Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Approximate EM Algorithms for Parameter and State Estimation in Nonlinear Stochastic Models

Graham C. Goodwin *Fellow, IEEE,* Juan C. Agüero *Student, IEEE*

Abstract— **Due to the availability of rapidly improving computer speeds, industry is increasingly using nonlinear process models in calculations that appear further down the control hierarchy. Indeed, nonlinear models are now frequently used for real-time control calculations. This trend means that there is growing interest in the availability of high speed state and parameter estimation algorithms for nonlinear models. One family of algorithms that can be used for this purpose is based on the, so called, Expectation Maximization Scheme. Unfortunately, in its basic form, this algorithm requires large computational resources. In this paper we review the EM algorithm and propose several approximate schemes aimed at retaining the essential flavour of this class of algorithm whilst ensuring that the computations are tractable. We will also compare the EM algorithm with several simpler schemes via a number of examples and comment on the trade-offs that occur.**

I. Introduction

Nonlinear stochastic models appear in many applications. Examples include mathematical models used in economics [8], [33], [6], ship steering [4] etc. Usually these models arise as nonlinear continuous time stochastic differential equations. The issue of discretization of these models is discussed elsewhere - see for example the work of [34], and the recent work of [37]. Here, we adopt a general nonlinear discrete time stochastic model of the following form:

$$x_{t+1} = f(x_t, u_t, \theta) + g(x_t, u_t, \theta) w_t \tag{1}$$

$$y_t = h(x_t, u_t, \theta) + l(x_t, u_t, \theta) v_t \tag{2}$$

where $u_t \in \mathbb{R}^{n_u}$, $x_t \in \mathbb{R}^n$, $y_t \in \mathbb{R}^{n_y}$ are a measured input, states, and measured output signal respectively. The random processes $\{w_t\}$, and $\{v_t\}$ are assumed to be independent zero mean Gaussian noise with covariance matrices Q, and R respectively. It is assumed that the functions $f(\cdot)$, $g(\cdot)$, $h(\cdot)$, and $l(\cdot)$ are analytic functions [3]. For simplicity, we also assume that $g(\cdot)$ and $l(\cdot)$ are non-singular. Note that the results should be able to be extended to the singular case -see the recent work of [28], [29].

In the current paper, we will focus on parameter and state estimation for nonlinear discrete-time stochastic system using Expectation-Maximization (**EM**) based algorithms. **EM** is an iterative two step procedure [10], [21] where (loosely speaking) one estimates the state based in the current parameter estimate in the first step, and then updates the parameters by maximizing a function that depends on the joint probability density function $p(X_N, Y_N|\theta)$ evaluated at the estimated states. For the case of linear systems, this approach yields an explicit form for both steps of the algorithm (See [27], [16], [23]). The method has also been extended to the case of bilinear systems in [13]. Another class of systems that can be treated in this fashion is those of Hammerstein form [20]. To illustrate, say that we have the following scalar function of powers of $\bar{u}_t$, $u_t = \sum_{i=1}^{r} \alpha_i \bar{u}_t^i$, where $\bar{u}_t$ is the input to the nonlinear part of the system and u_t is the input to a linear system. Then this is equivalent to a linear model having an "r" dimensional input vector $U_t = \begin{bmatrix} \bar{u}_t & \bar{u}_t^2 & \cdots & \bar{u}_t^r \end{bmatrix}^T$. We can then apply the usual EM algorithm applicable to the linear case leading to an explicit result for the E and M steps.

Unfortunately, however, in most estimation problems, neither of the steps in the EM algorithm have an explicit solution [19], [31]. For the case of general non-linear systems, no explicit form exists. In this case, Monte-Carlo sampling techniques have been utilized [2], [11] to perform the E-step. It is also possible to utilize the Extended Kalman Filter (**EKF**) for the E-step [12], [26]. Also, for certain classes of systems it is possible to obtain a closed form solution for the M-step. For example, in [12], [26] a particular kind of neural network model using radial basis functions is considered.

In this paper we will utilize linearization and Kalman Filter techniques. However, a distinctive feature of our approach is the choice of trajectory about which we linearize. We choose this trajectory via a nonlinear maximum a-posteriori estimator. We show that this yields an approximate **EM** scheme which performs very well. Indeed, simulation studies suggest that the new scheme has several advantage over the earlier methods described above.

The layout of the remainder of the paper is as follows. In section II, we describe the basic **EM** algorithm. In section III, we review various approximations that have been used in the context of the **EM** algorithm. In section IV, we describe a novel **EM** type algorithm which, inter alia, uses a maximum a-posteriori estimate as the basis of a local linearization. In section V we present several examples.

II. The EM algorithm

The schemes that are studied in the current paper are closely related to Maximum-Likelihood (**ML**). **ML** and related algorithms such as Prediction Error (**PEM**) methods are popular algorithms in identification [17], [20]. In the **ML** framework, the following log-likelihood function is

Graham C. Goodwin and Juan C. Agüero are with the University of Newcastle, Newcastle, Australia.

