FAST, NON-ITERATIVE ESTIMATION OF HIDDEN MARKOV MODELS

Håkan Hjalmarsson

S3-Automatic Control
Royal Institute of Technology
SE-100 44 Stockholm Sweden
hakan.hjalmarsson@s3.e.kth.se

Brett Ninness *

Electrical and Computer Engineering
University of Newcastle,
Callaghan 2308, Australia
brett@ee.newcastle.edu.au

ABSTRACT

The solution of many important signal processing problems depends on the estimation of the parameters of a Hidden Markov Model (HMM). Unfortunately, to date the only known methods for performing this estimation have been iterative, and therefore computationally demanding. By way of contrast, this paper presents a new fast and non-iterative method that utilizes certain recent 'state spaced subspace system identification' (4SID) ideas from the control theory literature. A short simulation example presented here indicates this new technique to be almost as accurate as Maximum-Likelihood estimation, but an order of magnitude less computationally demanding than the Baum–Welch (EM) algorithm.

1. INTRODUCTION

A stationary discrete Markov model is one in which a series of symbols, call them $\{S_t\}$, evolve with time (labeled as t) to take on values from a set $\{q_1, \cdots, q_n\}$ and according to a random law:

$$\mathbf{P}\{S_{t+1} = q_j \mid S_t = q_i\} = p_{i,j} = [P]_{i,j}$$

where $\mathbf{P}(x|y)$ is the probability of event x occurring conditional on the event y having occurred, and $[P]_{i,j} = p_{i,j}$ is a matrix of these probabilities. Specifically, $p_{i,j}$ is the probability of moving from state number i to state number j, and since the probability of moving to *some* state is one, then the rows of P must sum to one, so that P is a row stochastic matrix. An important consequence of this is that the vector e consisting of all ones is a right eigenvector of P with eigenvalue 1: $Pe = e$.

A discrete Hidden Markov Model (HMM) is one in which the states $\{S_t\}$ are not directly observed, but instead the symbols $\{O_t\}$ taking the possible values $\{y_1, \cdots, y_m\}$ are available, and which are only randomly linked to the 'underlying' $\{S_t\}$ via a law $\mathbf{P}(O_t = y_i \mid S_t = q_j) = b_j(i) = [B]_{i,j}$ where, again $[B]_{i,j} = b_j(i)$ is a matrix describing the random nature of the process, this time in terms of the discrete probability density functions $b_j(i)$ that hide the Markov states $\{S_t\}$ according to $b_j(i)$ being the probability of observing the output $O_t = y_i$ when the underlying Markov state is the j'th one $S_t = q_j$.

*This work was supported by the Australian Research Council and the Centre for Integrated Dynamics and Control (CIDAC)

Together with the initial probability distribution $\pi : [\pi]_i = \mathbf{P}\{S_0 = q_i\}$ the triple $\lambda = \{P, B, \pi\}$ completely describes the HMM. The solution of many signal processing problems such as speech recognition [3], target tracking, and certain communications problems depends on the use of Hidden Markov modeling and the associated estimation of the description λ from observations of a physical process.

One obvious approach to finding these estimates is to use the well known Maximum-Likelihood method wherein it is necessary to calculate the probability of an observed output sequence conditional on λ and then choose an estimate $\widehat{\lambda}$ of λ that maximises this estimate:

$$\widehat{\lambda} = \arg\max_{\lambda} \mathbf{P}(O_1, \cdots, O_N \mid \lambda). \tag{1}$$

Here, it has clearly been assumed that N observations of the output symbols $\{O_t\}$ are available. The indicated probability (likelihood) can be calculated by first noting that with the definition $\alpha_t(i) \triangleq \mathbf{P}(O_1, \cdots, O_t, S_t = q_i|\lambda)$

$$\mathbf{P}(O_1, \cdots, O_t \mid \lambda) = \sum_{i=1}^{n} \alpha_t(i) \tag{2}$$

where $\alpha_t(i)$ may be recursively calculated as

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{n} \alpha_t(i) p_{i,j} \right) b_j(O_{t+1}). \tag{3}$$

These equations are well known [3], the point of presenting them being to emphasise that the maximization of (1) is computationally non-trivial, there being a clear need to perform $N \times n^2$ multiplications in order to calculate the likelihood associated with a particular λ, quite apart from the computational load of actually iterating on the choice of λ to arrive at the maximizing value $\widehat{\lambda}$.

