A NEW MAXIMUM LIKELIHOOD GRADIENT ALGORITHM FOR ON-LINE HIDDEN MARKOV MODEL IDENTIFICATION

Iain B. Collings

Dept. of Elec. & Electr. Engineering,
University of Melbourne, Australia

Tobias Rydén

Dept. of Mathematical Statistics,
Lund University, Sweden

ABSTRACT

This paper presents a new algorithm for on-line identification of hidden Markov model (HMM) parameters. The scheme is gradient based, and provides parameter estimates which recursively maximise the likelihood function. It is therefore a recursive maximum likelihood (RML) algorithm, and it has optimal asymptotic properties. The only current on-line HMM identification algorithm with anything other than suboptimal rate of convergence is based on a prediction error (PE) cost function. As well as presenting a new algorithm, this paper also highlights and explains a counter-intuitive convergence problem for the current recursive PE (RPE) algorithm, when operating in low noise conditions. Importantly, this problem does not exist for the new RML algorithm. Simulation studies demonstrate the superior performance of the new algorithm, compared to current techniques.

1. INTRODUCTION

Recently, hidden Markov models (HMMs) have been used in a wide variety of applications, including speech processing [8], frequency tracking [10] and signal estimation in mobile communication systems [2]. In each of these areas, the related tasks of state estimation for known models and on-line model identification are of critical importance.

On-line identification algorithms are usually based on either recursive maximum likelihood (RML) or recursive prediction error (RPE) techniques. In the context of HMMs, two slightly different RML schemes are proposed in [5] and [6], and an RPE scheme is given in [1]. The difference between the RML schemes essentially concerns the scaling matrix that pre-multiplies the gradient of the log-likelihood; in [6] this matrix is derived using ideas of the EM (expectation-maximization) algorithm, while a different approach is taken in [5]. This difference is not unimportant, however, as the scaling matrix, together with the asymptotic covariance matrix of the log-likelihood gradient, determine the rate of convergence in the vicinity of the true parameter, see [4] and [9]. The fastest achievable rate is $n^{-1/2}$, with n being the number of observations. The algorithm in [6] does not generally achieve this rate. The algorithm in [5] does, but it is less efficient than the maximum likelihood estimate (MLE), since the gradient used in the algorithm actually differs from the gradient of the log-likelihood.

The RPE scheme [1] as well as the *new* RML scheme proposed in this paper both incorporate scaling matrices that are (estimated) inverse covariance matrices of the gradient of the objective function. This makes both algorithms achieve the optimal rate of convergence, although the asymptotic covariance matrix of the RPE algorithm is suboptimal. Additionally, and rather counter-intuitively, local convergence problems arise for the RPE algorithm in low noise conditions.

In this paper we do two things: highlight, and provide an explanation for, the failure of the RPE algorithm in low noise, and also present new RML algorithms to overcome the problem without sacrificing convergence rates. Simulation examples show that our new RML algorithm can satisfactorily identify HMM parameters, even in conditions where the RPE scheme becomes caught in local minima.

The general approach to estimation/identification in this paper is similar to that in [1], where the HMM is formulated in such a way as to allow gradient algorithms to be applied. In the case of time varying models, the RPE and RML schemes can be replaced by versions of the extended Kalman filter, in order to generate fully adaptive estimation/identification schemes. The parameters of interest are the state values and transition probabilities of the Markov chain. In this paper the transition probabilities are parametrised on a sphere so as to ensure that the derivatives are smooth, and that the estimates remain positive.

The paper is organised as follows. Sections 2 and 3 formulate the HMM and information-state models. Section 4 presents our new RML algorithm, working on the sphere for

supported by the Centre for Sensor Signal and Information Processing
supported by the Swedish Natural Science Research Council (contract no. M-AA/MA 10538-303)

the constrained transition probability estimates. In Section 6 simulation examples are given.

