A SIMPLE CALCULATION OF THE JOINT MOMENTS OF HIDDEN MARKOV MODELS

*Mehmet Karan** *Brian D. O. Anderson** *Robert C. Williamson* *

Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra 0200, Australia

ABSTRACT

In this paper, an algorithm presented by Streit to calculate the integer moments of the output sequence probabilities using the probability measures of the Hidden Markov Models (HMMs) is reformulated in a matrix algebra framework which allows a better understanding of the algorithm. This algorithm arises in the context of the classification of an unknown stochastic process by a set of HMMs using a test statistic which can be approximated by these integer moments. Using this new formulation, a simple way of calculating the joint moments directly from the parameters of the HMMs in a non-iterative way is presented, and an asymptotic analysis of the algorithm is performed.

1. INTRODUCTION

In [1], the problem of classification of an unknown stochastic process by a set of Hidden Markov Models (HMMs) is studied. A suboptimal classifier is proposed using a test statistic which can be approximated by the integer moments of the output sequence probabilities using the probability measures defined by the parameters of HMMs. The computational difficulty of calculating the integer moments was overcome by an algorithm given in the same paper. Streit's algorithm is an extension of usual forward-backward algorithm given in [2]. However, the formulation of the algorithm is somewhat cumbersome and does not give much insight. In particular, it is not clear how the algorithm behaves asymptotically. In this paper, we reformulate Streit's algorithm in a matrix algebra framework which leads to a simple way of calculating these integer moments. In addition to this the asymptotic analysis of this algorithm is done.

Another motivation for this paper is the conjecture stated in [1] which might lead a simple, indirect way of calculating the entropy rate, relative entropy rate and the Kullback-Leibler number of Hidden Markov Models using these integer moments of the output sequence probabilities. However, as it is shown in [3], one needs to calculate some specific fractional moments of the output sequence probabilities of the HMMs which are also difficult to compute.

Hidden Markov Models (HMMs) have been proved to be useful for practical applications, especially for speech recognition [2]. For these kinds of processes, it is assumed that there is an underlying Markov process which cannot be observed directly. A discrete state, discrete output HMM is defined using a state set $\mathcal{S} = \{1, 2, \ldots, N\}$ and an output set $\mathcal{O} = \{1, 2, \ldots, M\}$ where the output at time t of the HMM, $O(t) \in \mathcal{O}$, is either a deterministic or a probabilistic function of the state of the HMM at time t, $X(t) \in \mathcal{S}$. A HMM can be defined in terms of certain real parameters $\lambda = (A, B, \Pi)$. Here $A = [a_{ij}]_{N\times N}$ is the state probability transition matrix defined by

$$a_{ij} = \Pr\{X(t+1) = j \mid X(t) = i\} \quad i, j = 1, \ldots, N, \quad (1)$$

and $B = [b_{kj}]_{M\times N}$ is the output probability matrix where

$$b_{kj} = \Pr\{O(t) = k \mid X(t) = j\} \quad k = 1, \ldots, M \text{ and } j = 1, \ldots, N, \quad (2)$$

and $\Pi = [\ \pi_1,\ \pi_2,\ \ldots, \pi_N\ ]$ is the initial state probability vector defined by $\pi_i = \Pr\{X(0) = i\}$, $i = 1, \ldots, N$. If the HMM is stationary and A satisfies well known conditions ensuring the existence of a unique steady state probability vector for the states, one can postulate that Π is this vector, and then Π in λ is redundant.

If the parameters of a HMM are known, the probability of a consecutive output sequence of length T, $O_T = (O(1), \ldots, O(T))$ is found using the probability measure measure $P(\cdot)$ defined by

$$\mathrm{P}(O_T) = \Pi\, \mathcal{B}(O(1))\, A\, \mathcal{B}(O(2)) \ldots A\, \mathcal{B}(O(T))\, \underline{1}_N \quad (3)$$

where $\underline{1}_N$ is a N-dimensional column vector of ones and $\mathcal{B}(O(t))$ where $O(t) \in \{1, \ldots, M\}$, is a diagonal matrix obtained from the matrix B as

$$\mathcal{B}(O(t)) = \mathbf{diag}\{b_{O(t),1},\ b_{O(t),2},\ \ldots, b_{O(t),N}\} \quad . \quad (4)$$

