ELSEVIER

Forgetting properties for hidden Markov models

Brian D.O. Anderson[1]*

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

Abstract

Hidden Markov models provide the opportunity to capture a number of nonlinear and/or nongaussian signal processing problems. This paper discusses the existence of results applicable to hidden Markov model filters and fixed-lag smoothers which parallel results applicable to Kalman filters and fixed-lag smoothers, in relation to forgetting of initial conditions, effect of round-off errors, and appropriate choices of the lag for a fixed-lag smoother. Some related problems on maximum a posteriori sequence estimation are also discussed. Tools for addressing these problems are provided by extensions of the Perron–Frobenius theory to inhomogenous products of positive matrices, and inhomogenous matrix products in the so-called max plus algebra.

Zusammenfassung

Hidden Markov Modelle eröffnen die Möglichkeit, eine Anzahl von nichtlinearen und/oder nichtnormalverteilten Signalverarbeitungsproblemen zu erfassen. Dieser Artikel diskutiert vorhandene Ergebnisse, die für hidden-Markov-modell-basierte Filter und Glättungsfilter mit fester Verzögerung anwendbar sind. Diese Ergebnisse ähneln solchen, die für Kalman Filter und Glättungsfilter mit fester Verzögerung anwendbar sind, in Bezug auf das Vergessen der Anfangsbedingungen, die Auswirkungen von Rundungsfehlern und die Wahl der Verzögerung bei Glättungsfiltern. Es werden ebenfalls einige verwandte Probleme der Maximum-a-posteriori-Sequenzschätzung diskutiert. Werkzeuge zur Behandlung dieser Probleme werden durch Erweiterungen der Perron–Frobenius Theorie auf inhomogene Produkte positiver Matrizen und durch inhomogene Matrixprodukte in sogenannten Max Plus Algebren zur Verfügung gestellt.

Résumé

Les modèles de Markov cachés offrnt la possibilité de traiter un grand nombre de problèmes non-linéaires et/ou non-gaussiens en traitement du signal. Cet article décrit l'existence de résultats applicables aux filtres de modèles de Markov cachés et aux adoucisseurs à décalage fixe qui offrent un parallèle à des rèsultats applicables aux filtres de Kalman et aux adoucisseurs à décalage fixe, et aux ceci en relation à l'oubli des conditions initiales, effet des erreurs d'arrondi et des choix appropriés du décalage pour un adoucisseur à décalage fixe. Des problèmes connexes au sujet de l'estimation maximale a posteriori de séquences sont également présentés. Des outils permettant de traiter ces problèmes sont fournis par des extensions de la théorie de Perron–Frobenius aux produits de matrices positives inhomogènes, et aux produits de matrices inhomogènes dans l'algèbre max plus.

[1] The author wishes 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.

* Tel.: + 61-2-6279-8667; fax: + 61-2-6279-8688.

E-mail address: brian.anderson@anu.edu.au (B.D.O. Anderson).

1. Introduction

The purpose of this paper is to highlight certain important questions arising in the filtering and smoothing of signals from hidden Markov models, and to discuss the resolution of these questions in a manner drawing out close parallels with the resolution of the corresponding questions for Kalman filters and smoothers.

While detailed formulation of the questions is left to later sections, we outline the issues here.

In hidden Markov signal models, it is necessary to make some (possibly probabilistic) assumptions about the initial state. These assumptions may well be erroneous, and one would like reassurance that there will be some kind of recovery from the error as time evolves from the initial state. Again, when a hidden Markov model (HMM) filter operates, there will almost certainly arise round-off errors in the computation; one would like assurance that the cumulative effect of such errors will not overpower the calculations to yield totally misleading filter outputs. Clearly, similar issues arise in implementing a Kalman filter. They are resolved in the latter case by postulating that the signal model has certain properties which imply an exponential forgetting property for the filter, [4]. We report recent research in this paper that establishes the same conclusion for hidden Markov models [2,6,7, 10,11,14,16]. The ideas maybe have their roots in [9], which proves an exponential forgetting result as background for an HMM estimation problem.

A second question which we examine relates to fixed-lag smoothing. Fixed-lag smoothing offers an improvement over filtering, in terms of the quality or accuracy of estimates, but at the cost of there being a delay (relative to filtering) – the lag – in the production of the estimates, and more computational burden than for simple filtering. (Exact definitions of fixed-lag smoothed estimates are given in subsequent sections.) It is of interest to understand the trade off between the choice of lag and the improvement offered over filtering. For Kalman filtering, the bigger the lag the bigger the improvement, but the increment in improvement from an increase in lag falls off exponentially with lag. There is a dominant time constant – that of the Kalman filter – such that a four to five times multiple defines the maximum lag it is worthwhile to use, [4]. In this paper, an identical result is obtained for fixed-lag smoothers for HMMs, as reported in [2,16].

The concepts for Kalman filters and smoothers are reviewed in Section 2. Hidden Markov model filters and smoothers are treated in Sections 3 and 4, respectively.

The remaining section deals with a property of HMMs which does not parallel a property for Kalman filters, and indeed one for which research is not yet complete. In Kalman filtering, if a block of measurement data is collected over a time interval, and one seeks to estimate the most probable state trajectory over the same time interval consistent with that data, one can assemble such a trajectory from individual conditional mean smoothed estimates of the state at each time instant in the interval. In contrast, a most probable state trajectory of an HMM is not obtained (in general) by juxtaposing most probable estimates for the state at a sequence of distinct time instants. Different techniques are therefore needed to deal with *trajectory* estimation than those presented in Sections 3 and 4; we present a description of the problem using max-plus algebra ideas. We also discuss the issue of forgetting of initial condition data.

There is an interesting mathematical thread linking Sections 3–5. Matrices of nonnegative elements ("positive matrices") arise naturally in HMM problems; such matrices have special eigenvalue properties, as presented for example in [15], summed up in the so-called Perron–Frobenius theorem. Less well known are the properties of *inhomogenous products* of positive matrices (which are the subject of a sort of generalised Perron–Frobenius theorem). This is appealed to here.

It turns out there is an analog in max-plus algebras of the Perron–Frobenius theorem (termed in [13] "the (max, +) Perron–Frobenius theory"). The forgetting problem for initial condition data requires a max-plus version of the Perron–Frobenius extension to inhomogenous products of positive matrices in conventional matrix algebra; such an extension is not properly available at this point.

Throughout the paper, we consider discrete-time processes with a finite output space. The results

should be generalizable to both continuous-time processes and continuous outputs - indeed some of the earlier work on filtering moves in this direction. The techniques are of course less simple.

2. Kalman filtering and smoothing

In this section, we review certain properties of Kalman filters and smoothers; in the bulk of this paper, we aim to identify similar properties for HMM filters and smoothers. Throughout our presentation, we shall restrict attention to discrete-time systems. Almost all of the ideas remain valid for continuous time systems. A general reference for the material of this section is [4].

