Expectation Maximization Based Parameter Estimation by Sigma-Point and Particle Smoothing

Juho Kokkala
Department of Biomedical Engineering
and Computational Science
Aalto University
Espoo, Finland
Email: juho.kokkala@aalto.fi

Arno Solin
Department of Biomedical Engineering
and Computational Science
Aalto University
Espoo, Finland
Email: arno.solin@aalto.fi

Simo Särkkä
Department of Biomedical Engineering
and Computational Science
Aalto University
Espoo, Finland
Email: simo.sarkka@aalto.fi

***Abstract*—We consider parameter estimation in non-linear state space models by using expectation–maximization based numerical approximations to likelihood maximization. We present a unified view of approximative EM algorithms that use either sigma-point or particle smoothers to evaluate the integrals involved in the expectation step of the EM method, and compare these methods to direct likelihood maximization. For models that are linear in parameters and have additive noise, we show how the maximization step of the EM algorithm is available in closed form. We compare the methods using simulated data, and discuss the differences between the approximations.**

I. Introduction

We consider state space models of the following form:

$$\begin{aligned} \mathbf{x}_k &= \mathbf{f}(\mathbf{x}_{k-1}, \boldsymbol{\theta}) + \mathbf{q}_{k-1} \\ \mathbf{y}_k &= \mathbf{h}(\mathbf{x}_k, \boldsymbol{\theta}) + \mathbf{r}_k, \end{aligned} \tag{1}$$

where the initial point is given by $\mathbf{x}_0 \sim p(\mathbf{x}_0)$, and $\mathbf{x}_k$ is the discrete-time state sequence, $\mathbf{y}_k$ is the measurement sequence, $\mathbf{q}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ is the process noise sequence, and $\mathbf{r}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ is the measurement error sequence. When the parameters $\boldsymbol{\theta}$ are known, the *filtering* problem is to compute the posterior distribution of the state $\mathbf{x}_k$ at time step k given the history of the measurements up to the time step, $p(\mathbf{x}_k \mid \mathbf{y}_{1:k})$, and the corresponding *smoothing* problem is to compute the posterior distribution $p(\mathbf{x}_k \mid \mathbf{y}_{1:T})$ of the state $\mathbf{x}_k$ at time step k given all the measurements.

In the general case, analytical solutions to the filtering and smoothing problems are not available and one has to resort to approximative numeric algorithms. In this paper, we employ two types of approximative filtering and smoothing algorithms: (i) sigma-point methods (see [1]–[7]) where discrete sigma-points are used to form Gaussian approximations of the filtering and smoothing distributions, and (ii) particle filters [8]–[10] where the filtering and smoothing distributions of the states are approximated by discrete distributions. For a general overview of Bayesian filtering and smoothing, see [11].

In this paper, we focus on estimating the parameters $\boldsymbol{\theta}$ of the model (1) using likelihood-based approaches, where the likelihood $p(\mathbf{y}_{1:T} \mid \boldsymbol{\theta})$ is maximized. When direct maximization of the likelihood is not feasible, expectation–maximization (EM, see [12]–[15]) may be used to approximately maximize the likelihood. The EM algorithm consists of iterating the expectation step (E), which computes a bound for the log-likelihood function using the current parameter values, and the maximization step (M), where the bound is maximized with respect to parameters. In the E-step, one needs to evaluate the density of the latent variables, $p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T}, \boldsymbol{\theta})$, which in the case of state space models corresponds to solving the smoothing problem.

When an exact solution to the smoothing problem is not available, one may employ approximative smoothing algorithms for the E-step. In this context, Schön *et al.* [16] used particle smoothers. The use of sigma-point smoothers has been suggested by Väänänen [17] as well as by Gašperin and Juričić [18], who used the unscented transform and compared their approach to particle smoother EM. In this paper, we develop a unified view of EM algorithms based on either sigma-point or particle smoothers and compare these approximative EM algorithms to direct likelihood maximization. Furthermore, for models that are linear-in-parameters with additive Gaussian noise, we explicitly show how to perform the maximization in the M-step analytically.

In Bayesian inference, the interest lies in posterior distributions $p(\boldsymbol{\theta} \mid \mathbf{y}_{1:T})$ of the parameters instead of maximum likelihood point estimates. The EM algorithm can be extended to computing maximum *a posteriori* estimates by modifying the M-step to maximize the posterior instead of the likelihood. For estimating the distributions, a common approach is to use Markov chain Monte Carlo samplers (MCMC, see, *e.g.*, [19]). Particle MCMC [20] algorithms are a special class of MCMC samplers for state space models, where particle filter algorithms are used to produce new samples of the state variables $\mathbf{x}$ within the MCMC sampler. In the experiment section of this paper, we compare the point estimates produced by the direct likelihood-based and EM methods to posterior distributions computed by particle MCMC.

This paper is structured as follows. The schemes for sigma-point based (Sec. II) and particle filtering based (Sec. III) non-linear filtering and smoothing are gone through in brief. Direct likelihood-based parameter estimation is covered in Section IV. We revise the EM algorithm in the context of state space models (Sec. V). The linear-in-parameters case is explicitly written out separately. Section VII is dedicated to comparisons between sigma-point and particle EM in a highly non-linear one-dimensional example and a high-dimensional coordinated turn model. Finally, the results are discussed.

