)$. It can be seen that these quantities correspond to the classic HMM definition [34]. Specifically, the initial (not necessarily stationary) distribution is labeled π which is a vector of length $|\mathcal{Q}|$. Then, $p(Q_1 = i) = \pi_i$. where π_i is the i^{th} element of π. The observation probability distributions are notated $b_j(x) = p(X_t = x | Q_t = j)$ and the associated parameters depend on $b_j(x)$'s family of distributions. Also, the Markov chain is typically assumed to be time-homogeneous, with stochastic matrix A where $(A)_{ij} = p(Q_t = j | Q_{t-1} = i)$ for all t. HMM parameters are often symbolized collectively as $\lambda \triangleq (\pi, A, B)$ where B represents the parameters corresponding to all the observation distributions.

For speech recognition, the Markov chain $Q_{1:T}$ is typically hidden, which naturally results in the name *hidden* Markov model. The variables $X_{1:T}$ are typically observed. These are the conventional variable designations but need not always hold. For example, X_τ could be missing or hidden, for some or all τ. In some tasks, $Q_{1:T}$ might be known and $X_{1:T}$ might be hidden. The name "HMM" applies in any case, even if $Q_{1:T}$ are not hidden and $X_{1:T}$ are not observed. Regardless, $Q_{1:T}$ will henceforth refer to the hidden variables and $X_{1:T}$ the observations.

With the above definition, an HMM can be simultaneously viewed as a generator and a stochastic acceptor. Like any random variable, say Y, one may obtain a sample from that random variable (e.g., flip a coin), or given a sample, say y, one may compute the probability of that sample $p(Y = y)$ (e.g., the probability of heads). One way to sample from an HMM is to first obtain a complete sample from the hidden Markov chain (i.e., sample from all the random variables $Q_{1:T}$ by first sampling Q_1, then Q_2 given Q_1, and so on.), and then at each time point t produce a sample of X_t using $p(X_t|q_t)$, the observation distribution according to the hidden variable value at time t. This is the same as choosing first a sequence of urns and then a sequence of balls from each urn as described in [34]. To sample just from $X_{1:T}$, one follows the same procedure but then throws away the Markov chain $Q_{1:T}$.

It is important to realize that each sample of $X_{1:T}$ requires a new and different sample of $Q_{1:T}$. In other words, two different HMM observation samples typically originate from two different state assignments to the hidden Markov chain. Put yet another way, an HMM observation sample is obtained using the marginal distribution $p(X_{1:T}) = \sum_{q_{1:T}} p(X_{1:T}, q_{1:T})$ and not from the conditional distribution $p(X_{1:T}|q_{1:T})$ for some fixed hidden variable assignment $q_{1:T}$. As will be seen, this marginal distribution $p(X_{1:T})$ can be quite general.

Correspondingly, when one observes only the collection of values $x_{1:T}$, they have presumably been produced according to some specific but unknown assignment to the hidden variables. A given $x_{1:T}$, however, could have been produced from one of many different assignments to the hid-

†Recall Section 2.1.

$$
\begin{aligned}
p(x_{1:T}, q_{1:T}) &= p(x_T, q_T | x_{1:T-1}, q_{1:T-1}) p(x_{1:T-1}, q_{1:T-1}) && \text{Chain Rule of probability.} \\
&= p(x_T | q_T, x_{1:T-1}, q_{1:T-1}) p(q_T | x_{1:T-1}, q_{1:T-1}) p(x_{1:T-1}, q_{1:T-1}) && \text{Again, chain rule.} \\
&= p(x_T | q_T) p(q_T | q_{T-1}) p(x_{1:T-1}, q_{1:T-1}) && \text{Since } X_T \perp\!\!\!\perp \{X_{1:T-1}, Q_{1:T-1}\} | Q_T \text{ and.} \\
& && Q_T \perp\!\!\!\perp \{X_{1:T-1}, Q_{1:T-2}\} | Q_{T-1} \text{ which follow from Definition 3.1} \\
&= \ldots \\
&= p(q_1) \prod_{t=2}^{T} p(x_t | q_t) p(q_t | q_{t-1})
\end{aligned}
$$