2.1. *The signal model*

The signal model we consider is

$$\begin{aligned} x_{k+1} &= F_k x_k + G_k w_k, \\ z_k &= H_k x_k + v_k. \end{aligned} \tag{2.1}$$

In these equations, v_k and w_k represent zero mean Gaussian white noise sequences, usually independent, and with v_k usually having a nonsingular covariance:

$$E[w_k w_\ell^{\mathrm{T}}] = Q_k \delta_{k\ell}, \quad E[v_k v_\ell^{\mathrm{T}}] = R_k \delta_{k\ell}, \quad R_k > 0,$$
$$E[w_k v_\ell^{\mathrm{T}}] = 0, \quad \forall k,\ell. \tag{2.2}$$

Further, F_k, G_k, H_k, Q_k, R_k and R_k^{-1} are assumed bounded. (Often, these quantities are time-invariant.)

In addition, the pairs $[F_k, G_k Q_k^{1/2}]$ and $[F_k, H_k]$ are uniformly completely stabilisable and detectable, respectively (see also [5]).

At the initial time k_0 (which may be $-\infty$), x_{k_0} is presumed to be a gaussian random vector with mean $\bar{x}_{k_0}$ and variance P_0, and independent of the noise sequences.

2.2. *The Kalman filter*

The Kalman filter associated with the above signal model is given by

$$\begin{aligned} \hat{x}_{k+1/k} &= (F_k - K_k H_k)\hat{x}_{k/k-1} + K_k z_k \\ \hat{x}_{k_0/k_{0-1}} &= \bar{x}_{k_0}. \end{aligned} \tag{2.3}$$

In this equation, $\hat{x}_{k+1/k}$ denotes $E[x_{k+1}/z_k, z_{k-1}, \ldots, z_{k_0}]$ and K_k is the Kalman gain. It is defined as follows. Recursively solve

$$\Sigma_{k+1/k} = (F_k - K_k H_k)\Sigma_{k/k-1}(F_k - K_k H_k)^{\mathrm{T}} + G_k Q_k G_k^{\mathrm{T}} + K_k R_k K_k^{\mathrm{T}}$$
$$\Sigma_{k_0/k_{0-1}} = P_0, \tag{2.4}$$

$$K_k = F_k \Sigma_{k/k-1} H_k (H_k^{\mathrm{T}} \Sigma_{k/k-1} H_k + R_k)^{-1}. \tag{2.5}$$

The matrix $\Sigma_{k+1/k}$ has the additional interpretation

$$\Sigma_{k+1/k} = E\{[x_{k+1} - \hat{x}_{k+1/k}] \times [x_{k+1} - \hat{x}_{k+1/k}]^{\mathrm{T}}\}. \tag{2.6}$$

2.3. *Forgetting-type properties*

Under the assumptions on the signal model formulated above, there are three key forgetting properties. They all stem from the one nontrivial consequence of the assumptions, that the homogenous equation

$$\lambda_{k+1} = (F_k - K_k H_k)\lambda_k \tag{2.7}$$

is exponentially asymptotically stable. The three properties are as follows:

- *initial state forgetting.* $\hat{x}_{k+1/k}$ depends in a manner decaying exponentially with $k - k_0$ on the assumed value of $\bar{x}_{k_0}$. Incorrect initialisation therefore is not a problem, since it is forgotten exponentially fast.
- *round-off and similar errors can only accumulate to a bounded extent.* At each step of the filter computation, round-off errors can be introduced. It is important that the cumulative effect of all past round-off errors not overpower the calculations, but in some way be limited; exponential stability of (2.7) ensures that the effect is bounded (but could be large)
- *forgetting of old measurements.* Just as the initial condition is forgotten exponentially fast, so too are old measurements. This means that $\hat{x}_{k+k_1-1}$ depends on z_{k_1} in $O(\alpha^k)$ fashion, for some $\alpha < 1$.

Note that it is possible to contemplate Kalman filtering problems in which the conditions ensuring exponential stability of (2.7) are not fulfilled. See [4], Section 6.1 and [1].

Whenever there is exponential convergence, there can be interest in simple computations that give information about the rate. We simply make one point here, and for simplicity, in the time-invariant case only. One can show that

$$|\det(F - KH)| \leqslant |\det F|.$$

(The equality sign is exceptional.) Crudely, this says that the Kalman filter is at least as stable as the signal model.

2.4. *Interpretation of* $\hat{x}_{k+1/k}$

We have earlier indicated that

$$\hat{x}_{k+1/k} = E[x_{k+1} \mid z_k, \ldots, z_{k_0}]. \tag{2.8}$$

As a conditional mean estimate, $\hat{x}_{k+1/k}$ also is the estimate minimizing the conditional error variance. (Because of gaussianity, the conditional error variance and unconditioned error variance are the same.)

Because of gaussianity, $\hat{x}_{k+1/k}$ is also a maximum a posteriori estimate, i.e. $\arg\max_{x_{k+1}} p(x_{k+1} \mid z_k, \ldots, z_0)$.

Finally, let $x^*_{k_0}, x^*_{k_0+1}, \ldots, x^*_{k-1}, x^*_k$ be the most probable trajectory given $z_j, j < k$. Then (non-trivially)

$$x^*_k = \hat{x}_{k/k-1}. \tag{2.9}$$

2.5. *Smoothing*

The Kalman filtering equations above provide an estimate of x_k given measurements up to time $k-1$. One can conceive of collecting more measurements, say up to time $k + \Delta$, and then estimating x_k, via an estimate

$$\hat{x}_{k/k+\Delta} = E[x_k \mid z_{k_0}, z_{k_0+1}, \ldots, z_{k+\Delta}]. \tag{2.10}$$

The disadvantage is that one must wait longer for the estimate to become available, (till time $k + \Delta$ plus the computation time); this may be unacceptable if the estimate is being used for control (or decision making where rapid response is important). However, the advantage is that with more measurements, a more accurate estimate can be anticipated.

The equations for obtaining smoothed estimates are set out in [4]; we will not reproduce them here, but rather indicate several properties of smoothing.

When Δ is fixed and k is variable, we talk of a fixed-lag smoother with lag Δ. It turns out that

- the bigger Δ is, the smaller is $E\{[\hat{x}_{k/k+\Delta} - x_k][\hat{x}_{k/k+\Delta} - x_k]^T\}$, the error covariance of the smoothed estimate,
- for fixed k as Δ increases, $E\{[\hat{x}_{k/k+\Delta} - x_k][\hat{x}_{k/k+\Delta} - x_k]^T\}$ falls off exponentially towards some limiting value, indicating that for practical purposes, there is some $\Delta_{\max}$ such that $\Delta = \Delta_{\max}$ gives all the practical benefit to be had from smoothing, and taking $\Delta > \Delta_{\max}$ confers no extra benefit,
- the exponential decay rate referred to above is *the same as that associated with the Kalman filter*; i.e. if the Kalman filter forgets its initial state with a time constant of 1 s, a choice for Δ of 4 or 5 s will extract all the benefit from smoothing which it is possible to extract.

What sort of improvement can smoothing offer over filtering? It is hard to give a universal answer to this question. However, it appears that the higher is the signal-to-noise ration, the greater is the percentage improvement offered by smoothing over filtering [3].

As for filtering, see (2.9), one can connect smoothed estimates at single instants of terms with complete trajectories. Let $x^*_{k_0}, \ldots, x^*_{k+\Delta}$ be the most probable trajectory given z_j, $k_0 \leqslant j \leqslant k + \Delta$. Then

$$x^*_k = \hat{x}_{k/k+\Delta}$$

and

$$\hat{x}_{k/k+\Delta} = \underset{x_k}{\operatorname{argmax}}\ p(x_k / z_{k_0}, \ldots, z_{k+\Delta}).$$

3. Hidden Markov models and filtering

In this section, we shall define a signal model (HMM) and the associated filter. We will also establish an exponential forgetting property paralleling that for Kalman filters. We follow [16].