Email Addresses: Graham.Goodwin@newcastle.edu.au, Juan.Aguero@newcastle.edu.au

0-7803-9568-9/05/$20.00 ©2005 IEEE

maximized:

$$l(\theta) = \log p(Y_N|\theta) \tag{3}$$

where Y_N denotes the given data set containing the system outputs i.e. $Y_N := \{y_1, y_2, \ldots, y_N\}$. For future use, we also introduce the state sequence $X_N := \{x_0, x_1, \ldots, x_N\}$.

In simple cases, the likelihood function (3) can be easily obtained as the following marginal distribution:

$$l(\theta) = \log \int p(X_N, Y_N|\theta)dX_N \tag{4}$$

based on the joint probability density function $p(X_N, Y_N|\theta)$.

However, in general, the evaluation of the integral in (3) presents significant difficulties and hence approximations are required, e.g. based on particle filtering. It is also possible to avoid the integration involved in (3) by utilizing the **EM** algorithm.

The **EM** algorithm may be summarized as follows [10]:

1) Choose an initial estimate $\hat{\theta}_0 \in \Omega$, where Ω is a constraint set in the parameter space.
   Then, for $i = 0, 1, \cdots$
2) Compute the auxiliary function $\mathbf{Q}(\theta, \hat{\theta}_i)$ which is the expected value of the complete data log-likelihood with respect to the random variable X_N (usually called "hidden data" in the statistics literature) given the observed data Y_N and the previous estimate $\hat{\theta}_i$:

$$\mathbf{Q}(\theta, \hat{\theta}_i) = \mathop{\mathrm{E}}_{X_N}\{\log[p(X_N, Y_N|\theta)]|Y_N, \hat{\theta}_i\} \tag{5}$$

3) Set $\hat{\theta}_{i+1} = \arg\max_{\theta\in\Omega} \mathbf{Q}(\theta, \hat{\theta}_i)$.
4) Go to step 2, and continue until convergence.

Steps 2 and 3 are usually known as the E-Step and M-Step respectively. Under quite general conditions [10], [36], [7], the **EM** algorithm can be proved to converge to a stationary point which in many practical applications will be a local maximum of the likelihood function [21].

The basic idea of the **EM** algorithm is to decompose the log-likelihood function as:

$$\begin{aligned}\log[p(Y_N|\theta)] =& \mathop{\mathrm{E}}_{X_N}\{\log[p(X_N, Y_N|\theta)]|Y_N, \theta_i\} \\ &- \mathop{\mathrm{E}}_{X_N}\{\log[p(X_N|Y_N, \theta)]|Y_N, \theta_i\} \\ =&\mathbf{Q}(\theta, \theta_i) - \mathbf{H}(\theta, \theta_i)\}\end{aligned} \tag{6}$$

where the fact that $\mathop{\mathrm{E}}_{X_N}\{\log[p(Y_N|\theta)]|Y_N, \theta_i\} = \log[p(Y_N|\theta)]$ has been utilized.

By using the Jensen inequality [24], it is possible to prove that $\mathbf{H}(\theta, \theta_i) \leq \mathbf{H}(\theta_i, \theta_i)$ [10]. Then, if we can find a method to choose θ such that $\mathbf{Q}(\theta, \theta_i) \geq \mathbf{Q}(\theta_i, \theta_i)$, we have a simple mechanism to iteratively maximize the log-likelihood function $l(\theta)$.

III. Review of approximations used in the EM algorithm

In most estimation problems, neither of the steps in the **EM** algorithm have an explicit solution [19], [31]. Many different approximations have been proposed in the statistics literature. Indeed, there is a substantial literature on this topic. A brief selection of the methods is contained in [32], [18], [35], [31], [22] and the references therein. In this section, we describe some of these approaches which are more pertinent to our subsequent analysis.

For the E - step, Monte-Carlo methods have been proposed (see for example [31] and the references therein). However, this approach has difficulties when the number of data N grows [2]. In general the E-step can be approximated as

$$\mathbf{Q}(\theta, \theta_i) \approx \frac{1}{m}\sum_{j=1}^{m} \log[p(X_N^{(j)}, Y_N|\theta)] \tag{7}$$

where $X_N^{(j)}$ is a sample from the joint probability density function $p(X_N^{(j)}, Y_N|\theta_i)$. It is also possible to use $m = 1$, provided $X_N^{(1)} = \hat{X}_N$ is some "good" summary of $p(X_N, Y_N|\theta_i)$, such as a mode or expected value. These approaches are usually called "**EM-type**" algorithms [35]. It is also possible [12], [26] to use the Extended Kalman Filter to obtain samples of the joint distribution.