2. MOMENTS OF OUTPUT SEQUENCE PROBABILITIES OF HMMS

Consider two HMMs $\lambda_i = (A^{(i)}, B^{(i)}, \Pi^{(i)})$ and $\lambda_j = (A^{(j)}, B^{(j)}, \Pi^{(j)})$ with N_i and N_j states and the same

*The authors wish to acknowledge the funding of the activities of the Cooperative Research Centre for Robust and Adaptive Systems by the Australian Commonwealth Government under the Cooperative Research Centres Program

0-7803-1775-0/94 $3.00 © 1994 IEEE

number M of outputs. Then the finite-time integer moments of the output sequence probabilities using the probability measure $P_j(\cdot)$ defined by the parameters of the HMM λ_j as in (3) with respect to $P_i(\cdot)$ are defined by

$$M_{ji}(k,T) = E\{P_j(O_T)^k\} = \sum_{O_T} P_i(O_T)P_j(O_T)^k \quad (5)$$

where O_T is an output sequence of length T. Here k is a positive integer and the expectation is calculated with respect to the probability measure $P_i(\cdot)$. Note that as T gets larger, the direct calculation of $M_{ji}(k,T)$ becomes computationally intractable. In [1], this computational difficulty was solved by deriving an algorithm for the calculation of $M_{ji}(k,T)$. Now, we will reformulate this algorithm in a matrix algebra framework.

As done in [1], in order to find $M_{ji}(k,T)$, let us calculate another quantity, $R(k,T)$ which is defined by

$$R(k,T) = \sum_{O_T} \prod_{\nu=0}^{k} P_\nu(O_T), \quad (6)$$

where $P_\nu(\cdot)$ is the probability measure defined by the parameters of the HMM $\lambda_\nu = (A^{(\nu)}, B^{(\nu)}, \Pi^{(\nu)})$ which has N_ν states and M outputs for $\nu = 0,1,\ldots,k$. Observe that $M_{ji}(k,T)$ can be calculated from $R(k,T)$ by selecting λ_0 as λ_i and λ_ν as λ_j for $\nu = 1,\ldots,k$.

The quantity $R(k,T)$ in (6) can be rewritten in terms of the state sequences $X_T^{(\nu)}$, which are of length T, as

$$\begin{aligned} R(k,T) &= \sum_{O_T} \left\{ \prod_{\nu=0}^{k} \{ \sum_{X_T^{(\nu)}} P_\nu(O_T|X_T^{(\nu)}) \; P_\nu(X_T^{(\nu)}) \} \right\} \\ &= \sum_{X_T^{(0)}} \cdots \sum_{X_T^{(k)}} \left\{ \sum_{O_T} \prod_{\nu=0}^{k} P_\nu(O_T|X_T^{(\nu)}) \right\} \\ &\qquad \times \left\{ \prod_{\nu=0}^{k} P_\nu(X_T^{(\nu)}) \right\} \\ &= \sum_{X^{(0)}(T)} \cdots \sum_{X^{(k)}(T)} \theta_T(X^{(0)}(T),\ldots,X^{(k)}(T)) \quad (7) \end{aligned}$$

where $X^{(0)}(T),\ldots,X^{(k)}(T)$ are state sequences of length T and $\theta_T(X^{(0)}(T),\ldots,X^{(k)}(T))$ is found from the the following recursive equation:

$$\begin{aligned} \theta_t(X^{(0)}(t),\ldots,X^{(k)}(t)) = & \left\{ \sum_{O(t)} \prod_{\nu=0}^{k} P_\nu(O(t)|X^{(\nu)}(t)) \right\} \\ \times \sum_{X^{(0)}(t-1)} \cdots \sum_{X^{(k)}(t-1)} & \left\{ \prod_{\nu=0}^{k} P_\nu(X^{(\nu)}(t)|X^{(\nu)}(t-1)) \right. \\ & \left. \times\, \theta_{t-1}(X^{(0)}(t-1),\ldots,X^{(k)}(t-1)) \right\} \end{aligned} \quad (8)$$

with the initial condition

$$\theta_1(X^{(0)}(1),\ldots,X^{(k)}(1)) = \left\{ \sum_{O(1)} \prod_{\nu=0}^{k} P_\nu(O(1)|X^{(\nu)}(1)) \right\} \left\{ \prod_{\nu=0}^{k} P_\nu(X^{(\nu)}(1)) \right\} \quad (9)$$