3.1. Signal model

Consider a first-order discrete-time and discrete-state Markov process X_k, the subscript k denoting time.

For simplicity, we shall define the states to be the values $1, 2, \ldots, N$, and assume the process is stationary. At each time instant k, a corresponding signal Y_k is observed, again having discrete values, in the range $1, 2, \ldots, M$. We will adopt the convention that a lower-case x_k denotes the actual state value, and likewise for y_k. The probability vectors for X_k and Y_k are updated by the constant system matrices A (the state transition probability matrix) and C (observation matrix), with $A = \{a_{ij}\} = \{\Pr(X_{k+1} = i \mid X_k = j)\}$, and $C = \{c_{mn}\} = \{\Pr(Y_k = m \mid X_k = n)\}$. Further, unless otherwise stated, $a_{ij} > 0$ and $c_{mn} > 0$, $\forall i, j, n \in \{1, 2, \ldots, N\}$, $\forall m \in \{1, 2, \ldots, M\}$.

Remark 1. In the present definition, A and C are column-stochastic matrices, i.e. $\sum_{i=1}^{N} a_{ij} = 1$ and $\sum_{m=1}^{M} c_{mn} = 1$. This may be contrary to other notations and must be kept in mind to avoid possible confusion.

Remark 2. For the case of independent and identically distributed (iid) processes, in which all state transitions are equally likely, A has identical entries in each row and consequently columns in A are identical. For a similar situation involving the C matrix, the interpretation is that the output process is independent of the state process.

3.2. Evolution of filtered distributions

Let $\Pi_{k|k}$ and $\Pi_{k+1|k}$ be the filtered and one-step prediction probability vectors, with the ith entry being $\Pr(X_k = i \mid Y_0, Y_1, \ldots, Y_k)$ and $\Pr(X_{k+1} = i \mid Y_0, Y_1, \ldots, Y_k)$, respectively. By using a combination of Bayes' rule of conditional probability and the Markov property, the time evolution relations for the filtered probability vector are found to be

$$\Pi_{k+1|k} = A\Pi_{k|k} \tag{3.1}$$

$$\Pi_{k+1|k+1} = \frac{1}{1_N^{\mathrm{T}} C_{y_{k+1}} \Pi_{k+1|k}} C_{y_{k+1}} \Pi_{k+1|k} \tag{3.2}$$

where $1_N^{\mathrm{T}} C_{y_{k+1}} \Pi_{k+1|k} = [1 \ldots 1] C_{y_{k+1}} \Pi_{k+1|k}$ is a scaling normalising constant to ensure the entries of $\Pi_{k+1|k+1}$ sum to 1, and $C_{y_{k+1}} = \mathrm{diag}(c_{l1}, c_{l2}, c_{l3}, \ldots, c_{lN})$, when $y_{k+1} = l$.

3.3. Exponential forgetting

By iterating (3.1) and (3.2), the filtered probability vector at time k can be expressed in terms of an arbitrarily chosen initial distribution $\Pi_{0|0}$:

$$\Pi_{k|k} = \frac{1}{1_N^{\mathrm{T}} C_{y_k} A \Pi_{k-1|k-1}} C_{y_k} A \Pi_{k-1|k-1}$$

$$= \frac{1}{1_N^{\mathrm{T}} C_{y_k} A C_{y_{k-1}} A \ldots C_{y_1} A \Pi_{0|0}} (C_{y_k} A C_{y_{k-1}} A \ldots$$

$$C_{y_1} A) \Pi_{0|0} = \frac{1}{1_N^{\mathrm{T}} T_{1,k} \Pi_{0|0}} T_{1,k} \Pi_{0|0} \tag{3.3}$$

where

$$T_{1,k} = C_{y_k} A C_{y_{k-1}} A \ldots C_{y_1} A. \tag{3.4}$$

We now proceed to derive initial-condition forgetting of the filter by appealing to the generalised Perron–Frobenius result (see Appendix A, in particular Corollaries A.1 and A.2 and also [15]) for an inhomogeneous product of matrices. Broadly, this theorem states that, under certain very general conditions, a product of positive matrices, of the form $T_{1,r} = H_r H_{r-1} \ldots H_1$, tends to a dyadic matrix[2] as $r \to \infty$.

As stated previously, $A > 0$ and $C > 0$; therefore (3.4), a product of successive $C_y A$, is strictly positive, and the requirements of the aforementioned theorem are automatically satisfied. This means that, as $k \to \infty$,

$$T_{1,k} \to U(k) V^{\mathrm{T}} \tag{3.5}$$

for some positive column vector $U(k)$, and positive row vector V^{T}. Without loss of generality, let $v_1 = 1$. Hence the normalised filter probability vector becomes

$$\Pi_{k|k} = \frac{1}{1_N^{\mathrm{T}} T_{1,k} \Pi_{0|0}} T_{1,k} \Pi_{0|0}$$

[2] The term dyadic means that a matrix is of rank 1.

$$\rightarrow \frac{1}{1_N^T T_{1,k} \Pi_{0|0}} \begin{bmatrix} u_1(k) \\ u_2(k) \\ \vdots \\ u_n(k) \end{bmatrix} (1\, v_2\, v_3 \ldots v_n) \Pi_{0|0}$$

$$\rightarrow \frac{1}{\sum_{i=1}^N u_i(k)} U(k). \tag{3.6}$$

This is independent of the initial distribution $\Pi_{0|0}$. The rate of convergence of (3.6) is exponential and rates are computable [15].

3.4. *Relaxing the positivity constraints on A and C*

In the above analysis we have assumed that A and C have all positive entries. We can actually relax this constraint to one requiring a form of observability, [16].

In mathematical terms the exponential convergence to a dyad remains if we require that there exists an L such that for any selection $y_1, \ldots, y_L$, each row of the matrix

$$T_{1,L} = C_{y_L} A C_{y_{L-1}} A \ldots C_{y_1} A$$

is either entirely zero, or entirely nonzero. This means that at time L, each possible state $x_L = i$ is either such that it could have risen from $y_L, y_{L-1}, \ldots, y_1$ and *any* state $x_0 = j, j = 1, \ldots, N$ or $x_L = i$ is incompatible with the sequence $y_L, y_{L-1}, \ldots y_1$ and all possible $x_0 = j$. Put another way, whether or not one can be in state i at time L depends only on the last L measurements, and has nothing to do with the state at time 0.

Obviously, if this property holds, it also holds with L replaced by any $L' > L$.

Two important special cases are as follows.

Suppose A is a primitive matrix, which is equivalent to having the underlying state process irreducible and aperiodic. Suppose further that $C > 0$ (a given measurement can come from any state). Then $T_{1,L}$ will be positive, where L is the least integer for which A^L is positive.

Second, suppose A is positive, but C is not necessarily positive. Then $T_{1,L}$ will have the prescribed property for all L.

3.5. *Other forgetting properties*

It is clear that in addition to exponential forgetting of an initial condition, there is exponential forgetting of a particular past measurement, and of a particular round-off or other computational error.