The traditional **EM** method requires that the function $Q(\theta, \theta_i)$ be maximized with respect to θ. One variant of the M-Step is to simply find θ_{i+1} such that $\mathbf{Q}(\theta_{i+1}, \theta_i) > \mathbf{Q}(\theta_i, \theta_i)$. This class of algorithm is usually called Generalized Expectation Maximization (**GEM**) [10]. In this context, a variant of the traditional **EM** has been proposed in [32], [18], [19]. One idea described in [19], is to use the first iteration of Newton's algorithm for the M-step. The idea in Newton's algorithm is to maximize, at each iteration, a quadratic approximation, about θ_i, of the function $\mathbf{Q}(\theta, \theta_i)$ [5]. Even though it is possible to carry out the M - Step exactly by using a Newton type algorithm, this is not attractive in the sense that it leads to iterations within iterations. This method is usually called the Gradient Expectation Maximization, **GradEM**, algorithm.

In the **GradEM** algorithm the following recursion is utilized:

$$\theta_{i+1} = \theta_i - \mathcal{H}(\theta_i, \theta_i)^{-1}\nabla(\theta_i, \theta_i) \tag{8}$$

where $\mathcal{H}(\theta, \theta_i) = \frac{\partial^2\mathbf{Q}(\theta,\theta_i)}{\partial\theta\partial\theta^T}$ and $\nabla(\theta, \theta_i) = \frac{\partial\mathbf{Q}(\theta,\theta_i)}{\partial\theta}$ denotes the Hessian, and the first derivative of $Q(\theta, \theta_i)$ with respect to θ respectively.

The local and global convergence of this algorithm have been analyzed in [19]. Additionally, since Newton's method often converges quickly, the local properties of the **GradEM** algorithm are almost identical to those of the **EM** algorithm. In fact, any strict local maximum point of the log-likelihood function $l(\theta)$ locally attracts both algorithms at the same rate of convergence [19].

Note that because the function $\mathbf{H}(\theta, \theta_i)$ in (6) achieves its maximum at $\theta = \theta_i$, then its first derivatives vanish at that point. It then follows that

$$\begin{aligned}\left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\theta=\theta_i} &= \left.\frac{\partial \mathbf{Q}(\theta,\theta_i)}{\partial \theta}\right|_{\theta=\theta_i} - \left.\frac{\partial \mathbf{H}(\theta,\theta_i)}{\partial \theta}\right|_{\theta=\theta_i} \\ &= \left.\frac{\partial \mathbf{Q}(\theta,\theta_i)}{\partial \theta}\right|_{\theta=\theta_i} \end{aligned} \tag{9}$$

and hence the iteration (8) can also be written as

$$\theta_{i+1} = \theta_i - [\mathcal{H}(\theta,\theta_i)]^{-1} \left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\theta=\theta_i} \tag{10}$$

We see that this is similar to a Newton algorithm to maximize the log-likelihood function $l(\theta)$. The only difference is that instead of using the Hessian of the log-likelihood function, the Hessian of the function $\mathbf{Q}(\theta,\theta_i)$ is used. Unfortunately, the **GradEM** algorithm does not necessarily lead to a sequence θ_i which increases the log-likelihood function $l(\theta)$. (It does not necessarily go "up-hill") [21]. This problem can be solved by using more than one iteration in the Newton algorithm, or by using different variants of the **GradEM** algorithm [21]. Moreover, Newton's algorithm usually suffers from other problems (e.g. the inverse of the Hessian may not exist at some points, the algorithm is attracted by local maxima just as much as it is attracted by local minima, etc.). These problems can be solved by using modifications of Newton's algorithm. Some choices are described in [5]. Alternatively, following similar reasoning to that used above to motivate the **GradEM** algorithm, one can use the first iterations of any optimization procedure (relaxation, quasi-Newton, etc.) in order to obtain an estimate that gives a greater value for $\mathbf{Q}(\theta,\theta_i)$.

IV. A NOVEL **EM** BASED SCHEME FOR NON-LINEAR STOCHASTIC MODELS

As discussed above, it is generally impossible to obtain closed form expressions for the E and M steps. Indeed, for the nonlinear case, the E-step inherently involves some form of nonlinear filtering. However, this involves a heavy computational load. In an effort to avoid this, we describe below an approximate algorithm which is aimed at exploiting, as far as possible, the closed form results available for the linear case.

We describe the proposed algorithm under the two headings of E and M step.

- **The E-step:** The E-step requires that we calculate the following

$$\begin{aligned}\mathbf{Q}(\theta,\theta_i) &= \underset{X_N}{\mathrm{E}}\{\log[p(X_N,Y_N,\theta)]|Y_N,\theta_i\} \\ &= \int \log[p(X_N,Y_N|\theta)]p(X_N|Y_N,\theta_i)dX_N \end{aligned} \tag{11}$$

This equation suggests that we actually have to solve two sub-problems, namely, (i) *obtain an expression for the conditional distribution of the states (evaluated at $\hat{\theta}_i$) given the data* and (ii) *take the expected value of* $\log[p(X_N,Y_N|\theta)]$ *over this distribution.*