2. PROBLEM FORMULATION

2.1. State space signal model

Let $\{X_k\}$ be a discrete-time, homogeneous, first order Markov chain taking values in a set with N elements. Without loss of generality, we can be identify these N elements with the set of unit vectors $\{e_1, e_2, \ldots, e_N\}$, where $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)' \in \mathbb{R}^N$ with 1 in the i^{th} position. (The prime denotes transpose.) It is well known that the dynamics of $\{X_k\}$ can be written as follows [3]:

$$X_{k+1} = A'X_k + M_{k+1}, \tag{1}$$

where A is the $N \times N$ transition probability matrix with elements $a_{ij} = P(X_{k+1} = e_j \mid X_k = e_i)$, and M_k is a martingale increment. Of course $a_{ij} \geq 0$ and $\sum_{j=1}^N a_{ij} = 1$ for each i.

Also, consider the observation process

$$Y_k = g(X_k) + W_k, \tag{2}$$

where without loss of generality $g(X_k) = \langle g, X_k \rangle$, since X_k is in a finite set (where $\langle ., . \rangle$ denotes the inner product), and $g \in \mathbb{R}^N$ is the vector of state values of the Markov chain. In this paper we consider that W_k is a white (or more precisely, independent and identically distributed) Gaussian noise (WGN) process, with zero mean and standard deviation σ_w. Let $b_i(y)$ be the probability density function of Y_k conditional on $X_k = i$, and write $b(y) = (b_1(y), \ldots, b_N(y))'$. Thus, in our setting,

$$b_i(y) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left\{-\frac{1}{2}\left(\frac{y - g_i}{\sigma_w}\right)^2\right\}. \tag{3}$$

2.2. Model parametrisation

We are concerned with recursive identification of the parameters g and A; we assume σ_w to be known. It is clear that the set of valid rows in a transition probability matrix constitute a simplex. There is, however, an alternative parametrisation, on a sphere, that is superior to the one on a simplex, as observed in [1]. Specifically, let $a_{ij} = s_{ij}^2$. Each row in the corresponding matrix S belongs to the manifold

$$\mathbf{S}^{N-1} = \left\{(x_1, \ldots, x_N) \; : \; \sum_{i=1}^N x_i^2 = 1\right\}. \tag{4}$$

The parameter of interest can then be written

$$\theta = (g_1, \ldots, g_N, s_{11}, \ldots, s_{1N}, s_{21}, \ldots, s_{NN})'. \tag{5}$$

Clearly, both A and $b(y)$ depend on this parameter, that is $A = A(\theta)$ and $b(y) = b(y;\theta)$. The dimension of θ is $N + N^2$, representing N state values and N^2 transition probabilities. The important aspect of the work in this paper is that the identification scheme involves decoupling which greatly reduces the computational complexity. In cases where the parameter θ is not as in (5), the framework described below can still be used, with appropriate changes to the gradient expressions which follow.

3. PARAMETRISED INFORMATION-STATE SIGNAL MODEL

Let the information-state, $\alpha_k(\theta)$, be the vector of conditional probabilities $P_\theta(X_k = i \mid Y_1, \ldots, Y_k)$ under θ, defined as follows:

$$\alpha_k(\theta) = E_\theta[X_k \mid Y_1, \ldots, Y_k]\ . \tag{6}$$

This vector may be recursively updated using the following equation:

$$\alpha_k(\theta) = \frac{B(Y_k;\theta)A'(\theta)\alpha_{k-1}(\theta)}{\langle B(Y_k;\theta)A'(\theta)\alpha_{k-1}(\theta), \mathbf{1}\rangle} = G(\alpha_{k-1}(\theta), Y_k;\theta)\ , \tag{7}$$

where $B(y;\theta) = \text{diag}(b(y;\theta))$ is the diagonal matrix with entries given by $b(y;\theta)$, and $\mathbf{1}$ is an $N \times 1$ vector of ones.

Note that from now on, we will drop the obvious dependence of the variables on θ, for ease of notation.