3.6. *Conditional mean and maximum a posteriori estimation*

For the formulation we have given, there is not much point in talking about a conditional mean for the state. The maximum a posteriori state is however

$$\operatorname*{argmax}_i \{p(x_k = i \mid y_0, \ldots, y_k)\},$$

i.e. the index of the maximum entry of $\Pi_{k|k}$.

Suppose that the effective forgetting time for initial conditions is L. Suppose also that the vector $\Pi_{L|L}$ for every sequence of measurements of length L has a unique maximum entry. Since the effective forgetting time for initial conditions is L, this means that the index of the maximum entry in $\Pi_{k|k}$ for all $k > L$ depends simply on $y_{k-L+1}, \ldots, y_k$ and no earlier measurements, i.e. *as far as MAP estimation is concerned, there is finite-time forgetting of initial conditions.*

In contrast to the Kalman filtering situation, if a maximum a posteriori trajectory $x_0^*, \ldots x_k^*$ is obtained, given measurements $y_0, \ldots, y_k$, there is no guarantee that

$$x_k^* = \operatorname*{argmax}_i \{p(x_k = i \mid y_0, \ldots, y_k)\}.$$

4. Hidden Markov model smoothing

For the hidden Markov model of the previous section, we can conceive of solving a smoothing problem, i.e. obtaining recursively the vector $\Pi_{k|k+\Delta}$, the ith entry of which is

$$p(x_k = i \mid y_0, y_1, \ldots, y_{k+\Delta}).$$

Fixed-lag smoothing corresponds to Δ fixed and k varying, and, as for Kalman filtering, there is the disadvantage of delay in computation but the

advantage of more accurate determination of x_k (or its conditional probability vector). Ref. [16] develops the relevant equations.

4.1. *Exponential decay in smoothing*

Let Z be a matrix with N^2 rows and N columns with zeros in all rows except row 1, $N+1$, $2N+1$ and $(N-1)N+1$. In these rows, there are all zeros except for a unity entry in columns $1, 2, \ldots, N$, respectively. Then the unnormalised equation for $\Pi_{k|k+\Delta}$ turns out to be

$$\tilde{\Pi}_{k|k+\Delta} = [I_N \otimes (1_N^{\mathrm{T}} T_{k+1|k+\Delta})] Z \Pi_{k|k}, \tag{4.1}$$

where

$$T_{k+1|k+\Delta} = C_{y_{k+\Delta}} A C_{y_{k+\Delta-1}} A \ldots C_{y_k} A \tag{4.2}$$

and the vector $\tilde{\Pi}_{k|k+\Delta}$ is a scalar multiple of $\Pi_{k|k+\Delta}$. For fixed k and varying Δ, one can consider how (4.1) behaves. The key is of course (4.2), and the essence of the matter is that $T_{k+1|k+\Delta}$ for Δ varying inherits the convergence properties of $T_{1,k}$ for k varying, which were relevant in considering HMM filtering. For k fixed, as $\Delta \to \infty$ exponential convergence occurs:

$$T_{k+1|k+\Delta} \to U(\Delta) V^{\mathrm{T}}$$

for some fixed row vector V^{T}; then

$$\begin{aligned} \tilde{\Pi}_{k|k+\Delta} &\to [I_N \otimes (\Sigma u_i(\Delta)) V^{\mathrm{T}}] Z \Pi_{k|k} \\ &= [\Sigma u_i(\Delta)] (\mathrm{Diag}\, v_i) \Pi_{k|k} \end{aligned}$$

and on normalising, the Δ dependent term, viz. $\Sigma u_i(\Delta)$, disappears. This means that all the improvement that can be gained by smoothing is attained after some finite Δ, and no significant gains can be made as the lag is extended further.

4.2. *Example*

Some examples of the above are in [16]. Here, we want to recall a much earlier paper [12], more than 20 years old. In that paper, the hidden Markov model was actually a continuous-time model, with a continuous measurement set. The state process was defined by a random telegraph signal (switching between $+1$ and -1 according to a Poisson law), with measurements comprising the state contaminated by gaussian white noise.

The paper developed smoothing equations, and demonstrates with simulations the improvement flowing from selecting Δ away from zero, noting the exponential decay of the rate of improvement. However, no theory could be given to explain the results.

Reproduced as Fig. 1 are simulation results reported originally in [12] which confirm strikingly the improvement that smoothing offers over filtering (lag = 0), and also display the exponential dependence of the behavior on the smoothing lag Δ.

The results are actually obtained by running a discrete-time approximation of the continuous time smoother on noisy measurements of a continuous time random telegraph wave. Transitions between the two states in the signal model are governed by a Poisson process and occur on average every 83 time units. The signal-to-noise ratio is fairly high, and with zero lag, the filter gives an error rate (for MAP estimation) of 45 in 1000. By allowing a lag of about 30 time

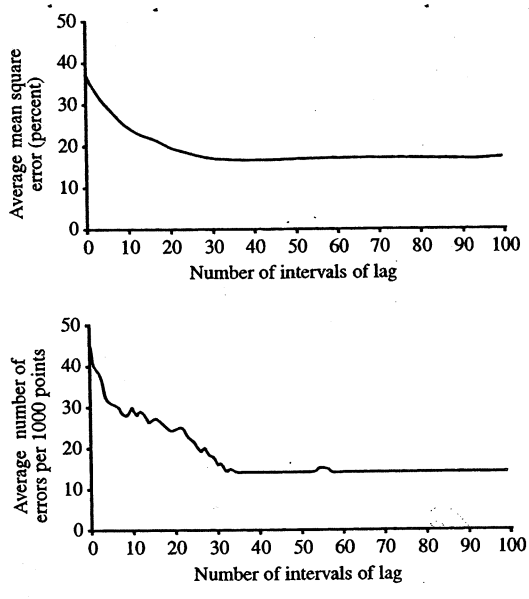

Fig. 1. Plots of mean-square error and error rate against smoothing lag for MAP smoothed state estimate from a discrete-time filter of a continuous time random telegraph wave.

units (less than the time constant for the underlying process), it is possible to gain all the possible improvement over filtering; this improvement, yielding an error rate of about 15 in 1000, is very substantial.

Ref. [12] also contains simulation results demonstrating that for lower signal-to-noise ratios, the improvement offered by fixed-lag smoothing over filtering is not so great. (This is also the case for Kalman filtering and smoothing). No quantitative explanation of this has however yet been offered in the HMM case.

4.3. Maximum a posteriori estimation

In the light of the HMM filtering results it is no surprise that if $x_0^*, \dots, x_{k+\Delta}^*$ denotes the most probable state trajectory given the measurements $y_0, \dots, y_{k+\Delta}$, then in general there is no guarantee that

$$x_k^* = \underset{i}{\operatorname{argmax}} \{p(x_k = i \mid y_0, y_1, \dots, y_{k+\Delta})\}.$$

5. Maximum a posteriori trajectories

5.1. MAP trajectories and max plus algebra

In Kalman filtering, the solutions of the filtering and smoothing problems yield (individual points on) maximum a posteriori trajectories. This is not so in the HMM context. In this section, we shall present the MAP trajectory determination problem using max plus algebra ideas (see [8] for an introduction to max plus algebra – this reference also describes the MAP trajectory determination problem in max plus algebra terms).