Thus, the E-step requires us to evaluate

$$\begin{aligned}\mathbf{Q}(\theta,\theta_i) &= \underset{X_N}{\mathrm{E}}\{\log[p(X_N,Y_N,\theta)]|Y_N,\theta_i\} \\ &= \underset{X_N}{\mathrm{E}}\{V_0(\theta) + \sum_{t=1}^{N} V_t(x_t,x_{t-1},\theta)|Y_N,\theta_i\} \end{aligned} \tag{12}$$

where for the case of Gaussian noise

$$\begin{aligned} V_0(\theta) =& \log[p(x_0|\theta)] \\ =& const - \frac{1}{2}\log[|P_0|] \\ &- \frac{1}{2}(x_0-\mu)^T P_0^{-1}(x_0-\mu) \\ V_t(x_t,x_{t-1},\theta) =& \log[p(y_t|x_t,\theta)] + \log[p(x_t|x_{t-1},\theta)] \end{aligned} \tag{13}$$

and where

$$\begin{aligned}\log[p(y_t|x_t,\theta)] = &-\frac{1}{2}\log[|l(x_t,u_t,\theta)Rl(x_t,u_t,\theta)^T|] \\ &- \frac{1}{2}(y_t - h(x_t,u_t,\theta))^T[l(x_t,u_t,\theta)Rl(x_t,u_t,\theta)^T]^{-1} \\ &(y_t - h(x_t,u_t,\theta)) \end{aligned} \tag{14}$$

$$\begin{aligned}\log[p(x_{t+1}|x_t,\theta)] = &-\frac{1}{2}\log[|g(x_t,u_t,\theta)Qg(x_t,u_t,\theta)^T|] \\ &- \frac{1}{2}(x_{t+1} - f(x_t,u_t,\theta))^T[g(x_t,u_t,\theta)Qg(x_t,u_t,\theta)^T]^{-1} \\ &(x_{t+1} - f(x_t,u_t,\theta)) \end{aligned} \tag{15}$$

We outline below a sub-optimal strategy aimed at obtaining a good approximation to (12). We assume we have available very good initial estimates $\{\bar{x}_t\}$, $\{\bar{w}_t\}$, $\{\bar{v}_t\}$ of the states, process noise, and measurement noise (more will be said about this below). Then one can linearize the state space model about these initial estimates as follows:

$$\begin{aligned} x_{t+1} \approx& f(\bar{x}_t,u_t,\theta) + g(\bar{x}_t,u_t,\theta)\bar{w}_t \\ &+ \frac{\partial f(\bar{x}_t,u_t,\theta)}{\partial x_t}[x_t - \bar{x}_t] + \frac{\partial g(\bar{x}_t,u_t,\theta)}{\partial x_t}[x_t - \bar{x}_t]\bar{w}_t \\ &+ g(\bar{x}_t,u_t,\theta)[w_t - \bar{w}_t] \end{aligned} \tag{16}$$

$$\begin{aligned} y_t \approx& h(\bar{x}_t,u_t,\theta) + l(\bar{x}_t,u_t,\theta)\bar{v}_t \\ &+ \frac{\partial h(\bar{x}_t,u_t,\theta)}{\partial x_t}[x_t - \bar{x}_t] + \frac{\partial l(\bar{x}_t,u_t,\theta)}{\partial x_t}[x_t - \bar{x}_t]\bar{v}_t \\ &+ l(\bar{x}_t,u_t,\theta)[v_t - \bar{v}_t] \end{aligned} \tag{17}$$

Based on the above, we now address the two sub-problems in the E-step:

(i) *Obtain an approximation to the conditional distribution of the states, given the data (evaluated at $\hat{\theta}_i$)*: The linear form of (16), (17) suggests that one can directly apply the (time-varying) Kalman Filter to approximate the conditional distribution of x_t given the data evaluated at $\hat{\theta}_i$.

(ii) *Take the expected value of* $\log[p(X_N, Y_N|\theta)]$: Next we turn to the problem of evaluating the expectation in (12). This step is greatly facilitated by the availability of a Gaussian approximation to the conditional state distribution as provided in step (i). In particular, one can use a Taylor's series expansion of $V_t(x_t, x_{t-1}, \theta)$ about $\bar{x}_t$, $\bar{x}_{t-1}$; i.e.

$$\begin{aligned} V_t \approx & V_t(\bar{x}_t, \bar{x}_{t-1}) + \frac{\partial V_t(\bar{x}_t, \bar{x}_{t-1})}{\partial x_t}[x_t - \bar{x}_t] \\ & + \frac{\partial V_t(\bar{x}_t, \bar{x}_{t-1})}{\partial x_{t-1}}[x_{t-1} - \bar{x}_{t-1}] \\ & + \frac{1}{2}[x_t - \bar{x}_t]^T \frac{\partial^2 V_t(\bar{x}_t, \bar{x}_{t-1})}{\partial x_t \partial x_t}[x_t - \bar{x}_t] \\ & + \frac{1}{2}[x_{t-1} - \bar{x}_{t-1}]^T \frac{\partial^2 V_t(\bar{x}_t, \bar{x}_{t-1})}{\partial x_{t-1} \partial x_{t-1}}[x_{t-1} - \bar{x}_{t-1}] \\ & + \frac{1}{2}[x_t - \bar{x}_t]^T \frac{\partial^2 V_t(\bar{x}_t, \bar{x}_{t-1})}{\partial x_t \partial x_{t-1}}[x_t - \bar{x}_{t-1}] + \cdots \end{aligned} \tag{18}$$

An advantage of having a Gaussian approximation to the state distribution is that *all* moments are simple functions of the mean and covariance [30], [17]. Hence, irrespective of the order of the Taylor's series used, one can readily evaluate the expected value in (12).