Now, define a column vector Θ_t whose entries are $\theta_t(X^{(0)}(t),\ldots,X^{(k)}(t))$ which are ordered lexicographically. For example, when $k = 1$, $N_0 = 2$ and $N_1 = 2$, the vector Θ_t becomes

$$\Theta_t = [\theta_t(1,1), \theta_t(1,2), \theta_t(2,1), \theta_t(2,2)]' \quad . \quad (10)$$

Then the recursive equation in (8) can be written as

$$\Theta_t = F \; \Theta_{t-1} \quad (11)$$

where the matrix F is obtained from

$$F = C \; \bar{A}' \; . \quad (12)$$

Here, in (12), C is a diagonal matrix, whose diagonal entries are $\Gamma(r_0,\ldots,r_k)$ ($r_\nu = 1,\ldots,N_\nu$ where $\nu = 1,\ldots,k$) which are also ordered lexicographically and which are defined by

$$\begin{aligned} \Gamma(r_0,\ldots,r_k) &= \sum_{O(t)} \left\{ \prod_{\nu=0}^{k} P_\nu(O(t)|X^{(\nu)}(t) = r_\nu) \right\} \\ &= \sum_{l=1}^{M} \left\{ \prod_{\nu=0}^{k} b_{l,r_\nu}^{(\nu)} \right\} \end{aligned} \quad (13)$$

The matrix $\bar{A}$ in (12) is defined by

$$\bar{A} = A^{(0)} \otimes A^{(1)} \otimes \ldots \otimes A^{(k)} \quad (14)$$

where $\otimes$ denotes the Kronecker product. Similarly, the initial condition in (9) can be rewritten in this framework as

$$\Theta_1 = C \; (\Pi^{(0)} \otimes \Pi^{(1)} \otimes \ldots \otimes \Pi^{(k)})' \quad (15)$$

Then using the relationship for $R(k,T)$, the finite-time moments, $M_{ji}(k,T)$, of output sequence probabilities of HMMs $\lambda_i = (A^{(i)}, B^{(i)}, \Pi^{(i)})$ and $\lambda_j = (A^{(j)}, B^{(j)}, \Pi^{(j)})$, , can be calculated by summing the entries of the column vector Θ_T, which can be obtained from the recursive equation in (11) by setting λ_0 to λ_i, and λ_ν to λ_j for $\nu = 1,\ldots,k$.

3. A SIMPLE CALCULATION OF INTEGER MOMENTS OF OUTPUT SEQUENCE PROBABILITIES OF HMMS

An analysis of the algorithm in (11) for calculating the finite-time integer moments of HMMs is possible using nonnegative matrix theory. Assume that F is a primitive matrix, i.e. it is irreducible and has only one eigenvalue of maximum modulus. Note that since the matrices C and $\bar{A}$ are nonnegative (because of the nonnegativity constraints on the parameters of the HMMs), the matrix F is a nonnegative matrix. Let $\rho(F)$ be the eigenvalue of the matrix F which has the maximum modulus, then from Perron-Frobenius theorem, $\rho(F)$ is positive. Also, the right and left eigenvectors of F corresponding to the eigenvalue $\rho(F)$, namely u and v, are positive and normalized such that $u'\,v = 1$. The asymptotic behaviour of algorithm in (11) can be understood using the following fact [4] :

$$\lim_{t\to\infty} [\rho(F)^{-1} F]^t = u\,v' \qquad (16)$$

Let $\bar{\rho}(F)$ be an eigenvalue of F which has the second largest modulus, then an approximate value for the calculation of finite-time moments of HMMs can be obtained from the fact that

$$\left\| \left[\frac{F}{\rho(F)}\right]^t - uv' \right\|_\infty \le K\beta^t \quad t = 1,2,3,\ldots \qquad (17)$$

where $\beta = \bar{\rho}(F)/\rho(F) < 1$ and K is a constant whose value depends on the matrix F. Thus, if the matrix F is primitive, then the finite-time moments $M_{ji}(k,T)$ can be approximated as

$$M_{ji}(k,T) \approx \rho(F)^T \ \underline{1}_{\bar{N}}^{'}\, u\, v'\, \Theta_1 \qquad (18)$$

where where the approximation error decreases asymptotically as in (17). In (18), the vector Θ_1 is given in (15) and $\bar{N}$ is equal to $N_i N_j^k$.

As can be seen from (18) the integer moments decrease exponentially and the rate of decrease is given by the maximum eigenvalue of the matrix F defined in (12).