We retain the notation of Sections 3 and 4. Then

$$\Pr[x_0 = j_0, x_1 = j_1, \dots, x_k = j_k, y_0 = i_0, \dots, y_k = i_k] = \sum_{p=1}^{k} c_{i_p j_p} a_{j_p j_{p-1}} \Pr(x_0 = j_0). \tag{5.1}$$

Let L denote the logarithm of the above joint probability. Define matrices B_γ, and vectors b_γ, $\gamma = 1, 2, \dots, M$, by

$$B_\gamma(\alpha,\beta) = \ln[c_{\gamma\alpha} a_{\alpha\beta}], \tag{5.2}$$

$$b_\gamma(\alpha) = \ln[c_{\gamma\alpha} \Pr(x_0 = \alpha)]. \tag{5.3}$$

Then

$$L = B_{i_k}(j_k, j_{k-1}) + B_{i_{k-1}}(j_{k-1}, j_{k-2}) + \cdots + b_{i_0}(j_0). \tag{5.4}$$

Estimating the MAP trajectory over $[0, k]$ is a matter of finding, for a given set $i_0, \dots, i_k$, the set $j_0^*, \dots, j_k^*$ which maximizes (5.1) or equivalently (5.4).

Based on standard concepts of max plus algebra, see Appendix B, there holds

$$\max_{j_k, \dots, j_0} L = 1^{\mathrm{T}} \otimes B_{i_k} \otimes \cdots \otimes B_{i_1} \otimes b_{i_0}. \tag{5.5}$$

If matrices A and C are positive, all entries of the B_γ and b_γ are finite. However, if some entries of A and C are zeros, then certain entries of B_γ or b_γ are $-\infty$ (which is allowable in max plus algebra; in the associated transition graph, a weight of $-\infty$ corresponds to absence of an arc).

The Viterbi algorithm, [17] conventionally used for MAP trajectory determination, is in effect a device for calculating very efficiently the max plus product; in the course of the calculation, the maximizing sequence $j_0^*, \dots, j_k^*$ is computed. The maximizing sequence identifies a particular path in the transition graph for the right-hand side of (5.4), viz. with maximum weight end-to-end.

5.2. Forgetting

We would like to obtain an analog of the forgetting concept of Section 3. Recall that in Section 3, we established an exponential forgetting of initial condition property of $\Pi_{k/k}$, while under certain conditions

$$\underset{i}{\operatorname{argmax}} [p(x_k = i / y_0, \dots, y_k)],$$

might depend only on the most recent L measurement values $y_{k-L+1}, \dots, y_k$. This is a form of finite memory.

For MAP trajectories, the analogous question can be posed as follows. Suppose k is large, and we have somehow obtained the MAP trajectory

$j_0^*, \ldots, j_k^*$. Let P be fixed, and we want to focus on $j_{k-P+1}^*, \ldots, j_k^*$, i.e. the $P+1$ last points in the MAP trajectory. Is it the case that for some Q, these P last points depend only on the $P+Q$ last measurements, viz. $i_k, i_{k-1}, \ldots, i_{k-P-Q+1}$?

Evidently, the examination of the MAP state in Section 3 amounts to considering just $P = 1$.

We can state the following (new) result:

Theorem 5.1. *With the HMM defined by matrices A and C as earlier described, and the matrices B_γ and b_γ as defined in* (5.2) *and* (5.3), *suppose that any product of length Q of the B_{i_α} is a dyad (in the max plus algebra sense). Suppose that $k > P + Q$. Then the section $(j_{K-P+1}^*, \ldots, j_K^*)$ of the maximum a posteriori trajectory depends just on the values $i_{k-P-Q+1}, \ldots, i_K$ of the measurements over time $k - P - Q + 1$ to k.*

Proof. Observe that

$$1^T \otimes B_{i_k} \otimes \cdots \otimes B_{i_{k-P+1}} \otimes B_{i_{k-P}} \otimes \cdots \otimes B_{i_{k-P-Q+1}} \otimes B_{i_{k-P-Q}} \otimes \cdots \otimes b_{i_0}$$
$$= (1^T \otimes B_{i_k} \otimes \cdots \otimes B_{i_{k-P+1}}) \otimes c \otimes d^T \otimes (B_{i_{k-P-Q}} \otimes \cdots \otimes b_{i_0}),$$

where c and d are vectors depending on $i_{k-P-Q+1}$ to i_{k-P}, existing by hypothesis).

$$= (1^T \otimes B_{i_k} \otimes \cdots \otimes B_{i_{k-P+1}} \otimes c) \otimes \alpha.$$

Here α is a scalar (viz. $d^T \otimes B_{i_{k-P-Q}} \otimes \cdots \otimes b_{i_0}$).

The path in the transition graph corresponding to $1^T \otimes B_{i_k} \otimes \cdots \otimes B_{i_{k-P+1}} \otimes c$ which has maximum weight is independent of the scalar α.

This path, $(j_{K-P}^*, \ldots, j_k^*)$ is such that

$$1^T \otimes B_{i_k} \otimes \cdots \otimes B_{i_{k-P+1}} \otimes c$$
$$= B_{i_k}(j_k^*, j_{k-1}^*) + \cdots + B_{i_{k-P+1}}(j_{k-P+1}^*, j_{k-P}^*) + c(j_{k-P}^*)$$
$$= \max_{j_{k-P}, \ldots, j_k} \{B_{i_k}(j_k, j_{k-1}) + \cdots + B_{i_{k-P+1}}(j_{k-P+1}, j_{K-P}) + c(j_{k-P}\}.$$

The matrices and vector inside $\{\ldots\}$ are dependent on measurements $i_{k-P-Q+1}$ to i_{k-P} (for $c(.)$) and $i_{k-P+1}, \ldots, i_k$ for the B matrices. Accordingly, the section $j_{k-P}^*, \ldots, j_k^*$ of the maximizing trajectory depends only on these same measurements.

The above theorem naturally throws up the following question. Suppose there is prescribed a finite set of $N \times N$ matrices, viz. $\{B_1, \ldots, B_M\}$. Under what circumstances does there exist an integer q such that the (max plus algebra) product of any selection of q of the B_i is expressible as a dyad. (Note: if such a condition holds, it is trivial that for any $q' > q$, a product of any selection of q' of the B_i will also be expressible as a dyad).

An exact answer is not known. However, there are several pertinent remarks which should be made.

(a) The Perron–Frobenius theory for a positive and some nonnegative square matrices A guarantees that as $p \to \infty$, $A^p \to \lambda^p u v^T$ for some positive real λ and positive vectors u and v, with the convergence exponentially fast. Analogously, max plus algebra guarantees that as $p \to \infty$, for matrices A satisfying a max plus version of the primitivity conditions for nonnegative matrices $A^p \to \lambda^p P$ where P is a sum of dyads, and often comprises just one dyad; further, the convergence occurs in a finite number of steps, i.e. there exists Q such that for all $p \geqslant Q$, $A^p = \lambda^p P$.

(b) The Perron–Frobenius theory, as explained in Appendix A, generalises to cope with inhomogeneous products of nonnegative matrices, so that a form of exponentially fast convergence to rank one occurs.