Finally, we return to the question as to how one can best obtain $\{\bar{x}_t\}$, $\{\bar{w}_t\}$ and $\{\bar{v}_t\}$ about which we linearize the system. Our proposal is to use a Maximum-A-Posteriori (MAP) estimate. Such an estimate can be obtained by maximizing

$$\log[p(X_N, Y_N|\theta)] = V_0(\theta) + \sum_{t=1}^{N} V_t(\theta) \tag{19}$$

with respect to $\{x_0, \cdots, x_N\}$. This leads to the maximum a-posteriori estimates $\{\bar{x}_0, \cdots, \bar{x}_N\}$. Also, by inverting (1), (2) we also obtain estimates $\{\bar{w}_0, \cdots, \bar{w}_N\}$, and $\{\bar{v}_0, \cdots, \bar{v}_N\}$ for the sequences $\{w_t\}$, and $\{v_t\}$ respectively. Note that one requires that the joint distribution be symmetric and unimodal to ensure that the MAP estimate is equal to the mean [9].

We make a further simplification by recognizing that practical systems have finite memory, i.e. distant data tells us little about the current state. Hence, we use Rolling Horizon state estimation [25]. The key idea here is to consider only data from $[\text{Min}(1, k-L), \text{Max}(N, k+L)]$ when estimating x_k. Here L is an a-priori estimate of the "memory" of the system. For example, if we choose $L = 10$, then, each MAP estimate is based on a maximization over (a maximum of) 21 variables.

Remark 4.1: Notice that, in the case of linear stochastic systems, the expressions given above are exact, i.e. the proposed algorithm for the nonlinear case builds on the known closed form expression for the linear case [27], [15]. ▽▽▽

- **The M-step:** Having approximated $\mathbf{Q}(\theta, \theta_i)$, the M-step simply requires that this function be maximized with respect to the parameters. This requires some form of iterative algorithm. We propose using the **GradEM** algorithm briefly described earlier.

V. Examples

A. Example 1

Consider the following simple example:

$$x_{t+1} = ax_t + bu_t + w_t \tag{20}$$

$$y_t = cx_t + du_t + ex_t^2 + v_t \tag{21}$$

where $w_t \sim \mathrm{N}(0, Q)$, $v_t \sim \mathrm{N}(0, R)$, and $a = 0.8$, $c = 10$, $b = d = e = 1$, and $Q = R = 0.1$. The measured input, u_t, was chosen as Gaussian white noise of covariance $\sigma_u^2 = 1$. $N = 300$ data points were collected from the system.

Notice that by doing a similarity transformation $x_t = \alpha z_t$, we obtain the following equivalent system:

$$z_{t+1} = az_t + \frac{b}{\alpha}u_t + \frac{1}{\alpha}w_t \tag{22}$$

$$y_t = c\alpha z_t + du_t + e\alpha^2 z_t^2 + v_t \tag{23}$$

Hence, we see that the model is overparameterized since (at least) two parameter values give the same description. For the sake of comparison we will make a state transformation to the estimated model (the one obtained after applying the estimation algorithm), in order to have a normalized state representation such that $\hat{c} = 10$.

For this system we have that:

$$p(x_0|\theta) = \frac{\exp\left\{-\frac{1}{2P_0}(x_0 - \mu)^2\right\}}{\sqrt{2\pi P_0}} \tag{24}$$

$$p(y_t|x_t, \theta) = \frac{\exp\left\{-\frac{1}{2R}(y_t - cx_t - du_t - ex_t^2)^2\right\}}{\sqrt{2\pi R}} \tag{25}$$

$$p(x_{t+1}|x_t, \theta) = \frac{\exp\left\{-\frac{1}{2Q}(x_{t+1} - ax_t - bu_t)^2\right\}}{\sqrt{2\pi Q}} \tag{26}$$

and thus

$$\begin{aligned} V_0 &= -\frac{1}{2}\log[2\pi P_0] - \frac{1}{2P_0}(x_0 - \mu)^2 \\ V_t &= -\frac{1}{2}\log[2\pi Q] - \frac{1}{2Q}(x_t - ax_{t-1} - bu_{t-1})^2 \\ &\quad - \frac{1}{2}\log[2\pi R] - \frac{1}{2R}(y_t - cx_t - du_t - ex_t^2)^2 \end{aligned} \tag{27}$$