Pertaining to this latter issue, the form of (2) and (3) make it clear that no closed form solution exists for $\widehat{\lambda}$ given by (1) and given this, one of the most popular methods for instead iteratively finding $\widehat{\lambda}$ is the so-called 'Baum–Welch' method [3] which is a particular instance of the Expectation-Maximisation (EM) algorithm.

Since the main purpose of this paper is to propose an alternative non-iterative approach to finding an estimate of λ, it is instructive to first motivate the attraction of such an alternative by illustrating the performance of the Baum–Welch procedure on a simple example where five hundred samples

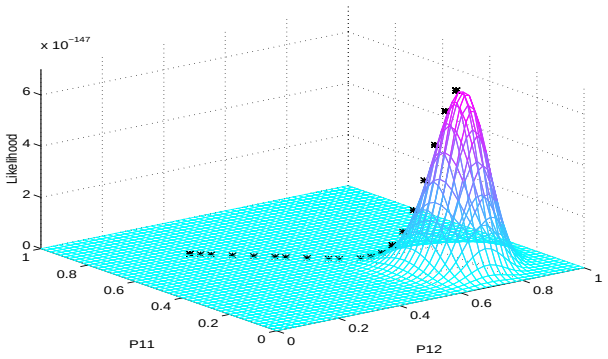

Figure 1: *20 iterations of the Baum Welch algorithm.*

of a 2-state Markov chain with state transition probabilities defined by $p_{1,1} = 0.3, p_{2,1} = 0.7$ are simulated, and then 'hidden' according to two output symbols being observed and being related to the underlying states according to the observation probabilities being $b_1(1) = 0.8, b_2(1) = 0.2$.

The likelihood surface for this situation, together with the first 20 iterations of the Baum-Welch algorithm in trying to find the maximum of this surface are shown in figure 1. Clearly, for this particular example at least 20 iterations are required to come close to convergence, which is a serious computational burden.

The rest of this paper will concern itself with illustrating a new non-iterative, and computationally cheap algorithm for estimating discrete state Hidden Markov Models.

2. STATE SPACE DESCRIPTION OF HMM'S

Central to the new methods of this paper is the idea of representing the HMM as a linear system in state space form:

$$\begin{aligned} x_{t+1} &= P^T x_t + w_t \quad (4) \\ y_t &= C x_t + \nu_t \quad (5) \end{aligned}$$

where $\{w_t\}$ and $\{\nu_t\}$ are particular random processes to be commented on in a moment. For this representation (4), (5) to work as a HMM description, the state vector x_t is constrained (without loss of generality) to contain all zeros save for a 1 in the location corresponding to the enumeration of which state the Markov chain is in at time t. As well, the observation y_t is constrained to be either zero or one. In this case, the k'th element of the row vector C is the probability of the output $\{y_t\}$ being 1 given that the state $x_t = q_k$. That is, C corresponds to the row of B associated with the output symbol $O_t = 1$.