II. Sigma-Point Filtering and Smoothing

In assumed density Gaussian filtering (see [3], [7], [11]), the idea is to assume that the filtering distribution is approximately Gaussian. That is, we assume that there exist means $\mathbf{m}_{k|k}$ and covariances $\mathbf{P}_{k|k}$ such that

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k}) \approx \mathcal{N}(\mathbf{x}_k \mid \mathbf{m}_{k|k}, \mathbf{P}_{k|k}). \tag{2}$$

The filtering equations of the resulting Gaussian filter [3], [5] consist of the **prediction step** and the **update step**. The prediction step for the state mean and covariance is:

$$\begin{aligned}
\mathbf{m}_{k|k-1} &= \mathbb{E}[\mathbf{f}(\mathbf{x}_{k-1})], \\
\mathbf{P}_{k|k-1} &= \mathbb{E}[(\mathbf{f}(\mathbf{x}_{k-1}) - \mathbf{m}_{k|k-1}) \\
&\qquad \times (\mathbf{f}(\mathbf{x}_{k-1}) - \mathbf{m}_{k|k-1})^{\mathsf{T}}] + \mathbf{Q},
\end{aligned} \tag{3}$$

where all the expectations are taken with respect to the distribution $\mathbf{x}_{k-1} \sim \mathcal{N}(\mathbf{m}_{k-1|k-1}, \mathbf{P}_{k-1|k-1})$. The corresponding update step for the mean and covariance of the state distribution given the measurement $\mathbf{y}_k$ is:

$$\begin{aligned}
\boldsymbol{\mu}_k &= \mathbb{E}[\mathbf{h}(\mathbf{x}_k)], \\
\mathbf{S}_k &= \mathbb{E}[(\mathbf{h}(\mathbf{x}_k) - \boldsymbol{\mu}_k)\,(\mathbf{h}(\mathbf{x}_k) - \boldsymbol{\mu}_k)^{\mathsf{T}}] + \mathbf{R}, \\
\mathbf{C}_k &= \mathbb{E}[(\mathbf{x}_k - \mathbf{m}_{k|k-1})\,(\mathbf{h}(\mathbf{x}_k) - \boldsymbol{\mu}_k)^{\mathsf{T}}], \\
\mathbf{K}_k &= \mathbf{C}_k\,\mathbf{S}_k^{-1}, \\
\mathbf{m}_{k|k} &= \mathbf{m}_{k|k-1} + \mathbf{K}_k\,(\mathbf{y}_k - \boldsymbol{\mu}_k), \\
\mathbf{P}_{k|k} &= \mathbf{P}_{k|k-1} - \mathbf{K}_k\,\mathbf{S}_k\,\mathbf{K}_k^{\mathsf{T}},
\end{aligned} \tag{4}$$

where all the expectations are taken with respect to the distribution $\mathbf{x}_k \sim \mathcal{N}(\mathbf{m}_{k|k-1}, \mathbf{P}_{k|k-1})$. The **backward pass** can be calculated by the Rauch–Tung–Striebel smoother as follows:

$$\begin{aligned}
\mathbf{m}_{k+1|k} &= \mathbb{E}[\mathbf{f}(\mathbf{x}_k)], \\
\mathbf{P}_{k+1|k} &= \mathbb{E}[(\mathbf{f}(\mathbf{x}_k) - \mathbf{m}_{k+1|k}) \\
&\qquad \times (\mathbf{f}(\mathbf{x}_k) - \mathbf{m}_{k+1|k})^{\mathsf{T}}] + \mathbf{Q}, \\
\mathbf{D}_{k+1} &= \mathbb{E}[(\mathbf{x}_k - \mathbf{m}_{k|k})\,(\mathbf{f}(\mathbf{x}_k) - \mathbf{m}_{k+1|k})^{\mathsf{T}}], \\
\mathbf{G}_k &= \mathbf{D}_{k+1}\,[\mathbf{P}_{k+1|T}]^{-1}, \\
\mathbf{m}_{k|T} &= \mathbf{m}_{k|k} + \mathbf{G}_k\,(\mathbf{m}_{k+1|T} - \mathbf{m}_{k+1|k}), \\
\mathbf{P}_{k|T} &= \mathbf{P}_{k|k} - \mathbf{G}_k\,(\mathbf{P}_{k+1|T} - \mathbf{P}_{k+1|k})\,\mathbf{G}_k^{\mathsf{T}},
\end{aligned} \tag{5}$$

where all the expectations are taken with respect to the distribution $\mathbf{x}_k \sim \mathcal{N}(\mathbf{m}_{k|k}, \mathbf{P}_{k|k})$.

For implementing the EM algorithm, we need the smoothing distribution $p(\mathbf{x}_k \mid \mathbf{y}_{1:T}) \approx \mathcal{N}(\mathbf{x} \mid \mathbf{m}_{k|T}, \mathbf{P}_{k|T})$. Additionally, we also need the pairwise smoothing distributions, which can be expressed in terms of the above results:

$$\begin{aligned}
&p(\mathbf{x}_k, \mathbf{x}_{k-1} \mid \mathbf{y}_{1:T}) \approx \\
&\mathcal{N}\left(\begin{pmatrix}\mathbf{x}_k \\ \mathbf{x}_{k-1}\end{pmatrix} \middle| \begin{pmatrix}\mathbf{m}_{k|T} \\ \mathbf{m}_{k-1|T}\end{pmatrix}, \begin{pmatrix}\mathbf{P}_{k|T} & \mathbf{P}_{k|T}\mathbf{G}_{k-1}^{\mathsf{T}} \\ \mathbf{G}_{k-1}\mathbf{P}_{k|T} & \mathbf{P}_{k-1|T}\end{pmatrix}\right).
\end{aligned} \tag{6}$$

EM algorithms appearing in literature (see, *e.g.*, [13], [15]) sometimes suggest a separate recursion to be carried out for the cross-terms. The formulation in Equation (6) makes any additional recursions unnecessary, as the terms can be evaluated directly during the smoother backward pass.

A. Sigma-Points in Cubature Integration

The state mean and covariance is dependent upon computing Gaussian integrals of the form:

$$\mathbb{E}[\mathbf{f}(\mathbf{x})] = \int_{\mathbb{R}^n} \mathbf{f}(\mathbf{x})\,\mathcal{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{P})\,\mathrm{d}\mathbf{x}, \tag{7}$$

where $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^d$ and $\mathcal{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{P})$ is a multi-dimensional Gaussian density with mean $\mathbf{m}$ and covariance matrix $\mathbf{P}$. One way of computing these integrals is by using multidimensional generalizations of Gaussian quadratures, which are sometimes referred to as Gaussian *cubatures* [21], which give approximations of the form

$$\mathbb{E}[\mathbf{f}(\mathbf{x})] \approx \sum_i w_i\,\mathbf{f}(\mathbf{x}_i), \tag{8}$$

where the weights w_i and the sigma-points $\mathbf{x}_i$ are functions of the mean $\mathbf{m}$ and covariance $\mathbf{P}$ of the Gaussian term. The sigma-points are selected as follows:

$$\mathbf{x}_i = \mathbf{m} + \mathbf{L}\,\boldsymbol{\xi}_i, \tag{9}$$

where $\boldsymbol{\xi}_i$ are method specific unit sigma-points, and $\mathbf{L}$ is a matrix square-root factor such that $\mathbf{P} = \mathbf{L}\,\mathbf{L}^{\mathsf{T}}$. The differences in the methods come from different choices of weights and unit sigma-points, where common methods are based on the Gauss–Hermite quadrature [3], [5] and the unscented transform [1], [2].