For the algorithms which follow, we also require the derivative of the information-state update equation. Let $\eta_k(i;\theta) = D_{g_i}\alpha_k(\theta)$ and $\rho_k(i,j;\theta) = D_{s_{ij}}\alpha_k(\theta)$, where D denotes differentiation. These derivatives may be recursively updated as follows:

$$\begin{aligned}\eta_k(i) &= \frac{[D_{g_i}B(Y_k)]A'\alpha_{k-1} + B(Y_k)A'\eta_{k-1}(i)}{\langle \mathbf{1}, B(Y_k)A'\alpha_{k-1}\rangle} \\ &\quad -B(Y_k)A'\alpha_{k-1}\frac{\langle \mathbf{1}, [D_{g_i}B(Y_k)]A'\alpha_{k-1} + B(Y_k)A'\eta_{k-1}(i)\rangle}{\langle \mathbf{1}, B(Y_k)A'\alpha_{k-1}\rangle^2} \\ &= G_i^1(\alpha_{k-1}, \eta_{k-1}(i), Y_k;\theta)\ ,\end{aligned} \tag{8}$$

where

$$D_{g_i}B(Y_k;\theta) = \left(\frac{Y_k - g_i}{\sigma_w^2}\right)B(Y_k;\theta)\,\text{diag}(e_i)\ , \tag{9}$$

and

$$\begin{aligned}\rho_k(i,j) &= \frac{B(Y_k)[D_{s_{ij}}A']\alpha_{k-1} + B(Y_k)A'\rho_{k-1}(i,j)}{\langle \mathbf{1}, B(Y_k)A'\alpha_{k-1}\rangle} \\ &\quad -B(Y_k)A'\alpha_{k-1}\frac{\langle \mathbf{1}, B(Y_k)[D_{s_{ij}}A']\alpha_{k-1} + B(Y_k)A'\rho_{k-1}(i,j)\rangle}{\langle \mathbf{1}, B(Y_k)A'\alpha_{k-1}\rangle^2} \\ &= G_{ij}^1(\alpha_{k-1}, \rho_{k-1}(i,j), Y_k;\theta)\ ,\end{aligned} \tag{10}$$

with

$$D_{s_{ij}}A'(\theta) = 2s_{ij}(e_j - \text{diag}(s_{i\cdot})s'_{i\cdot})e'_i \ . \qquad (11)$$

Also, $s_{i\cdot} = (s_{i1}, \ldots, s_{iN})$, and it should be pointed out that the derivatives are taken in the direction perpendicular to the constraint surface (4).

Now, by collecting the elements of $\eta_k(\theta)$ and $\rho_k(\theta)$ into one column vector $\zeta_k(\theta) = \text{vec}(\eta_k(\theta), \rho_k(\theta))$ we obtain the recursive update

$$\begin{aligned} \zeta_k(\theta) &= \text{vec}\begin{pmatrix} G^1_i(\alpha_{k-1}(\theta), \eta_{k-1}(\theta), Y_k; \theta) \\ G^1_{ij}(\alpha_{k-1}(\theta), \rho_{k-1}(\theta), Y_k; \theta) \end{pmatrix} \\ &= G^1(\alpha_{k-1}(\theta), \zeta_{k-1}(\theta), Y_k; \theta) \ . \end{aligned} \qquad (12)$$

4. THE NEW RML ALGORITHM

The log-likelihood $\ell_n(\theta) = \log p(Y_1, \ldots, Y_n; \theta)$ can be re-written as

$$\begin{aligned} \ell_n(\theta) &= \textstyle\sum_{k=1}^n \log p(Y_k \mid Y_{k-1}, \ldots, Y_1; \theta) \\ &= \textstyle\sum_{k=1}^n \log\langle b(Y_k; \theta), A'(\theta)\alpha_{k-1}(\theta)\rangle = \sum_{k=1}^n u_k(\theta) \end{aligned} \qquad (13)$$

where $u_k(\theta)$ denotes a log-likelihood increment. Thus, the conditional score function $D_\theta u_k(\theta)$ (not given here due to space limitations) can be written in terms of η_k, ρ_k, $D_{g_i}b(Y_k; \theta)$ and $D_{s_{ij}}A'(\theta)$, given in (8) to (11).