Given this representation, the nature of the random processes $\{w_t\}$ and $\{\nu_t\}$ such that the output $\{y_t\}$ is the output of a HMM is not obvious. However, with some calculation it can be shown that $\mathbf{E}\{w_t\} = 0$, $\mathbf{E}\{\nu_t\} = 0$ and also that $\lim_{t\to\infty} \mathbf{E}\left\{\nu_t^2\right\} = Cm - CMC^T$, $\lim_{t\to\infty} \mathbf{E}\left\{w_t w_t^T\right\} = M - P^T M P$ where M is a diagonal matrix with the elements of the vector m on the diagonal, and m itself is defined by $\lim_{t\to\infty} \mathbf{E}\{x_t\} = \alpha\mu_1 \triangleq m$ where μ_1 is the eigenvector of P^T with eigenvalue 1 and α is a scalar constant. This immediately gives the second order asymptotic properties of $\{x_t\}$ and $\{y_t\}$ as $\lim_{t\to\infty} \mathbf{E}\{y_t^2\} = Cm$ and

$$\lim_{t\to\infty} \mathbf{E}\left\{x_t x_t^T\right\} = \lim_{t\to\infty} \operatorname{diag} \mathbf{E}\{x_t\} = \operatorname{diag}\{m\} \triangleq M$$

where $\operatorname{diag}\{x\}$ denotes a diagonal matrix with the elements of the vector x along the diagonal.

Therefore, HMM's can be afforded a state space description which is asymptotically wide-sense stationary. The paper now reviews completely separate work from the control theory literature which has recently been attracting great attention since it provides computationally cheap means for estimating such systems in state space form.

3. 4SID ESTIMATION

Of enormous recent interest in the area of system identification methods designed for control theory applications has been the study of so-called State Space Subspace Identification (4SID) [4]. Despite this recent interest, the ideas involved actually go back many years, at least to Akaike [1] whose approach was targeted at stochastic estimation problems pertinent to this paper.

For the purposes of explaining this, suppose one is presented with observations $\{y_t\}$ of a stationary stochastic process and is faced the task of estimating a state-space representation of this process in *innovations* form:

$$\begin{aligned} x_{t+1} &= A x_t + K e_t, \quad (6) \\ y_t &= C x_t + e_t \quad (7) \end{aligned}$$

where $\{e_t\}$ is a stationary white noise process. Via the idea of 'predictor space', Akaike [1] made clear for the first time that such a representation always exists, and in so doing suggested a way that it may be estimated from observations of $\{y_t\}$. This estimation method is now known with some simple modifications (involving user chosen weighting matrices) as 4SID estimation.

To explain these ideas, assume the availability of an output record $\{y_1, \cdots, y_N\}$ and form the matrices

$$[Y_p]_{k,j} = y_{t-k+j}, \quad ; k = 1, \cdots, t; j = 1, \cdots, N - 2t + 1 \quad (8)$$

$$[Y_f]_{k,j} = y_{t+k+j}, \quad ; k = 1, \cdots, t; j = 0, \cdots, N - 2t \quad (9)$$

and then try to predict Y_f from Y_p as (the subscripts p and f are meant to indicate 'past' and 'future' data)

$$Y_f = H Y_p.$$

The idea here is that the rows of H represent weights such that the columns of Y_f are the mean square optimal $1, \cdots, 2t$ step ahead predictors of y based on the past in the columns of Y_p. Take one column as an example:

$$\underbrace{\begin{pmatrix} y_{t+1} \\ \vdots \\ y_{2t} \end{pmatrix}}_{y_f} = H \underbrace{\begin{pmatrix} y_t \\ \vdots \\ y_1 \end{pmatrix}}_{y_p} + \underbrace{\begin{pmatrix} e_t \\ \vdots \\ e_1 \end{pmatrix}}_{v}.$$

The issue now is to find the 'predictor space' defined by Akaike in [1] to be the space spanning the projection (defined by the inner product of expectation over the underlying

probability space that the random variables are defined on) of y_f on y_p. In order to work this out, it is easiest to change to a new orthogonal basis: $\bar{y}_f = Ly_f$, $\bar{y}_p = Jy_p$ where the elements in $\bar{y}_f$ are uncorrelated with each other and are of unit norm (likewise for $\bar{y}_p$): $\mathbf{E}\{\bar{y}_f \bar{y}_f^T\} = I$, $\mathbf{E}\{\bar{y}_p \bar{y}_p^T\} = I$.