III. Particle Filtering and Smoothing

In particle filtering [9], also called sequential importance resampling, the filtering distribution $p(\mathbf{x}_k \mid \mathbf{y}_{1:k})$ is approximated by a weighted set of discrete particles $\{(w_k^{(i)}, \tilde{\mathbf{x}}_k^{(i)}) : i = 1, \ldots, N\}$, which corresponds to the approximation

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta(\mathbf{x}_k - \tilde{\mathbf{x}}_k^{(i)}), \tag{10}$$

where N is the number of particles and $\delta(\cdot)$ is the Dirac delta function. In the particle filter, the prediction step is replaced by sampling new particle values according to an importance distribution:

$$\tilde{\mathbf{x}}_{k+1}^{(i)} \sim \pi(\mathbf{x}_{k+1} \mid \tilde{\mathbf{x}}_k^{(i)}, \mathbf{y}_{k+1}), \tag{11}$$

and the update step consists of updating the particle weights based on the dynamic model and measurement likelihoods as follows:

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\mathbf{y}_k \mid \tilde{\mathbf{x}}_k^{(i)}) p(\tilde{\mathbf{x}}_k^{(i)} \mid \tilde{\mathbf{x}}_{k-1}^{(i)})}{\pi(\tilde{\mathbf{x}_k} \mid \tilde{\mathbf{x}}_{k-1}^{(i)}, \mathbf{y}_k)}. \tag{12}$$

To avoid the so-called degeneracy problem where only one or few particles hold significant weight, the particles may be resampled after the weight update step. In resampling, new particles $\tilde{\mathbf{x}}_k$ are sampled from the old particles with probabilities $w_k^{(i)}$, and the weights are reset to $1/N$. In adaptive resampling, the resampling step is performed only if the effective sample size (see [22]) is below a threshold. A particle smoother aims to form a similar discrete approximation of the smoothing distribution:

$$p(\mathbf{x}_k \mid \mathbf{y}_{1:T}) \approx \sum_{i=1}^{N} w_{k|T}^{(i)} \delta(\mathbf{x}_k - \tilde{\mathbf{x}}_k^{(i)}). \tag{13}$$

The basic particle smoother proposed by Hürzeler and Künsch [8] as well as by Doucet *et al.* [9] computes the smoothing weights from the filtering weights in a backward pass while preserving the particle values $\tilde{\mathbf{x}}$. The equation for computing the smoothing weights is

$$w_{k|T} = \sum_{j=1}^{N} w_{k+1|N}^{(j)} \frac{w_k^{(i)} p(\tilde{\mathbf{x}}_{k+1}^{(j)} \mid \tilde{\mathbf{x}}_k^{(i)})}{\sum_{l=1}^{N} w_k^{(l)} p(\tilde{\mathbf{x}}_{k+1}^{(j)} \mid \tilde{\mathbf{x}}_k^{(l)})}. \tag{14}$$

In EM, we also need an approximation for the pairwise joint distribution of consecutive states:

$$p(\mathbf{x}_{k+1}, \mathbf{x}_k \mid \mathbf{y}_{1:T}) \approx \sum_{i=1}^{N} \sum_{j=1}^{N} w_{k|T}^{(ij)} \delta(\mathbf{x}_k - \tilde{\mathbf{x}}_k^{(i)}) \delta(\mathbf{x}_{k+1} - \tilde{\mathbf{x}}_{k+1}^{(j)}). \tag{15}$$

Schön *et al.* [16] show that weights for the pairwise joint distribution can be obtained as follows:

$$w_{k|T}^{(ij)} = w_{k+1|T}^{(j)} \frac{w_k^{(i)} p(\tilde{\mathbf{x}}_{k+1}^{(j)} \mid \tilde{\mathbf{x}}_k^{(i)})}{\sum_{l=1}^{N} w_k^{(l)} p(\tilde{\mathbf{x}}_{k+1}^{(j)} \mid \tilde{\mathbf{x}}_k^{(l)})}. \tag{16}$$

Furthermore, these weights can be easily obtained as an intermediate step during the particle smoother.

IV. Direct Likelihood-Based Parameter Estimation

We use the term 'direct likelihood-based' for parameter estimation schemes that directly utilize the (approximate) likelihood returned by the non-linear filtering scheme. This is contrary to 'indirect likelihood' schemes which only aim to approximate this likelihood (such as the EM algorithm, see Section V). Direct likelihood-based estimation maximizes the log-likelihood,

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \log p(\mathbf{y}_{1:T} \mid \boldsymbol{\theta}), \tag{17}$$

or in case of maximum *a posteriori* estimation:

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} \left(\log p(\mathbf{y}_{1:T} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \right). \tag{18}$$

The log-likelihood may be approximated based on the filtering outcome (see Eq. 4):

$$\begin{aligned} &\log p(\mathbf{y}_{1:T} \mid \boldsymbol{\theta}) \approx \\ &- \frac{1}{2} \sum_{k=1}^{T} \log |2\pi \mathbf{S}_k| - \frac{1}{2} \sum_{k=1}^{T} (\mathbf{y}_k - \boldsymbol{\mu}_k)^{\mathsf{T}} \mathbf{S}_k^{-1} (\mathbf{y}_k - \boldsymbol{\mu}_k). \end{aligned} \tag{19}$$

Furthermore, the gradient of the log-likelihood can be written in terms of the filtering equations (see, *e.g.*, [11], [23]) allowing the use of efficient conjugate-gradient optimization algorithms. Alternatively, Fisher's identity (see [11], [16]), which links the gradient to the expectation of the gradient of the complete data log likelihood, may be used. In addition, particle filters may also be used to approximate the likelihood and its gradient (see [10]).