Although in (18) a simple way of calculating the asymptotic moments of HMMs is given, it is not obvious how to test whether the matrix F is primitive or not by looking at the parameters of the HMMs a priori. Of course, it is possible to check whether the matrix F is irreducible or not and all eigenvalues of F can be found, and it can be checked whether the maximum eigenvalue has multiplicity one or not. However, it is desirable to simplify the primitivity condition on F to some conditions on the parameters of the HMMs. This is possible if the matrix F is restricted to be a positive matrix (rather than being just nonnegative). Note that when F is a positive matrix, it is primitive also.

Now, if the output probability matrices $B^{(i)}$ and $B^{(j)}$ are positive, then the diagonal entries of the matrix C are positive as well. If the state transition probability matrices $A^{(i)}$ and $A^{(j)}$ are positive matrices, then the matrix $\bar{A}$ is a positive matrix. Hence, the matrix F which is a product of C and the transpose of $\bar{A}$, is positive. Consequently, if the matrices $A^{(i)}$, $A^{(j)}$ and $B^{(i)}$ and $B^{(j)}$ are positive matrices, then the finite-time moments can be calculated by the approximate formula in (18).

The following example illustrates this fact.

Consider two HMMs $\lambda_i = (A^{(i)}, B^{(i)}, \Pi^{(i)})$ and $\lambda_j = (A^{(j)}, B^{(j)}, \Pi^{(j)})$ whose parameters are given by

$$A^{(i)} = \begin{bmatrix} 0.7 & 0.1 & 0.1 & 0.1 \\ 0.2 & 0.4 & 0.3 & 0.1 \\ 0.1 & 0.3 & 0.2 & 0.4 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}, \quad A^{(j)} = \begin{bmatrix} 0.2 & 0.4 & 0.1 & 0.3 \\ 0.4 & 0.3 & 0.1 & 0.2 \\ 0.3 & 0.1 & 0.4 & 0.2 \\ 0.2 & 0.1 & 0.3 & 0.4 \end{bmatrix},$$

$$B^{(i)} = \begin{bmatrix} 0.25 & 0.3 & 0.2 & 0.1 \\ 0.25 & 0.2 & 0.4 & 0.5 \\ 0.1 & 0.4 & 0.1 & 0.1 \\ 0.4 & 0.1 & 0.3 & 0.3 \end{bmatrix}, \quad B^{(j)} = \begin{bmatrix} 0.5 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.4 & 0.1 \\ 0.2 & 0.1 & 0.3 & 0.5 \end{bmatrix},$$

$$\Pi^{(i)} = [0.379\ 0.238\ , 0.195\ , 0.186\ ], \qquad \Pi^{(j)} = [1/4\ , 1/4\ , 1/4\ , 1/4\ ]. \qquad (19)$$

k	$\log(M_{ji}(k,T))$			
	T = 5		T = 50	
	(11)	(18)	(11)	(18)
1	-6.76197	-6.76043	-67.56654	-67.56492
2	-13.29795	-13.29624	-132.78779	-132.78566

Table 1. Comparison of the exact and approximate formulae for integer moments of HMMs

Table 1 shows the logarithm finite-time integer moments of $P_j(\cdot)$, the probability measure defined by the parameters of λ_j with respect to $P_i(\cdot)$, that are calculated using the exact formula in (11) and the approximate formula in (18), when T is equal to 5 and 50. The advantage of the result given in (18) is that after finding the maximal eigenvalue of the matrix, and the corresponding left and right eigenvectors, the integer moments $M_{ji}(k,T)$ can be calculated approximately for all T given a fixed k.

As can be seen from the Table 1, the approximation error in (18) for this example is very small even for small T.

4. CONCLUSION

In this paper, the algorithm in [1] to calculate the integer moments of probability measures of Hidden Markov models is reformulated. The asymptotic behaviour of this algorithm is analysed and an approximate expression which is easy to calculate for these moments is derived.

REFERENCES

[1] R. L. Streit, "The moments of matched and mismatched Hidden Markov Models," *IEEE Transactions on ASSP*, vol. 38, no. 4, pp. 610–622, 1990.

[2] L. R. Rabiner and B. H. Juang, "An introduction to Hidden Markov Models," *IEEE ASSP Magazine*, pp. 4–16, 1986.

[3] M. Karan, B. D. O. Anderson, and R. C. Williamson, "A note on the calculation of a probabilistic distance between hidden markov models," in *International Workshop on Intelligent Signal Processing and Communication Systems, Systems, ISPACS '93*, pp. 93–97, 1993.

[4] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.