Finally, in order to work out the projection of $\bar{y}_f$ on $\bar{y}_p$ it will be advantageous if we can choose L and J such that

$$\mathbf{E}\left\{\bar{y}_f \bar{y}_p^T\right\} = \mathrm{diag}\{\sigma_1, \sigma_2, \cdots, \sigma_n, 0, \cdots, 0\}$$

so that the predictor space will easily be seen to be spanned by the first n elements in $\bar{y}_p$.

Finding an L and J to satisfy these requirements can be done by defining

$$R_{pp} = \mathbf{E}\left\{y_p y_p^T\right\}, \quad R_{ff} = \mathbf{E}\left\{y_f y_f^T\right\}, \quad R_{fp} = \mathbf{E}\left\{y_f y_p^T\right\}$$

and then calculating the SVD:

$$\begin{aligned} R_{ff}^{-1/2} R_{fp} R_{pp}^{-1/2} &= (U_1, U_2) \begin{pmatrix} S_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix} \\ &= U_1 S_1 V_1^T \end{aligned} \tag{10}$$

so as to be able to choose $J = V^T R_{pp}^{-1/2}$, $L = U^T R_{ff}^{-1/2}$ in which case

$$\mathbf{E}\left\{\bar{y}_f \bar{y}_p^T\right\} = S = \begin{pmatrix} S_1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Now, in practice the covariances R_{pp}, R_{pf}, R_{ff} are not available, but they can be estimated as

$$R_{pp} \approx \frac{1}{N} Y_p Y_p^T, \quad R_{ff} \approx \frac{1}{N} Y_f Y_f^T, \quad R_{fp} \approx \frac{1}{N} Y_f Y_p^T. \tag{11}$$

As well, in practice $\bar{y}_p$'s random variable behavior over the whole probability space it is defined on is unknown. However, it is possible to define its realisations from the observed sample data as

$$\bar{Y}_p = JY_p = \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix} R_{pp}^{-1/2} Y_p$$

so that realisations of the predictor space (state space) are the columns of

$$\widehat{X} = V_1^T R_{pp}^{-1/2} Y_p. \tag{12}$$

That is, $\widehat{X}$ is a fat matrix with estimates of the state realisations stacked up as columns: $\widehat{X} = (\widehat{x}_t, , \cdots, \widehat{x}_{N-t})$. This can be used to estimate the C matrix by solving

$$Y_t \triangleq (y_{t+1}, y_{t+2}, \cdots, y_{N-t+1}) = C\widehat{X}$$

in a least squares sense to give

$$\widehat{C} = (Y_t \widehat{X}^T)(\widehat{X}\widehat{X}^T)^{-1}. \tag{13}$$

This then leads to an estimate of the innovations as $\widehat{e} = Y_t - \widehat{C}\widehat{X}$ which can be substituted into the state update equation to give

$$\begin{aligned} \underbrace{(\widehat{x}_{t+1}, \cdots, \widehat{x}_{N-t})}_{\Psi} &= A \underbrace{(\widehat{x}_t, \cdots, \widehat{x}_{N-t-1})}_{\phi} + K\widehat{e} \\ &= (A, K) \underbrace{\begin{pmatrix} \phi \\ \widehat{e} \end{pmatrix}}_{\Phi} \end{aligned} \tag{14}$$

and then A and K can be estimated in a Least Squares sense as

$$(\widehat{A}, \widehat{K}) = \Psi\Phi^T(\Phi\Phi^T)^{-1}. \tag{15}$$

4. NON-ITERATIVE ESTIMATION OF HMM'S

The contribution of this paper is to suggest that based on the development of § 2 illustrating a state-space description (4), (5) for a HMM which was illustrated in § 2 to be asymptotically second order stationary, and based on the overview in § 3 of new 4SID methods for estimating state space model structures from wide-sense stationary observations, then the two ideas can be combined to devise a new fast and non-iterative method for estimating HMM's.