V. Expectation–Maximization Based Parameter Estimation

When the direct likelihood based approach is not feasible, the indirect approach by expectation–maximization allows us to use the smoothing scheme for maximizing the likelihood without explicitly using it as the optimization target function. Expectation–maximization is an iterative algorithm for finding maximum likelihood (or maximum *a posteriori*) parameter estimates in a setting with some unobserved variables, in state space context the state variables $\mathbf{x}$. In the following, we use the formulation by Neal and Hinton [14] and notation of Schön *et al.* [16].

A. The EM Algorithm

The EM algorithm is based on the following lower bound of the log-likelihood:

$$\log p(\mathbf{y}_{1:T} \mid \boldsymbol{\theta}) \geq \int q(\mathbf{x}_{0:T}) \log \frac{p(\mathbf{x}_{0:T}, \mathbf{y}_{1:T} \mid \boldsymbol{\theta})}{q(\mathbf{x}_{0:T})} \, \mathrm{d}\mathbf{x}_{0:T} \tag{20}$$

with an arbitrary distribution q. The algorithm consists of iteratively maximizing the lower bound with respect to q (holding $\boldsymbol{\theta}$ fixed) and with respect to $\boldsymbol{\theta}$ (holding q fixed). Furthermore, when $\boldsymbol{\theta} = \boldsymbol{\theta}^{(n)}$ is fixed, the maximum with respect to q is obtained by

$$q(\mathbf{x}_{0:T}) = p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T}, \boldsymbol{\theta}^{(n)}), \tag{21}$$

and thus the bound equals

$$\begin{aligned} &\int p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T}, \boldsymbol{\theta}^{(n)}) \log \frac{p(\mathbf{x}_{0:T}, \mathbf{y}_{1:T} \mid \boldsymbol{\theta})}{p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T}, \boldsymbol{\theta}^{(n)})} \, \mathrm{d}\mathbf{x}_{0:T} \\ &= \int p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T}, \boldsymbol{\theta}^{(n)}) \log p(\mathbf{x}_{0:T}, \mathbf{y}_{1:T} \mid \boldsymbol{\theta}) \, \mathrm{d}\mathbf{x}_{0:T} \\ &\quad - \int p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T}, \boldsymbol{\theta}^{(n)}) \log p(\mathbf{x}_{0:T} \mid \mathbf{y}_{1:T}, \boldsymbol{\theta}^{(n)}) \, \mathrm{d}\mathbf{x}_{0:T}. \end{aligned}$$

The latter term is independent of $\boldsymbol{\theta}$ and may thus be omitted when maximizing with respect to $\boldsymbol{\theta}$. The first term is the conditional expectation of $\log p(\mathbf{x}_{0:T}, \mathbf{y}_{1:T} \mid \boldsymbol{\theta})$ with respect to $\mathbf{y}$ and $\boldsymbol{\theta}^{(n)}$, denoted by

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)}) = \mathbb{E}[\log p(\mathbf{x}_{0:T}, \mathbf{y}_{1:T} \mid \boldsymbol{\theta}) \mid \mathbf{y}_{1:T}, \boldsymbol{\theta}^{(n)}]. \tag{22}$$

Thus, maximization of Equation (20) with respect to q corresponds to computing $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)})$, which is called the expectation step, and maximization of Equation (20) with respect to $\boldsymbol{\theta}$ corresponds to maximizing $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)})$ with respect to $\boldsymbol{\theta}$, which is called the maximization step. The resulting EM algorithm consists of initializing the parameters to $\boldsymbol{\theta}^{(0)}$ and for $n = 0, 1, \ldots$ iterating the following two steps:

- E-step: compute $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)})$.
- M-step: $\boldsymbol{\theta}^{(n+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)})$.

If a maximum *a posteriori* estimate is desired, maximization of $\mathcal{Q}$ in the M-step is replaced by maximization of $\mathcal{Q} + \log p(\boldsymbol{\theta})$. Due to the Markov property of state space models, $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)})$ further factorizes such that $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)}) = \mathrm{I}_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)}) + \mathrm{I}_2(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)}) + \mathrm{I}_3(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)})$, where

$$\mathrm{I}_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)}) = \mathbb{E}[\log p(\mathbf{x}_0 \mid \boldsymbol{\theta}) \mid \mathbf{y}_{1:T}, \boldsymbol{\theta}^{(n)}], \tag{23}$$

$$\mathrm{I}_2(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)}) = \sum_{k=1}^{T} \mathbb{E}[\log p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) \mid \mathbf{y}_{1:T}, \boldsymbol{\theta}^{(n)}], \tag{24}$$

$$\mathrm{I}_3(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)}) = \sum_{k=1}^{T} \mathbb{E}[\log p(\mathbf{y}_k \mid \mathbf{x}_k) \mid \mathbf{y}_{1:T}, \boldsymbol{\theta}^{(n)}]. \tag{25}$$

Hence, computation of $\mathcal{Q}$ requires the smoothing distributions $p(\mathbf{x}_t \mid \mathbf{y}_{1:T}, \boldsymbol{\theta}^{(n)})$ and the joint smoothing distributions $p(\mathbf{x}_k, \mathbf{x}_{k+1} \mid \mathbf{y}_{1:T}, \boldsymbol{\theta}^{(n)})$. As these are not available analytically in the general case, one has to resort to approximations. In the following, we review how to approximate $\mathcal{Q}$ either by sigma-point based smoothers or by particle smoothers.

Figure 1 visualizes the likelihood curve and the corresponding EM bound approximation for two iteration steps. The approximative nature of EM is clearly visible in the dashed lines, while the particle filter evaluated likelihood reminds that even the sigma-point likelihood is just an approximation (for details, see Sec. VII-A).