In the light of (a) and (b), it is reasonable to postulate a form of *max plus algebra* Perron–Frobenius theorem for *inhomogeneous* products. Such a result is exactly what is needed for the hypothesis of the theorem to be fulfilled.

Simulation data suggest the reasonableness of the result. However, just as there are exceptions to the Perron–Frobenius result (which are dealt with by imposing requirements of for example irreducibility or primitivity, to exclude the awkward cases), so is this true in the max plus algebra situation.

The successful operation of a Viterbi decoder rests on the presumption that the conclusion of the main Theorem is valid, at least under normal operation, or for almost all symbol sequences. This

raises the issue that there may be a probabilistic version of the result needed to underpin the theorem. It might be something like the following.

Conjecture. *Consider a set of M generic square matrices $B_1 \ldots, B_M$ and consider max plus algebra products Π_q of length q of any of the B_i. Then there exist positive constants α and β with $\beta < 1$ such that*

$$\Pr\{\Pi_q \text{ is expressible as a dyad}\} = 1 - \alpha\beta^q.$$

An important additional task would be to pin down the convergence times (effectively Q in the theorem and (roughly) β^{-1} in the conjecture); these should underbound the length of a Viterbi decoder. It would be comforting if these convergence times could be related to the time constants of the signal model (as is roughly possible for the Kalman filter and indeed the normal HMM filter, when their forgetting properties are being considered).

6. Conclusion

In this paper, we have indicated a number of parallels between Kalman filtering and smoothing results on the one hand and HMM filtering and smoothing on the other. These results are linked to forgetting of initial condition data, coping with round-off error, and choice of a lag for a fixed-lag smoother.

We have also indicated open issues. The two big ones are as follows:

- Can one establish for HMMs that the benefit offered by smoothing over filtering is greater at higher SNRs?
- Can one establish a deterministic or stochastic max plus algebra version of the Perron–Frobenius theory for inhomogeneous matrix products, from which would flow a forgetting result for MAP trajectory estimation in HMMs, and thereby justify the use of finite length Viterbi decoder?

Of course, for both these problems, quantitative or rather than just qualitative conclusions would be welcome. The first question is addressed in a preprint available from the authors.

Appendix A. Generalisations of the Perron–Frobenius theorem for inhomogeneous product of matrices

In this appendix, we summarise certain key ideas relating to products of nonnegative matrices and positive matrices from [13].

Definition A.1. An $(n \times n)$ matrix $T \geqslant 0$ is said to be row-allowable if it has at least one positive entry in each row. It is said to be column-allowable if T' is row-allowable. It is said to be allowable if it is both column- and row-allowable.

Remark 3. A column-allowable T has the property that $x > 0$ implies $x'T > 0$.

Definition A.2. For two vectors $x' = (x_1 \ldots x_n) > 0$ and $y' = (y_1 \ldots y_n) > 0$, a pseudo-metric[3] can be defined as

$$d(x', y') = \ln\left[\frac{\max_i (x_i/y_i)}{\min_i (x_i/y_i)}\right] = \max_{i,j} \ln\left[\frac{x_i y_j}{x_j y_i}\right]. \quad \text{(A.1)}$$

This is a measure of the alignment between two given vectors, x and y, and $d(x', y') = 0$ iff $x = \lambda y$, for some scalar $\lambda > 0$.

For $x, y > 0$ and column-allowable T (to ensure $x'T > 0$ and $y'T > 0$) the function $d(.,.)$ has the following contraction properties:

(1) $d(x'T, y'T) \leqslant d(x', y')$.

(2) Given a T which also has at least one positive element in a coincident position in any two rows, $d(x'T, y'T) < d(x', y')$. This guarantees that for $x, y > 0$, multiplication by a strictly positive T always tends to align the two vectors x and y.

Remark 4. A direct consequence of above properties is that, under certain conditions as $k \to \infty$,

$$d(x'T_1 T_2 \ldots T_k, y'T_1 T_2 \ldots T_k) \to 0,$$

[3] A pseudo-metric has the properties of a metric save that $d(x', y') = 0$ can occur even if $x \neq y$.

where each T_i is allowable, $1 \leqslant i \leqslant k$, provided $x, y > 0$.

Definition A.3. The Birkhoff's contraction coefficient is defined as

$$\tau_B(T) = \sup_{x,y>0, x \neq \lambda y} \frac{d(x'T, y'T)}{d(x', y')}. \tag{A.2}$$

Birkhoff's contraction coefficient places an upper bound on the rate of contraction due to the multiplication with an allowable matrix T, by providing a ratio between the pre- and post-multiplied metric of two vectors by T. As will be shown below, the case when $\tau_B(.) = 0$ is of special interest.

Remark 5. For an allowable matrix T, an explicit form (for the long derivations, see [13]) for $\tau_B(T)$ is

$$\tau_B(T) = \frac{1 - \sqrt{\phi(T)}}{1 + \sqrt{\phi(T)}}, \tag{A.3}$$

where

$$\phi(T) = \begin{cases} \min_{i,j,k,l} \frac{t_{ik} t_{jl}}{t_{jk} t_{il}} & \text{if} \quad T > 0, \\ 0 & \text{if} \quad T \not> 0. \end{cases} \tag{A.4}$$

If T is column-allowable but not row-allowable, let $A = \{a_{ij}\}$ be the matrix formed by deleting any row of zeros, so that A is row-allowable, then $\phi(A) = \phi(T)$.

For column-allowable T and U, $\tau_B(.)$ has the following properties,

(1) $0 \leqslant \tau_B(T) \leqslant 1$.
(2) $\tau_B(TU) \leqslant \tau_B(T)\tau_B(U)$, hence $\tau_B(TU) \leqslant \tau_B(T)$ and $\tau_B(TU) \leqslant \tau_B(U)$.
(3) For a positive diagonal matrix U, $\tau_B(U) = 1$.
(4) $\tau_B(U) = 0$ iff U is also rank 1.

Definition A.4. The products $T_{p,r} = H_{p+1} H_{p+2} \cdots H_{p+r}$, formed from allowable matrices $\{H_k\}$ in some specified order for $p \geqslant 0$, $r \geqslant 1$, are said to be weakly ergodic if, as $r \to \infty$, $T_{p,r}$ approaches a rank 1 matrix, or equivalently, as $r \to \infty$, $\tau_B(T_{p,r}) \to 0$.

With the above definitions in mind, the next theorem follows naturally.

Theorem A.1 (Seneta). *If* $T_{p,r} = H_{p+1} H_{p+2} \cdots H_{p+r}$, *and all* H_k *are nonnegative and allowable, then* $\tau_B(T_{p,r}) \to 0$ (*i.e. weakly ergodic*) *for each* $p \geqslant 0$, *iff the following condition holds*:

$$T_{p,r} > 0 \quad \text{for all } r \geqslant r_0(p).$$

Further, there then holds:

$$\frac{t_{ik}^{(p,r)}}{t_{jk}^{(p,r)}} \to W_{i,j}^{(p)} > 0 \quad \text{as } r \to \infty$$

for all i, j, p, k where the limit is independent of k, i.e. the rows of $T_{p,r}$ *tend to proportionality as* $r \to \infty$. *Equivalently,* $T_{p,r} \to UV^T(r)$, *where U is a constant vector, and* $V^T(r)$ *some positive row vector.*

Corollary A.1. *For a reverse product, of the form* $T_{p,r} = H_{p+r} H_{p+r-1} \cdots H_{p+1} = (H_{p+1} H_{p+2} \cdots H_{p+r})^T$, *the same result holds except that as* $r \to \infty$ *the columns tend to proportionality, i.e.* $T_{p,r} \to U(r)V^T$, *where U and* V^T *have the same meaning as in Theorem* A.1.