We can now construct our new RML algorithm as follows:

$$\widehat{\theta}_k = \pi_{\mathcal{D}}\left(\widehat{\theta}_{k-1} + P_k\psi_k\right) \ , \qquad (14)$$

where $\pi_{\mathcal{D}}$ is a projection onto the constraint domain, $\psi_{k+1} = \text{vec}(D_{g_i}u_k(\theta) \ , \ D_{s_{ij}}u_k(\theta))$ and

$$\begin{aligned} &P_k = \tfrac{1}{\lambda_k}\left(P_{k-1} + P_{k-1}\psi_k[\psi_k P_{k-1}\psi_k + \lambda_k]^{-1}\psi'_k P_{k-1}\right) \\ &\alpha_k = G(\alpha_{k-1}, Y_k; \widehat{\theta}_k), \\ &\zeta_k = G^1(\alpha_{k-1}, \zeta_{k-1}, Y_k; \widehat{\theta}_k) \ , \end{aligned} \qquad (15)$$

5. FAILURE OF THE RPE IN LOW NOISE

As discussed in Section 1, the aim of this paper is twofold: to present a new RML algorithm for on-line HMM identification, and to highlight a rather counter-intuitive convergence problem with the current RPE approach when operating in low noise conditions. Importantly, the problem does not occur with our new RML algorithm. This section provides an explanation for the phenomenon.

The RPE on-line identification algorithm, of [1], minimises the square of the prediction error. The prediction at time k given the past, $\widehat{Y}_k$, is a function of $\widehat{\theta}_{k-1}$ and α_{k-1} (which is itself a function of $\widehat{\theta}_{k-1}$). It can clearly be seen that if the derivative of α_{k-1} with respect to θ is zero, then the best that can be hoped for from any RPE algorithm is an estimate of the *product* $g'A'$, rather than g and A separately (due to the fact that in this case g and A only effect the cost function through the product $g'A'$ and not through α. Note that $\langle g, A'\alpha\rangle$ can be written $g'A'\alpha$ in this case). Unfortunately, of course, there are many combinations of g and A which will lead to the same product $g'A'$. This leads to many global minima for the prediction error cost function. With this in mind, it can be seen from (8) and (10) that in low noise conditions, the derivative of α_k with respect to θ is in fact almost zero leading problems with the RPE approach. (Note that the terms *low noise conditions* and *small noise* are used to indicate the situation where $\sigma_w \ll \min_{i\neq j}|g_i - g_j|$.)

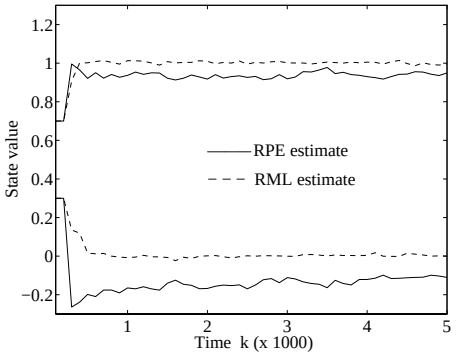
Importantly, this problem of many global minima of the prediction error cost function does not occur in the case of the maximum likelihood cost function, upon which our new RML algorithm is based. This is because importantly, unlike RPE, even in the low noise case g and A effect the ML cost function separately (that is, *not* as the product $g'A'$), as can be seen in (13).

6. SIMULATIONS

Example 1: A two state Markov chain embedded in WGN has been generated with diagonal transition probabilities of 0.9, and state levels 0 and 1. Also, $\sigma_w = 0.1$. The initial diagonal transition probability estimates were 0.1, and inital state levels 0.3 and 0.7. Figures 1 and 2 show typical parameter estimates for this data. The figures demonstrate that the estimates converge to the true values only for the new RML algorithm. The RPE algorithm has clearly converged to some other value. This value turns out to be such that the product $g'A'$ is correct, as was demonstrated previously.