B. Approximating the E-Step Using Sigma-Point Smoothers

In terms of sigma-point smoothing outcomes, the expression for $\mathcal{Q}$ for the non-linear state space model can be written as

$$\begin{aligned}
&\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)}) \\
&\approx -\frac{1}{2} \log |2\pi\, \mathbf{P}_0| - \frac{T}{2} \log |2\pi\, \mathbf{Q}| - \frac{T}{2} \log |2\pi\, \mathbf{R}| \\
&- \frac{1}{2} \operatorname{tr} \left\{ \mathbf{P}_0^{-1} \left[\mathbf{P}_{0|T} + (\mathbf{m}_{0|T} - \mathbf{m}_0)\, (\mathbf{m}_{0|T} - \mathbf{m}_0)^{\mathsf{T}} \right] \right\} \\
&- \frac{1}{2} \sum_{k=1}^{T} \operatorname{tr} \left\{ \mathbf{Q}^{-1}\, \mathbb{E}\big[(\mathbf{x}_k - \mathbf{f}(\mathbf{x}_{k-1})(\mathbf{x}_k - \mathbf{f}(\mathbf{x}_{k-1})^{\mathsf{T}} \mid \mathbf{y}_{1:T} \big] \right\} \\
&- \frac{1}{2} \sum_{k=1}^{T} \operatorname{tr} \left\{ \mathbf{R}^{-1}\, \mathbb{E}\big[(\mathbf{y}_k - \mathbf{h}(\mathbf{x}_k,)(\mathbf{y}_k - \mathbf{h}(\mathbf{x}_k,)^{\mathsf{T}} \mid \mathbf{y}_{1:T} \big] \right\},
\end{aligned} \tag{26}$$

where the model parameters are set to $\boldsymbol{\theta}$ and the expectations are over the distributions obtained from the Gaussian smoother with parameters $\boldsymbol{\theta}^{(n)}$. In practice, we can approximate the Gaussian smoother and the Gaussian integrals above as described in Section II.

C. Approximating the E-Step Using Particle Smoothers

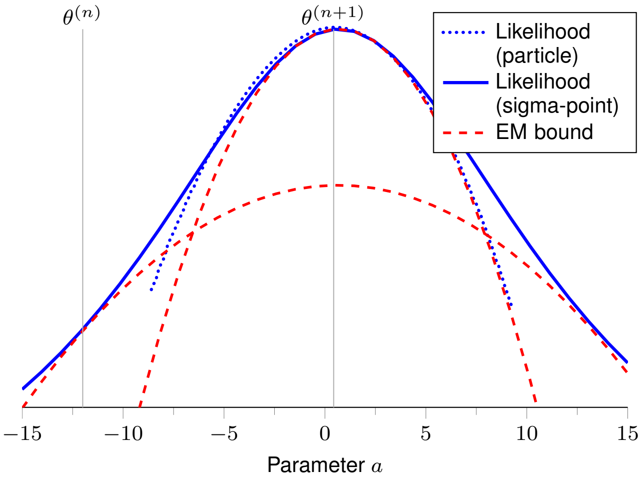

Fig. 1. Visualization of the one-step evolution of the EM algorithm for the univariate estimation of parameter a. The dotted line represents the particle filter likelihood estimate, while the solid line is the sigma-point filter likelihood approximation. The dashed lines correspond to the sigma-point EM bounds for iterations n and $n+1$.

The use of particle smoothers to approximate the function $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)})$ was proposed by Schön *et al.* [16]. From Equations (23–25), we obtain

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)}) &\approx \sum_{i=1}^{N} w_{0|T}^{(i)} \log p(\tilde{\mathbf{x}}_0^{(i)} \mid \boldsymbol{\theta}) \\
&+ \sum_{k=0}^{T-1} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{k|T}^{(ij)} \log p(\tilde{\mathbf{x}}_{k+1}^{(j)} \mid \tilde{\mathbf{x}}_k^{(i)}, \boldsymbol{\theta}) \\
&+ \sum_{k=1}^{T} \sum_{i=1}^{N} w_{k|T}^{(i)} \log p(\mathbf{y}_k \mid \tilde{\mathbf{x}}_k^{(i)}, \boldsymbol{\theta}),
\end{aligned} \tag{27}$$

where the weighted discrete approximation for the smoothing distributions (Eq. 13) and the pairwise joint smoothing distributions (Eq. 15) are obtained by running a particle filter and smoother (see Sec. III) over the data conditional on parameters $\boldsymbol{\theta}^{(n)}$.

VI. Evaluation of the M-Step for Linear-in-Parameters Models

In general, maximizing $\mathcal{Q}$ in the M-step requires a numerical optimization algorithm. However, if the parameters appear as linear coefficients inside the non-linear dynamic and measurement models, an analytical solution can be obtained. In practice, this type of models may occur in estimation of scale parameters or if the model functions are linear combinations of nonlinear functions. These linear-in-parameter models can be represented as follows:

$$\begin{aligned}
\mathbf{f}(\mathbf{x}, \boldsymbol{\theta}) &= \mathbf{A}(\boldsymbol{\theta})\, \tilde{\mathbf{f}}(\mathbf{x}), \\
\mathbf{h}(\mathbf{x}, \boldsymbol{\theta}) &= \mathbf{H}(\boldsymbol{\theta})\, \tilde{\mathbf{h}}(\mathbf{x}),
\end{aligned} \tag{28}$$

where the model parameters appear in the matrix coefficients $\mathbf{A}$ and $\mathbf{H}$ only, and the functions $\tilde{\mathbf{f}}(\mathbf{x})$ and $\tilde{\mathbf{h}}(\mathbf{x})$ hold the non-linearities. The expression of $\mathcal{Q}$ for this type of linear-in-parameters model with additive noise can be written as:

$$\begin{aligned}
&\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)}) = \\
&-\frac{1}{2}\log|2\pi\,\mathbf{P}_0| - \frac{T}{2}\log|2\pi\,\mathbf{Q}| - \frac{T}{2}\log|2\pi\,\mathbf{R}| \\
&-\frac{1}{2}\operatorname{tr}\Big\{\mathbf{P}_0^{-1}\Big[\mathbf{P}_{0|T} + (\mathbf{m}_{0|T} - \mathbf{m}_0)\,(\mathbf{m}_{0|T} - \mathbf{m}_0)^\mathsf{T}\Big]\Big\} \\
&-\frac{T}{2}\operatorname{tr}\Big\{\mathbf{Q}^{-1}\Big[\boldsymbol{\Sigma} - \mathbf{C}\,\mathbf{A}^\mathsf{T} - \mathbf{A}\,\mathbf{C}^\mathsf{T} + \mathbf{A}\,\boldsymbol{\Phi}\,\mathbf{A}^\mathsf{T}\Big]\Big\} \\
&-\frac{T}{2}\operatorname{tr}\Big\{\mathbf{R}^{-1}\Big[\mathbf{D} - \mathbf{B}\,\mathbf{H}^\mathsf{T} - \mathbf{H}\,\mathbf{B}^\mathsf{T} + \mathbf{H}\,\boldsymbol{\Theta}\,\mathbf{H}^\mathsf{T}\Big]\Big\},
\end{aligned}$$