Specifically, the new method is as follows. First, use the observations $\{y_1, \cdots, y_N\}$ of the HMM to form the past and future Toeplitz and Hankel matrices Y_p and Y_f as in (8), (9). Then

1. Form estimates R_{pp}, R_{ff} and R_{fp} as in (11).
2. Form the singular value decomposition (10) to provide an estimate of the predictor space as in (12).
3. Use this to estimate B from $\widehat{C}$ given by (13) as
$$\widehat{B} = \begin{pmatrix} \widehat{C} \\ e^T - \widehat{C} \end{pmatrix}$$
where e is a column vector of all ones.
4. Estimate P as $\widehat{A}^T$ given by (15).

As will presently be shown by simulation, the results of this scheme are encouraging. However, before presenting them, it is essential to comment on the main drawback of the above method, which is that it is designed (by the orthogonality principles imbued by the SVD calculations) to estimate an *innovations* form (6), (7) realisation of the HMM process which in general will be different to the *positive* realisation. By the latter is meant the representation which is of more interest in applications in which all the elements of A and C are positive, A is a column stochastic matrix and the elements of C are less than one.

Therefore attention must be focussed on finding the state space transformation matrix T such that $T^{-1}\widehat{P}T, T\widehat{B}$ are of the required positive form. This problem is the so-called 'positive realisation problem', on which there has been much recent progress (see [2] and the references therein). The current state of the problem is that although it is known how to construct a positive realisation (if it exists), in general this realisation will be far from minimal, which is insufficient for the purposes here of finding a realisation of given order.

However, this problem is not insurmountable, at least for special cases of fixed dimension. For example, consider the simplest $n = 2$ dimensional case as an example, and denote by Q the matrix $Q = T^{-1}\widehat{P}T$ that is the row stochastic matrix similar to $\widehat{P}$ that is being sought. It must be of the form

$$Q = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

for some probabilities $p, q \in [0, 1]$, and hence the eigenvalues of Q must be 1 and $1 - (p + q)$. Generically, when using the new algorithm derived here on a 2×2 example, one eigenvalue

of $\widehat{P}$ will be very close to 1 and the other, call it λ will be well way from 1. Since similarity transformations preserve eigenvalues, then $\lambda = 1-(p+q)$, so that in fact Q can be written as

$$Q = \underbrace{\begin{pmatrix} \lambda & 1-\lambda \\ 0 & 1 \end{pmatrix}}_{\Lambda} + q \underbrace{\begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}}_{\Sigma}. \qquad (16)$$

Now, since $\widehat{P}$ does not necessarily have an eigenvalue *exactly* at 1, finding a row stochastic Q that is *exactly* similar to $\widehat{P}$ is impossible, so instead the search for one approximately similar is made as

$$\widehat{P}T - TQ \approx 0 \qquad (17)$$

where the precise meaning of $\approx$ will be clarified in a moment. In this case, using the Kronecker tensor product of matrices $\otimes$ and the vec$(\cdot)$ operator which turns a matrix into a vector by stacking its columns on top of one another, then the parameterisation (16) allows the condition (17) to be re-formulated as

$$\left[\underbrace{(I \otimes \widehat{P}) - (\Lambda^T \otimes I)}_{A} - q \underbrace{(\Sigma^T \otimes I)}_{B} \right] \underbrace{\mathrm{vec}\, T}_{\tau} \approx 0.$$

It is now elected to formally define the ≈ 0 notion as one in which the Euclidean norm $\tau^T[A-qB]^T[A-qB]\tau$ is minimised. However, this involves a joint minimisation over τ and q which is very difficult, so an alternative strategy is taken in which the Cauchy–Schwarz inequality is used to obtain an overbound $\tau^T\tau \|A-qB\|_F$ (with $\|\cdot\|_F$ being matrix Frobenius norm) for the above expression, and then focus attention on choosing q such that $\|A-qB\|_F$ is minimised.