Corollary A.2. *Consider a sequence of positive integers* $\{k_s\}$, $s \geqslant 0$ *and* $k_{s+1} - k_s = g$, *g a constant. If we redefine product* $T_{p,r}$ *by grouping g matrices* $\{H_k\}$ *into composite terms, so that* $T_{p,r} = H_{p+1} H_{p+2} \cdots H_{p+r} = T_{p,k_0 - p} T_{k_0, k_1 - k_0} T_{k_1, k_2 - k_1} \cdots T_{k_{t-1}, k_t - k_{t-1}} T^*$, *for some allowable* T^* *where* k_t *is the nearest member of* $\{k_s\}$ *not greater than r, and if further,*

$$\phi(T_{k_s, k_{s+1} - k_s}) \geqslant \varepsilon^2,$$

then the rate at which $\tau_B(T_{p,r}) \to 0$ *as* $r \to \infty$ (*see* [13]) *can be overbounded, e.g. in the case of* $p = 0$, *as follows*:

$$\tau_B(T_{0,r}) \leqslant \left(\frac{1-\varepsilon}{1+\varepsilon}\right)^{-(k_0/g)-1} \left(\frac{1-\varepsilon}{1+\varepsilon}\right)^{r/g}. \tag{A.5}$$

Appendix B. Max plus algebra

Definition B.1 (Max plus algebra). The max plus algebra $(\mathbb{R}_{\max}, \oplus, \otimes)$ is defined as follows:

(1) $\mathbb{R}_{\max} \overset{\text{def}}{=} \mathbb{R} \cup \{-\infty\}$, where $\mathbb{R}$ is the set of real numbers,

(2) $\oplus$ is maximisation in conventional algebra, so that $x \oplus y \overset{\text{def}}{=} \max(x, y)$, and

(3) $\otimes$ is conventional addition, where $x \otimes y \overset{\text{def}}{=} x + y$.

The notation was chosen so that a number of results from conventional linear algebra can be directly transferred to max plus algebra by replacing the $+$ and $\times$ signs by $\oplus$ and $\otimes$, respectively. In particular, $\oplus$ is commutative over $\mathbb{R}_{\max}$, and $\otimes$ is distributive over $\oplus$. The symbol ε, with the numerical value of $-\infty$, is the neutral element with respect to maximisation; similarly, the symbol e, with the numerical value 0, denotes the neutral element with respect to addition. That is, $\forall x \in \mathbb{R}$:

- $x \oplus \varepsilon = x$, and
- $x \otimes e = x$.

For suitably dimensioned matrices $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$, the following operations are defined:

Definition B.2 (Scalar multiplication). The multiplication of a matrix A and a scalar c is

$$(c \otimes A)_{ij} = c \otimes a_{ij} = c + a_{ij}. \tag{B.1}$$

Definition B.3 (Matrix sum). The matrix sum $A \oplus B$ is defined to be

$$(A \oplus B)_{ij} = a_{ij} \oplus b_{ij} = \max(a_{ij}, b_{ij}). \tag{B.2}$$

Definition B.4 (Matrix product). The matrix product $A \otimes B$ is defined by

$$\begin{aligned}(A \oplus B)_{ij} &= (a_{i1} \otimes b_{1j}) \oplus (a_{i2} \otimes b_{2j}) \oplus (a_{i3} \otimes b_{3j}) \oplus \cdots \\ &\quad \oplus (a_{in} \otimes b_{nj}) \\ &= \bigoplus_k a_{ik} \otimes b_{kj} \\ &= \max_{k=1,2,\ldots,n} (a_{ik} + b_{kj}).\end{aligned} \tag{B.3}$$

Definition B.5 (Rank of a matrix). The rank of a matrix Z is the smallest number of dyads $U_i \otimes V_i$, $i = 1, 2, \cdots, k$, and that $Z = U_1 \otimes V_1^{\mathrm{T}} \oplus U_2 \otimes V_2^{\mathrm{T}} \oplus \cdots \oplus U_k \otimes V_k^{\mathrm{T}}$. Here, U_i and V_i denote vectors and the superscript T denotes transpose in the normal matrix sense. Note that various approaches to the definition of rank can be considered; the definition adopted here is the most convenient for our paper.

Remark B.1. As in conventional linear algebra, multiplication of two matrices can never increase rank and may lead to a reduction. That is, for two matrices A, B, $\operatorname{rank}(A \otimes B) \leqslant \min(\operatorname{rank}(A), \operatorname{rank}(B))$.

B.1. Graph concepts

To facilitate the analysis of convergence properties, we recall the following graph theoretic concepts, some which have been adapted from analysis of nonnegative matrices.

Definition B.6 (Directed graph). A directed graph (or digraph for short) $\mathcal{G}$ is defined as a pair $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of elements called nodes, numbered from 1 to n, and $\mathcal{E}$ is the set of directed arcs joining any node pair. An arc joining nodes i and j is denoted as $i \to j$.

Definition B.7 (Path). A path is defined as a sequence of nodes $(i_1, i_2, \cdots, i_p)$ such that there is an arc from node i_{j-1} to node i_j for $j = 2, 3, \ldots, p$. Equivalently, a path can also be defined as a sequence of arcs which connects a sequence of nodes. An elementary path is one in which no node appears more than once. We will denote a path either in full as $i_1 \to i_2 \to i_3 \to \cdots \to i_p$, or $i_1 \to \to i_p$, indicating only the terminating nodes for short.

Definition B.8 (Circuit). A circuit is a path for which the initial and terminating nodes are identical. Similarly, an elementary circuit is defined as one in which no intermediate nodes occur more than once.

Definition B.9 (Precedence graph corresponding to a matrix). The precedence graph corresponding to an $n \times n$ matrix A, denoted as $\mathcal{G}(A)$, is a weighted digraph with n nodes. An arc from node j to node i is present in $\mathcal{G}(A)$ if and only if $a_{ij} \neq \varepsilon$. The value of a_{ij} is then called the weight of this arc.

Definition B.10 (Weight, length, average weight, circuit mean). The weight $w(\rho)$ of a path $\rho = i_1 \to i_2 \to \cdots \to i_{l-1} \to i_l$ is the sum of the weights

of the individual arcs. The length $l(\rho)$ of the same path is equal to the number of arcs in the path. The average weight of a path is its weight divided by its length: $w(\rho)/l(\rho) = (a_{i_l i_{l-1}} + a_{i_{l-1} i_{l-2}} + \cdots + a_{i_3 i_2} + a_{i_2 i_1})/(l-1)$. The circuit mean is the average weight of a circuit.

Definition B.11 (Critical circuit). For a given precedence graph, any circuit of maximum average weight is called a critical circuit.