where the model parameters $\mathbf{A}, \mathbf{H}, \mathbf{m}_0$ are evaluated at $\boldsymbol{\theta}$ and the quantities $\boldsymbol{\Sigma}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ can be evaluated based on both the sigma-point and particle smoother results run with fixed parameter values $\boldsymbol{\theta}^{(n)}$. This generalizes the formulation that was presented for linear state space models in [11].

The convenient formulation of $\mathcal{Q}$ given above implies that if the parameters appear linearly in one of the model matrices (*e.g.*, are some subcomponents in the matrices), by setting the gradients of $\partial\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(n)})/\partial\boldsymbol{\theta}$ to zero for each $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{H}, \mathbf{Q}, \mathbf{R}, \mathbf{P}_0, \mathbf{m}_0\}$ separately, we get the following results:

- When $\boldsymbol{\theta} = \mathbf{A}$, we get $\mathbf{A}^* = \mathbf{C}\,\boldsymbol{\Phi}^{-1}$.
- When $\boldsymbol{\theta} = \mathbf{H}$, we get $\mathbf{H}^* = \mathbf{B}\,\boldsymbol{\Theta}^{-1}$.
- When $\boldsymbol{\theta} = \mathbf{Q}$, we get
$$\mathbf{Q}^* = \boldsymbol{\Sigma} - \mathbf{C}\,\mathbf{A}^\mathsf{T} - \mathbf{A}\,\mathbf{C}^\mathsf{T} + \mathbf{A}\,\boldsymbol{\Phi}\,\mathbf{A}^\mathsf{T}.$$
- When $\boldsymbol{\theta} = \mathbf{R}$, we get
$$\mathbf{R}^* = \mathbf{D} - \mathbf{H}\,\mathbf{B}^\mathsf{T} - \mathbf{B}\,\mathbf{H}^\mathsf{T} + \mathbf{H}\,\boldsymbol{\Theta}\,\mathbf{H}^\mathsf{T}.$$
- When $\boldsymbol{\theta} = \mathbf{m}_0$, we get $\mathbf{m}_0^* = \mathbf{m}_{0|T}$.
- Finally, the maximum with respect to the initial covariance $\boldsymbol{\theta} = \mathbf{P}_0$ is
$$\mathbf{P}_0^* = \mathbf{P}_{0|T} + (\mathbf{m}_{0|T} - \mathbf{m}_0)\,(\mathbf{m}_{0|T} - \mathbf{m}_0)^\mathsf{T}.$$

A. Evaluating the Quantities by Sigma-Point Smoothing

If a sigma-point smoother is used, the required matrix quantities can be evaluated by evaluating the following set of sums based on the smoothing outcome:

$$\boldsymbol{\Sigma} = \frac{1}{T}\sum_{k=1}^{T}\mathbf{P}_{k|T} + \mathbf{m}_{k|T}\,[\mathbf{m}_{k|T}]^\mathsf{T}, \tag{29}$$

$$\boldsymbol{\Phi} = \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\big[\tilde{\mathbf{f}}(\mathbf{x}_{k-1})\,\tilde{\mathbf{f}}^\mathsf{T}(\mathbf{x}_{k-1}) \mid \mathbf{y}_{1:T}\big], \tag{30}$$

$$\boldsymbol{\Theta} = \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\big[\tilde{\mathbf{h}}(\mathbf{x}_k)\,\tilde{\mathbf{h}}^\mathsf{T}(\mathbf{x}_k) \mid \mathbf{y}_{1:T}\big], \tag{31}$$

$$\mathbf{B} = \frac{1}{T}\sum_{k=1}^{T}\mathbf{y}_k\,\mathbb{E}\big[\tilde{\mathbf{h}}^\mathsf{T}(\mathbf{x}_k) \mid \mathbf{y}_{1:T}\big], \tag{32}$$

$$\mathbf{C} = \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\big[\mathbf{x}_k\,\tilde{\mathbf{f}}^\mathsf{T}(\mathbf{x}_{k-1}) \mid \mathbf{y}_{1:T}\big], \tag{33}$$

$$\mathbf{D} = \frac{1}{T}\sum_{k=1}^{T}\mathbf{y}_k\,\mathbf{y}_k^\mathsf{T}, \tag{34}$$

where $\tilde{\mathbf{f}}(\cdot), \tilde{\mathbf{h}}(\cdot)$ are the non-linear components of the dynamic and measurement model functions as defined in Equation (28). The expectations are evaluated by the Gaussian integral approximation (Eq. 7), where the densities are outcomes from the smoothing backward pass results (in Sec. II). The pairwise smoothing distributions needed for evaluating $\mathbf{C}$ are directly accessible in the smoothing outcome (see Eq. 6).

B. Evaluating the Quantities by Particle Smoothing

In the case of particle smoothing, the expectations in Equations (29–34) are evaluated as weighted sums of the particles:

$$\boldsymbol{\Sigma} = \frac{1}{T}\sum_{k=1}^{T}\sum_{i=1}^{N} w_{k|T}^{(i)}\tilde{\mathbf{x}}_k^{(i)}\,[\tilde{\mathbf{x}}_k^{(i)}]^\mathsf{T}, \tag{35}$$

$$\boldsymbol{\Phi} = \frac{1}{T}\sum_{k=1}^{T}\sum_{i=1}^{N} w_{k-1|T}^{(i)}\tilde{\mathbf{f}}(\tilde{\mathbf{x}}_{k-1}^{(i)})\,\tilde{\mathbf{f}}^\mathsf{T}(\tilde{\mathbf{x}}_{k-1}^{(i)}), \tag{36}$$