With the notation $a = \mathrm{vec}\, A$, $b = \mathrm{vec}\, B$ then $\|A-qB\|_F^2 = \frac{1}{2}(a-qb)^T(a-qb)$, the minimisation of which with respect to q and subject to the constraint that $q \in [0,1]$ may be solved in closed form using Lagrange multiplier techniques. With this value of q in hand, it is then possible to perform the SVD

$$A - qB = USV^T = (U_1, U_2) \begin{pmatrix} S_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix}$$

so that an appropriate T is one in which $\mathrm{vec}\, T \in \mathrm{Span} V_2$.

5. SIMULATION STUDY

To illustrate the efficacy of the new method proposed here, the simulation example begun in § 1 is revisited, with the Baum-Welch algorithm illustrated there used thirty times on thirty different data sets, and the new non-iterative 4SID algorithm proposed here also applied to the same data sets. For the new algorithm, the method just outlined in § 4 was used to find the positive form of the estimate. For each data set, the estimated P matrix, call it $\widehat{P}$ was compared to the true P, call it P_T by calculating $\|\widehat{P} - P_T\|_F$. The results for the Baum-Welch algorithm (dash-dot line) and the new subspace method (solid line) for each of the thirty data sets are shown as the top plot in figure 2. Although the Maximum likelihood estimate is on average more accurate, the difference is not great. Moreover the computational cost paid for achieving this higher accuracy shown in the bottom plot of

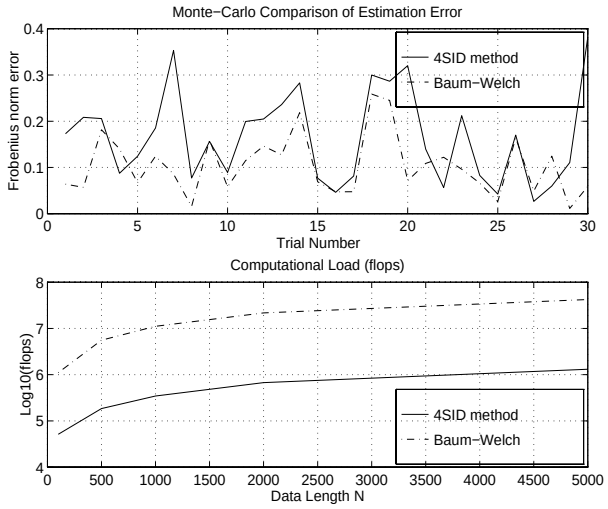

Figure 2: *Top plot is comparison of norm of estimation error for the new subspace method of this paper (solid line) and the ML estimate (dash-dot line) for thirty different HMM realisations. Bottom plot is a profile of the floating point computational load.*

figure 2 is seen to quite high - more than ten times the floating point calculations being required to calculate the Maximum Likelihood estimate.

6. CONCLUSION

This paper has presented a new method for the estimation of HMM's which has the advantage of being non-iterative and computationally cheap. Preliminary simulation study indicates it's performance to be similar to the Maximum Likelihood method. However, it suffers a serious drawback in that it provides estimates in innovations form, not positive form. For the simplest case of dimension 2, the paper illustrated a method for overcoming this problem. The admittedly ad-hoc nature of obtaining this positive form is not considered a serious drawback, since the central theme of this paper is to establish the potential of of an efficient new estimation method, not to provide a complete and general solution.

In any event, given the current vigorous effort [2] directed toward providing a general solution to the 'positive realisation problem', it may soon be possible to 'complete the picture' by avoiding the ad-hocness of the final positive formulation suggested here.

7. REFERENCES

[1] H. Akaike. Markovian representation of stochastic processes. *SIAM J. Control*, 13(1):162–173, 1975.

[2] B.D.O. Anderson et al. Nonnegative realization of a linear system. *IEEE Trans. on Cir. Sys.*, 43(2):134–141, 1996.

[3] L.R. Rabiner. A tutorial on hidden Markov models. *Proc. IEEE*, 77(2):257–285, 1989.

[4] P. van Overschee and B.D̃e Moor. *Subspace Identification for Linear Systems.* Kluwer Academic, 1996.