Assuming that we have good estimates $\{\bar{x}_t\}$, $\{\bar{u}_t\}$, $\{\bar{w}_t\}$ for some parameter value $\hat{\theta}_i$, then the linearized model about $\{\bar{x}_t\}$, $\{\bar{v}_t\}$, $\{\bar{w}_t\}$ is of the form:

$$x_{t+1} = ax_t + bu_t + w_t \tag{28}$$

$$y_t = [c + 2e\bar{x}_t]x_t + du_t - e\bar{x}_t^2 + v_t \tag{29}$$

To obtain the linearization point, we will use the MAP estimate as described in section IV. We choose $L = 10$ for the rolling horizon MAP estimator.

Given $\{\bar{x}_t\}$, and the approximate model (28), (29) we then utilize the Kalman Smoother to obtain a Gaussian approximation for the joint conditional distribution of $\{x_t, x_{t-1}\}$. Next, we take the expectation in (12). Here, we note from (27) that only $4th$ order moments are needed to obtain the result. Moreover, as explained in section IV, all moments are readily calculable from the mean and covariance results provided by the Kalman smoother. This completes the E-step.

Finally, for the M-step, we note that for this model there are closed form expressions, as in the linear case, since the model is linear in the parameters.

For the sake of comparison we have tested the following algorithms:

1) Certainty Equivalence EM (CE-EM): Here the MAP estimates $\bar{x}_t$, $\bar{w}_t$, $\bar{v}_t$ are assumed to lie at a point such that the entire probability mass is at these points. This means that the expectation in (12) can be replaced by simply evaluating (7) with $m = 1$.
2) EKF-EM: EM based on the use of the EKF as in [26].
3) MAP-EM: The new algorithm as described in section IV.

We fix $\hat{c} = 10$ as explained previously. Algorithm 1 appears to give satisfactory results provided the measured input variance is large (e.g. $\sigma_u^2 = 10$). However, when this variance is reduced to $\sigma_u^2 = 1$, the algorithm fails to converge. Figure 1 shows the convergence of algorithms 2 and 3 with initial values $\hat{a} = 0.5$, $\hat{b} = \hat{d} = \hat{e} = 0$, $\hat{Q} = \hat{R} = 10$. The results obtained by both algorithms are essentially identical for this example.

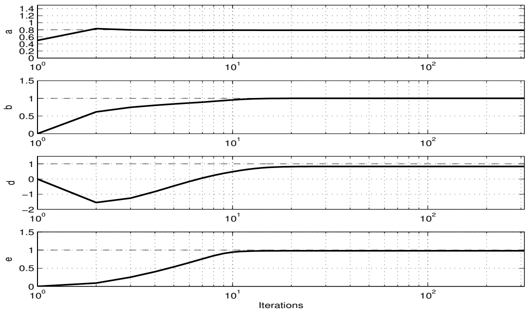


Fig. 1. Example V-A: Convergence of the parameter estimates obtained by algorithms 2 and 3.

B. Example 2

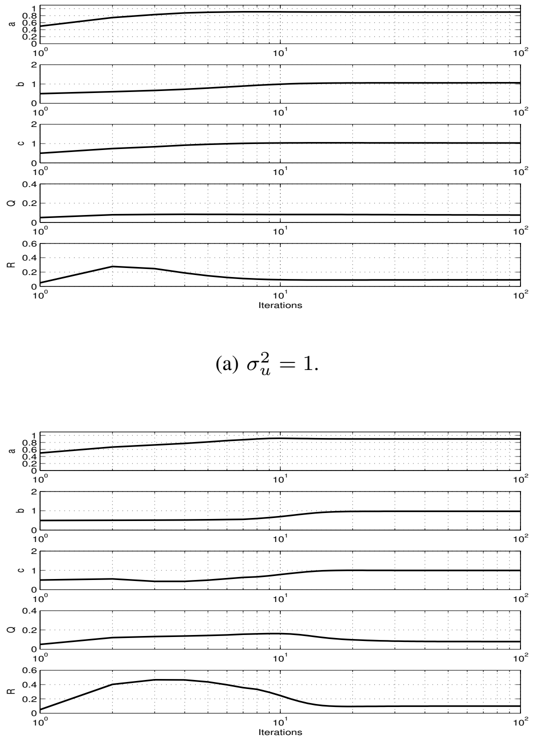
Consider the following system

$$x_{t+1} = ax_t + bu_t + w_t \tag{30}$$

$$y_t = c\,cos(x_t) + v_t \tag{31}$$

where $w_t \sim \mathrm{N}(0, Q)$, $v_t \sim \mathrm{N}(0, R)$, $u \sim N(0, \sigma_u^2)$, and $a = 0.9$, $b = c = 1$, $Q = R = 0.1$. $N = 100$ data points were collected from the system.