$$\boldsymbol{\Theta} = \frac{1}{T}\sum_{k=1}^{T}\sum_{i=1}^{N} w_{k|T}^{(i)}\tilde{\mathbf{h}}(\tilde{\mathbf{x}}_k^{(i)})\,\tilde{\mathbf{h}}^\mathsf{T}(\tilde{\mathbf{x}}_k^{(i)}), \tag{37}$$

$$\mathbf{B} = \frac{1}{T}\sum_{k=1}^{T}\mathbf{y}_k\sum_{i=1}^{N} w_{k|T}^{(i)}\tilde{\mathbf{h}}^\mathsf{T}(\tilde{\mathbf{x}}_k^{(i)}), \tag{38}$$

$$\mathbf{C} = \frac{1}{T}\sum_{k=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{k-1|T}^{(ij)}\tilde{\mathbf{x}}_k^{(j)}\,\tilde{\mathbf{f}}^\mathsf{T}(\tilde{\mathbf{x}}_{k-1}^{(i)}), \tag{39}$$

$$\mathbf{D} = \frac{1}{T}\sum_{k=1}^{T}\mathbf{y}_k\,\mathbf{y}_k^\mathsf{T}. \tag{40}$$

In addition to the closed-form ML solution presented here, similar expressions may be obtained for MAP estimation with a suitable conjugate prior. For example, in the one-dimensional case, scaled inverse chi-squared is used for the variances $\mathbf{Q}, \mathbf{R}$ and conditional normal prior for the coefficients $\mathbf{A}, \mathbf{H}$.

VII. Experiments

In this section, we consider two simulated examples to demonstrate the use of expectation–maximization in combination with sigma-point and particle smoothing. First, we consider a highly non-linear one-dimensional problem, for which we present results both from the direct likelihood-based approaches and the smoother-based EM schemes. The second example demonstrates the use of the smoother-based EM methods in a more practical high-dimensional tracking task.

A. Univariate Non-Stationary Growth

We consider a univariate non-stationary growth model (UNGM, [24], [25]), which is frequently used for demonstrating various filtering schemes. This linear-in-parameters dynamic model can be written as:

$$x_{k+1} = a\,x_k + b\,\frac{x_k}{1 + x_k^2} + c\,\cos(1.2k) + q_k, \tag{41}$$

where $q_k \sim \mathcal{N}(0, Q)$. The version with unknown parameters has been used in EM experiments by [16] and [18]. Usually, a quadratic measurement model is used. However, sigma-point filtering schemes are known to perform badly on this model [4]. Therefore we considered a linear observation model to be able to concentrate on comparisons between different parameter estimation schemes for estimating the model parameters. The observation model is thus given as:

$$y_k = d\,x_k + r_k, \tag{42}$$

where $r_k \sim \mathcal{N}(0, R)$. We assume the measurement scale parameter $d = 0.22$ and the initial distribution $x_0 \sim \mathcal{N}(0, 0.1^2)$ known, and estimate other parameters $\boldsymbol{\theta} = \{a, b, c, Q, R\}$. We draw ground-truth parameters from scaled inverse chi-squared priors (with the parametrisation of [19]) for the variance parameters Q and R, and conditional normal priors for $a, b,$ and c:

$$\begin{aligned} Q &\sim \text{Inv-}\chi^2(15, 10), & a \mid Q &\sim \mathcal{N}(0.5, 0.001\,Q), \\ R &\sim \text{Inv-}\chi^2(15, 1), & b \mid Q &\sim \mathcal{N}(25, 0.1\,Q), \\ & & c \mid Q &\sim \mathcal{N}(8, 0.025\,Q). \end{aligned}$$

Our data consist of 100 trajectories with $T = 100$ timesteps and the model parameters drawn from the above priors independently. The sigma-point filtering and smoothing scheme was implemented by using unscented transform [1], [2] with parameters $\kappa = 0$, $\alpha = 1$, and $\beta = 0$ (see [26]). For the particle scheme, we used 100 particles with resampling threshold 75, and the optimal importance distribution. Both EM algorithms were run for 1,000 iteration steps for each trajectory, starting from the prior scales (Q, R) and means (a, b, c). With particle EM, we also performed maximum *a posteriori* estimation, as with this conjugate prior the M-step is analytically tractable for MAP estimation. Figure 1 illustrates the connection between the direct likelihood-based approach and the indirect EM approach. The likelihood as a function of parameter a is shown for the first trajectory in the data, while holding the other parameters fixed. From the figure, we see that the sigma-point EM bounds are indeed lower bounds for the sigma-point likelihood approximation. For comparison, a likelihood curve estimated by the particle filter is also shown. Compared to the particle filter likelihood, the sigma-point filter clearly overestimates the likelihood away from the mode. Compared over all the trajectories, both EM approaches tend to converge to similar ML estimates. Figure 2 shows scatterplots of the final parameter estimates after 1,000 iterations against the ground-truth values. For comparison, also MAP estimates produced by the particle EM are shown (the MAP results for the sigma-point method are omitted to avoid crowding the plot). With both EM methods, there is a clear correlation between the estimates and the true values. The measurement noise scale R is however an exception, as there is high variation in the ML estimates, and the MAP estimates are all close to the prior scale.

TABLE I. CORRELATIONS OF THE FINAL EM ESTIMATES VERSUS THE DIRECT ML ESTIMATES FOR ALL THE TRAJECTORIES IN THE UNGM EXAMPLE.

Parameter	a	b	c	$\log Q$	$\log R$
Particle EM vs. ML	0.998	0.945	0.994	0.977	0.792
Sigma-point EM vs. ML	0.997	0.909	0.989	0.972	0.731

To compare the EM approaches to direct likelihood-based parameter estimation, we computed parameter estimates with direct maximum likelihood (Sec. IV). The likelihood was evaluated by a UKF sigma-point filter. Correlations between the direct ML estimates and the final EM estimates are shown in Table I. With one trajectory, the direct ML optimization did not converge, and thus the correlations are computed over 99 trajectories. The EM algorithms indeed converge to the maximum likelihood estimates, as the correlations are over 0.9 for all parameters except the measurement noise variance R. For all parameters, the correlations of the particle EM estimates are higher than the correlations of the sigma-point EM estimates, which shows that the particle EM approximates the true ML solution better than the sigma-point EM.