Definition B.12 (Transition graph corresponding to a matrix). As an extension to Definition B.9, a transition graph $\mathcal{T}(A)$ associated with A depicts the node-to-node transitions (a_{ij} being the weight of the directed arc $j \to i$) with explicit distinction of the starting and end nodes achieved by 'stretching' out the precedence graph $\mathcal{G}(A)$, as shown in Fig. 2.

Remark B.2. In view of Definitions B.4, B.9 and B.12, the matrix product $C = A \otimes B$ may be visualised as the concatenation of two transition graphs in the order[4] shown in Fig. 3. Each c_{ij} entry denotes the maximum weight of a path over all paths of length 2 (since this product consists of 2 terms only) from node j to node i. For example, $c_{12} = \max(a_{11} + b_{12}, a_{12} + b_{22})$. On the other hand, a value of ε indicates the non-existence of a path.

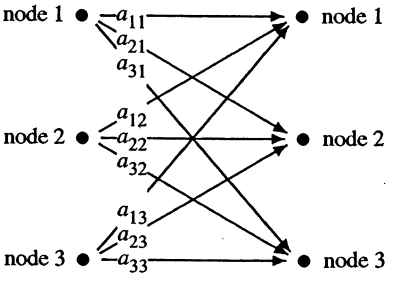

Fig. 2. The transition graph of A, a (3×3) matrix.

[4] The ordering of the two transition graphs follows from the definition of weights of the directed arcs in the respective precedence graphs.

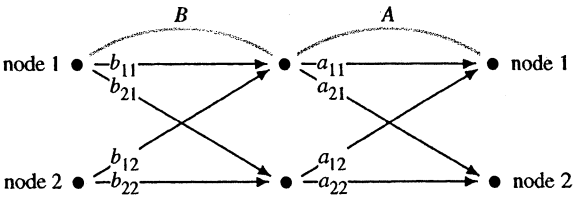

Fig. 3. Graphical illustration of multiplication of two (2×2) matrices in max plus algebra.

Definition B.13 (Irreducibility, strongly connectedness). A square matrix A is irreducible if no permutation matrix P exists such that the matrix $\tilde{A}$ defined as

$$\tilde{A} = P^{\mathrm{T}} \otimes A \otimes P$$

has an upper triangular block structure. In the max plus context, upper triangular means that the elements in the lower triangular portion all have the numerical value $\varepsilon = -\infty$.

The precedence graph associated with an irreducible matrix A is called strongly connected, to reflect the easily proved consequence of irreducibility that there always exists an elementary path between nodes i and j, $\forall i,j$. The converse also holds.

Definition B.14 (Aperiodicity). An irreducible square matrix A is aperiodic[5] if an N exists such that for all $k \geqslant N$ and for all i,j, it holds that $(A^k)_{ij} \neq \varepsilon$.

Lemma B.1. *An irreducible matrix A such that $a_{ii} \neq \varepsilon$ for at least one i, is aperiodic.*[6]

Proof. An extension of the argument in [8]. □

Definition B.15 (Critical graph). The critical graph $G^c(A)$ corresponding to an $n \times n$ matrix A consists of those nodes or arcs of $G^c(A)$ which belong to a critical circuit.

[5] This is analogous to the notion of primitivity in conventional linear algebra, which states that a nonnegative matrix A is primitive if there exists a k such that A^r is strictly positive for all $r \geqslant k$.

[6] This statement is sharper than a similar lemma [17], where a similar result was stated, but with the requirement that every diagonal entry be finite.

Definition B.16 (Cyclicity of a graph). The cyclicity of a maximally connected subgraph (m.s.c.s) of a graph is the greatest common divisor of the lengths of all its circuits. The cyclicity of the graph is the least common multiple of the cyclicities of all the m.s.c.s's.

Lemma B.2. *Suppose that A is aperiodic, or equivalently that G(A) is strongly connected. Then there exists Q and an eigenvalue λ and a matrix P such that*

$$\forall q \geqslant Q, \quad A^q = \lambda^q P$$

if and only if the cyclicity of each m.s.c.s. of $G^c(A)$ is 1. Moreover, Q is expressible as a sum of r dyads, where r is the number of m.s.c.s's of $G^c(A)$.

Corollary B.1. *Suppose A is aperiodic and G(A) has a single critical circuit of length 1. Then there exists Q and vectors u and v such that $\forall q \geqslant Q$,*

$$A^q = \lambda^q u \otimes v.$$

(Note that for A^q to have the dyadic form, it is not necessary that it have a single critical circuit of length 1.)

References

[1] B.D.O. Anderson, Stability properties of Kalman–Bucy filters, J. Franklin Inst. 291 (1971) 137–144.

[2] B.D.O. Anderson, New developments in the theory of positive systems, in: C.I. Byrnes, B.N. Datta, D.S. Gilliam, C.F. Martin (Eds.), Systems and control in the 21st century, Birkhauser, Boston, 1996, pp. 17–36 (ISBN: 0-8176-3881-4/3-7643-3881-4).

[3] B.D.O. Anderson, S. Chirarattananon, Smoothing as an improvement on filtering: a universal bound, Electron. Lett. 7 (1971) 524.

[4] B.D.O. Anderson, J.B. Moore, Optimal filtering, Prentice-Hall, Englewood Cliffs, NJ, 1979.

[5] B.D.O. Anderson, J. B. Moore, Detectability and stabilizability of discrete-time linear systems, SIAM J. Control Optim. 19 (1981) 20–32.

[6] A. Arapostathis, S.I. Marcus, Analysis of an identifi≈ation algorithm arising in the adaptive estimation of Markov chains, Math. Control Signals Systems 3 (1990) 1–29.

[7] R. Atar, O. Zeitouni, Lyapunov exponents for finite state nonlinear filtering, SIAM J. Control Optim. 35 (1997) 36–55.

[8] F. Baccelli, G. Cohen, G.J. Olsder, J.P. Quadrat, Synchronization and Linearity, Wiley, New York, 1992.

[9] L.E. Baum, T. Petrie, Statistical reference for probabilistic functions of finite state Markov chains, Ann. Math. Statist. 37 (1966) 1554–1563.

[10] Y. Bengio, P. Frasconi, Diffusion of context and credit information in Markovian models, J. Artificial Intelligence 3 (1995) 246–270.

[11] P.J. Bickel, Y. Ritov, Inference in hidden Markov models I: local asymptotic normality in the stationary case, Bernoulli 2 (1996) 199–228.

[12] D. Clements, B.D.O. Anderson, A nonlinear fixed-lag smoother for finite-state Markov processes, IEEE Trans. Inform. Theory IT-21 (1975) 446–452.

[13] S. Gaubert, M. Plus, Methods and applications of (*max*, +) linear algebra, INRIA Research Report No. 3088, 1997.

[14] F. Le Gland, L. Mevel, Geometric ergodicity in Hidden Markov Models, vol. 1028, Internal Publication, IRISA, France, 1996.

[15] E. Seneta, Nonnegative Matrices and Markov Chains, 2nd Edition, Springer, Berlin, 1981.

[16] L. Shue, B.D.O. Anderson, S. Dey, Exponential stability of filters and smoothers for hidden Markov models, IEEE Trans. Signal Processing 46 (1998) 2180–2194.

[17] A.J. Viterbi, J.K. Omura, Principles of Digital Communication and Coding, McGraw-Hill, New York, 1979.