We follow the procedure described in section IV, and in example V-A to complete all the steps in order to estimate the parameters, we use the following initial values: $\hat{a} = \hat{b} = \hat{c} = 0.5$, $\hat{Q} = \hat{R} = 0.05$.



(a) $\sigma_u^2 = 1$.

(b) $\sigma_u^2 = 4$.

Fig. 2. Example V-B: Convergence of the parameter estimates by using algorithm 3 (MAP-EM).

Notice that in this case, under a Gaussianity assumption, it is possible to exactly calculate the expected value in (12) since if $x \sim N(\mu, \sigma^2)$ then

$$\mathrm{E}\left\{cos(x)\right\} = e^{-\frac{\sigma^2}{2}}cos(\mu) \tag{32}$$

$$\mathrm{E}\left\{cos(x)^2\right\} = \frac{1}{2}\left[1 - e^{-2\sigma^2} + 2e^{-2\sigma^2}cos(\mu)^2\right] \tag{33}$$

Thus, it is possible to use algorithm EKF-EM. Notice that, this also means that in algorithm 3 (MAP-EM) the Taylor expansion step is not necessary. However, to demonstrate the utility of the power series idea explained in step (ii) in section IV, we use a fourth order approximation in (18). The results obtained by using this approximation are identical to the ones obtained by using the expressions above in the algorithm MAP-EM.

The results obtained by using the three algorithms, EKF-EM, CE-EM, and MAP-EM are presented in table I for two different values of the measured input variance, $\sigma_u^2 = 1$, and $\sigma_u^2 = 4$. The EKF-EM obtains reasonable results for $\sigma_u^2 = 1$, but the results are very poor when the input variance is increased to $\sigma_u^2 = 4$. The CE-EM does not give good results for any of the cases analyzed. On the other hand the MAP-EM algorithm gives good results for all the cases analyzed. The convergence of the parameter estimates obtained by using MAP-EM are shown in figure 2.

Remark 5.1: Notice that, for some applications, MAP estimates are more useful than other kind of estimates such as minimum square error [14]. In particular, in this example we have shown that we obtain better results when using

TABLE I

EXAMPLE V-B: ESTIMATED PARAMETERS

$N = 100$	$\sigma_u^2 = 1$			$\sigma_u^2 = 4$		
	EKF-EM	CE-EM	MAP-EM	EKF-EM	CE-EM	MAP-EM
$\hat{a}$	0.9094	0.8346	0.9023	0.9409	0.7283	0.9004
$\hat{b}$	0.7993	0.6420	1.0593	0.2974	0.5164	0.9730
$\hat{c}$	0.8625	0.6961	1.0289	0.5873	0.2089	0.9938
$\hat{Q}$	0.0106	$1.91 \cdot 10^{-12}$	0.0765	0.1647	$1.8 \cdot 10^{-12}$	0.0793
$\hat{R}$	0.2248	0.3410	0.0912	0.3878	0.5344	0.099

MAP. However, we acknowledge that the approach proposed in algorithm 3 is computationally more demanding than algorithm 2. Indeed, an important issue in any MAP estimation procedure is the optimization of a non-convex cost function. Thus, it is important to use a "good" initial state estimate in the optimization algorithm utilized to obtain the MAP estimate. In this particular example, we obtained an initial estimate for the states by using a rolling horizon approach, with $L = 2$, but with a coarsely quantized state space. We maximized the MAP cost function considering only the following values for the state $X = \{-10, -5, 0, 5, 10\}$. This means that we searched for the MAP estimate among (a maximum of) 5^5 different possibilities in every window. Then, we used this estimate to initialize the optimization algorithm to find the MAP estimate, but now in the complete state space. $\triangledown\triangledown\triangledown$

VI. CONCLUSIONS

This paper has described a novel algorithm for nonlinear state and parameter estimation. The algorithm is of the **EM** type and uses a **MAP** estimate as the basis of a local linearization in the E-step. Also, a Taylor's series expansion has been used for the joint log likelihood so that all moments can be generated from the Gaussian approximation provided by the linearized (about the **MAP** estimate) Kalman Filter. The algorithm has been tested in simulated examples and has been shown to offer superior performance compared with other approximate **EM** algorithms described in the literature.

REFERENCES

[1] J. C. Agüero and G. C. Goodwin. EM based algorithms for parameter estimation in linear and nonlinear stochastic models. *Technical Report EE04014, The University of Newcastle, Australia.*, 2004.

[2] C. Andrieu, A. Doucet, and V. Tadic. Online sampling for parameter estimation in general state space models. *13th IFAC symposium on system identification, Rotterdam, The Netherlands*, 2003.

[3] T. M. Apostol. *Mathematical Analysis*. Addison-Wesley Publishing Company, INC., 1957.

[4] K. J. Åström and C. G. Källström. Identification of ship steering dynamics. *Automatica*, Vol. 12, No 12:9–22, 1976.

[5] D. Bertsekas. *Nonlinear programming*. Athena Scientific, 2nd edition, 1999.