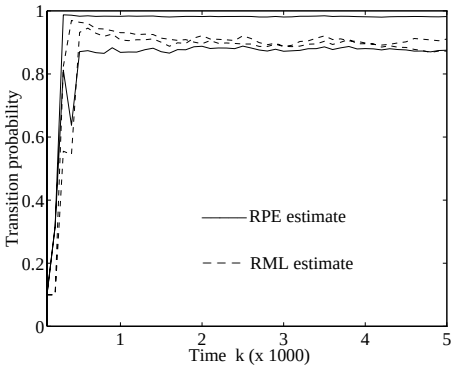
Example 2: This example confirms that the rate of convergence of model estimates has not been sacrificed for our new RML algorithm, when compared to the RPE algorithm, for cases when the RPE algorithm actually converges to the true values (i.e. in high noise cases). This is important because it demonstrates that the RML algorithm has $n^{-1/2}$-convergence. This means that RML is superiour to the previous algorithm in [6], in terms of convergence, as well as being superiour to the RPE algorithm which has problems in low noise conditions.

The model parameters are the same as in Example 1, except that $\sigma_w = 0.3$. The results are shown in Figures 3 and 4. The error function used on the vertical axes is $(1/k)\sum_{i=1}^k |\widehat{\theta}_i - \theta|$.



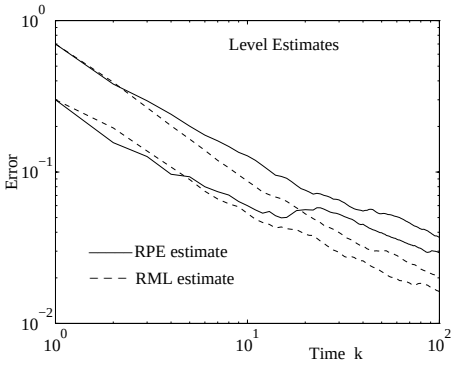
$\sigma_w = 0.1$, True values are 0 and 1

Figure 1: Level estimates of 2 state Markov chain



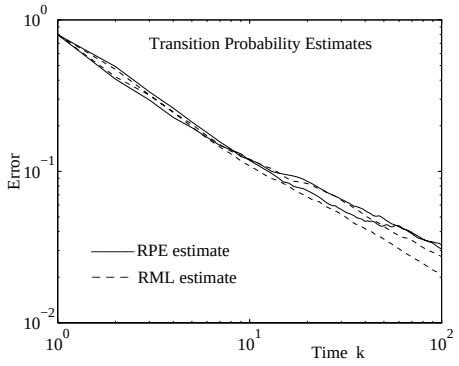
$\sigma_w = 0.1$, True value is 0.9

Figure 2: Transition probability estimates of 2 state Markov chain



$\sigma_w = 0.3$ (both algorithms converge)

Figure 3: Error function for level estimates



$\sigma_w = 0.3$ (both algorithms converge)

Figure 4: Error function for transition probability estimates

7. REFERENCES

[1] I. B. Collings, V. Krishnamurthy, and J. B. Moore, "On-line identification of hidden Markov models via recursive prediction error techniques", *IEEE Trans. Signal Process.*, vol. 42, pp. 3535–3539, 1994.

[2] I. B. Collings and J. B. Moore, "An HMM approach to adaptive demodulation of QAM signals in fading channels," *Int. J. Adapt. Control Signal Process.*, vol. 8, pp. 457–474, 1994.

[3] R. J. Elliott, "Exact adaptive filters for Markov chains observed in Gaussian noise," *Automatica*, vol. 30, no. 9, 1976.

[4] V. Fabian, "On asymptotic normality in stochastic approximation," *Ann. Math. Statist.*, vol. 39, pp. 1327–1332, 1968.

[5] U. Holst and G. Lindgren, "Recursive estimation in mixture models with Markov regime," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1683–1690, 1991.

[6] V. Krishnamurthy and J. B. Moore, "On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure," *IEEE Trans. Signal Process.*, vol. 41, pp. 2557–2573, 1993.

[7] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*. Cambridge, MA: MIT Press, 1983.

[8] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–285, 1989.

[9] D. Ruppert, "Almost sure approximations to the Robbins-Monro and Kiefer-Wolfowitz processes with dependent noise," *Ann. Prob.*, vol. 10, pp. 178–187, 1982.

[10] R. L. Streit and R. Barrett, "Frequency line tracking using hidden Markov models," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, pp. 586–598, 1990.