[6] F. Black and M. Scholes. The pricing of options and corporate liabilities. *The Journal of political economy*, Vol. 81, No 3:637–654, 1973.

[7] R. A. Boyles. On the convergence of the EM algorithm. *Journal of the royal statistical society*, Vol. 45, No 1:47–50, 1983.

[8] C. W. Clark. Mathematical models in the economics of renewable resources. *SIAM Review*, Vol. 21:81–99, 1979.

[9] H. Cox. On the estimation of the state variables and parameters for noisy dynamic systems. *IEEE transactions on automatic control*, Vol. 9, No 1:5–12, 1964.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from imcomplete data via the EM algorithm. *Journal of the royal statistical society*, 39, Issue 1, Series B:1–38, 1977.

[11] A. Doucet, N. de Freitas, and N. Gordon. *Sequential and Monte Carlo Methods in Practice*. Springer Verlag, 2001.

[12] Z. Ghaharamani and S. Roweis. Learning nonlinear dynamical systems using the expectation maximization algorithm. *Advances in neural information processing systems, Cambridge, MA: MIT Press*, Vol. 11:431–437, 1999.

[13] S. Gibson and B. M. Ninness. Maximum likelihood identification of bilinear systems. In *15th IFAC world congress, Barcelona, Spain*, July 2002.

[14] S. Godsill, A. Doucet, and M. West. Maximum a posteriori sequence estimation using monte carlo particle filters. *Ann. Inst. Statist. Math*, Vol. 53, No 1:82–96, 2001.

[15] G. C. Goodwin and A. Feuer. Estimation with missing data. In *International Congress on Modelling and Simulation*, November 1995.

[16] G. C. Goodwin and A. Feuer. Estimation with missing data. *Mathematical Modelling of Systems*, Vol. 5, No 3, 1999.

[17] G. C. Goodwin and R. Payne. *Dynamic System Identification: Experiment design and data analysis*. Academic Press, 1977.

[18] P. J. Green. On the use of the EM algorithm for penalized likelihood estimation. *Journal of the royal statistical society*, Vol. 52, No 3:443–452, 1990.

[19] K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the royal statistical society*, Series B, Vol. 57, No 2:425–437, 1995.

[20] L. Ljung. *System Identification: Theory for the user*. Prentice Hall, 2nd edition, 1999.

[21] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley, 1997.

[22] X. L. Meng and D. van Dyk. The EM algorithm– an old folk-song sung to a fast new tune. *Journal of the royal statistical society*, Vol. 59, No 3:511–567, 1997.

[23] B. M. Ninness and S. Gibson. Robust and simple algorithms for maximum likelihood estimation of multivariable systems. In *15th IFAC world congress, Barcelona, Spain*, July 2002.

[24] C. R. Rao. *Linear Statistical Inference and its Applications*. Wiley, New York, 1965.

[25] C. V. Rao, J. B. Rawlings, and D. Q. Mayne. Constrained state estimation for nonlinear discrete time systems: stability and moving horizon approximations. *IEEE transactions on automatic control*, Vol. 48, No 2:246–258, 2003.

[26] A. Roweis and Z. Ghaharamani. *Learning nonlinear dynamical systems using the expectation maximization algorithm*. Kalman Filtering and Neural Networks (S. Haykin, ed.), John Wiley and Sons, 2001.

[27] R. Shumway and D. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. In *Technical report 27, Division of Statistics, University of California, Davis*, 1981.

[28] V. Solo. An EM algorithm for singular state space models. *Proceedings of the 42th IEEE Conference on Decision and Control*, pages 3457–3460, 2003.

[29] V. Solo. An EM algorithm for singular state space models: II. *Proceedings of the 43rd IEEE Conference on Decision and Control*, pages 3611–3612, 2004.

[30] A. Stuart and J. K. Ord. *Kendall's advanced theory of statistics*, volume 1. Edward Arnold, 1994.

[31] M. A. Tanner. *Tools for statistical inference*. Springer, 1993.

[32] D. M. Titterington. Recursive parameter estimation using incomplete data. *Journal of the royal statistical society*, Vol. 46, No 2:257–267, 1984.

[33] S. J. Turnovski. Applications of continuous-time stochastic methods to models of endogenous economic growth. *Annual reviews in control*, Vol. 20:155–166, 1996.

[34] B. Wahlberg, L. Ljung, and T Söderström. On sampling of continuous time stochastic processes. *Control theory and advanced technology*, Vol. 9:99–112, 1993.

[35] G. C. G. Wei and M. Tanner. A Monte-Carlo implementation of the EM algortihm and the poor man's data augmentation algortihms. *Journal of the american statistical association*, Vol. 85, No 411:699–704, 1990.

[36] C. F. J. Wu. On the convergence properties of the EM algorithm. *The annals of statistics*, 11, No 1:95–103, 1983.

[37] J. Yuz and G. C. Goodwin. On sampled data models for nonlinear systems. *IEEE transactions on automatic control*, (October), 2005.