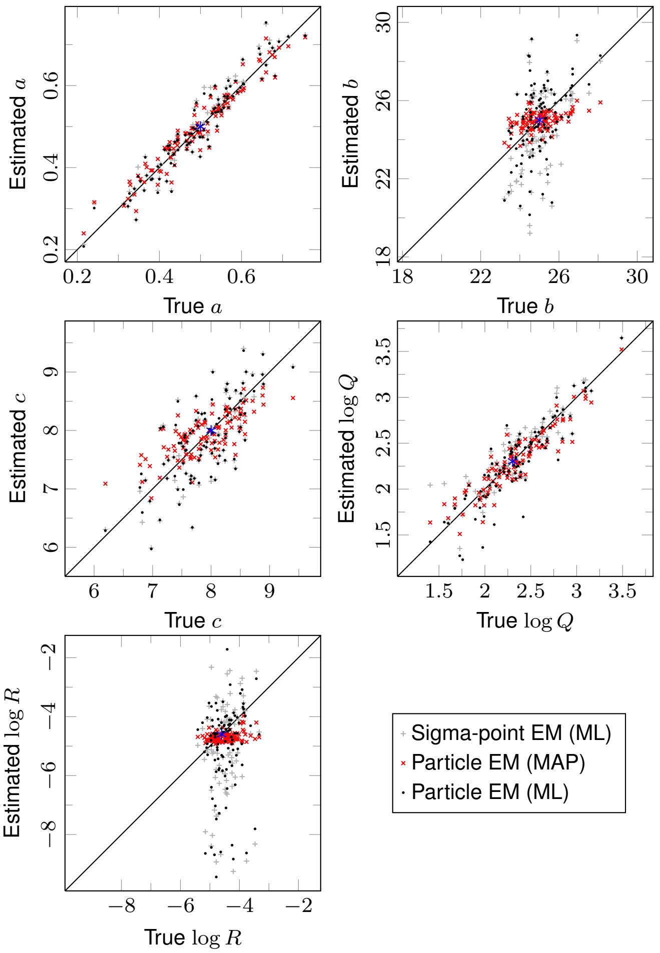

Fig. 2. The estimates of the parameters of the univariate non-stationary growth model after 1,000 EM iterations compared to ground-truth parameter values over 100 datasets with different ground-truth parameters.

For the first trajectory, we also compared the point estimates to the posterior distributions. We used the particle marginal Metropolis–Hastings variant of particle MCMC [20] to sample from the posterior. Figure 3 shows the histograms of samples from the marginal posterior distributions for all parameters, as well as the corresponding point estimates. The point estimates mostly fall within intervals with high posterior density, and the particle EM MAP estimates are close to the histogram peaks.

B. Tracking of a Manouvering Target

As a more practical second example we consider a tracking problem, where the dynamics are described by a coordinated

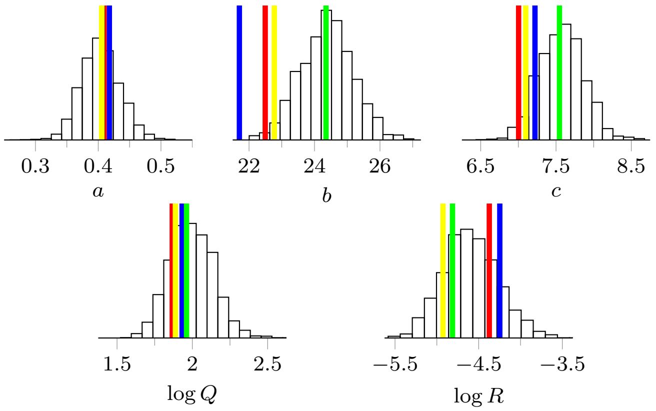

Fig. 3. Comparison of direct ML and EM-based point estimates to samples from marginal posterior distributions of the parameters. Sigma-point EM (🟦), direct ML (🟥), particle EM MAP (🟩), and particle EM ML (🟨).

turn model and the the observations by bearings only sensor measurements (see, [7], [26]–[28]).

The state $\mathbf{x} = (\mathrm{x}_1, \dot{\mathrm{x}}_1, \mathrm{x}_2, \dot{\mathrm{x}}_2, \boldsymbol{\omega})$ is five-dimensional and we consider bearings-only observations with two distinct sensor locations for the observations. We used the formulation of the model that is described in [7], [28]. The interest lies in estimating the dynamic noise ($\mathbf{q}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$) and measurement noise ($\mathbf{r}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$) covariances. In our case, the ground-truth dynamic model noise covariance was

$$\mathbf{Q} = \begin{pmatrix} q_\mathrm{c}\Delta t^3/3 & 0 & q_\mathrm{c}\Delta t^2/2 & 0 & 0 \\ 0 & q_\mathrm{c}\Delta t^3/3 & 0 & q_\mathrm{c}\Delta t^2/2 & 0 \\ q_\mathrm{c}\Delta t^2/2 & 0 & q_\mathrm{c}\Delta t & 0 & 0 \\ 0 & q_\mathrm{c}\Delta t^2/2 & 0 & q_\mathrm{c}\Delta t & 0 \\ 0 & 0 & 0 & 0 & q_\omega \end{pmatrix},$$

where $\Delta t = 0.01$, $q_\mathrm{c} = 0.1$ and $q_\omega = 0.01$. The ground-truth measurement noise covariance was $\mathbf{R} = \mathrm{diag}(0.05^2,\ 0.1^2)$. The sensor locations were $(-1, 0.5)$ and $(1, 1)$. The parameters of the initial distribution were $\mathbf{m}_0 = \mathbf{0}$ and $\mathbf{P}_0 = \mathrm{diag}(0.1^2,\ 0.1^2,\ 0.1^2,\ 0.1^2,\ 0.01^2)$. We simulated 100 different trajectories with $T = 100$ timesteps from the model and estimated the dynamical and measurement noise covariance matrices $\mathbf{Q}$ and $\mathbf{R}$, with all the other model parameters assumed known. The $\mathbf{Q}$ and $\mathbf{R}$ were initialized randomly to diagonal matrices where the square-roots of the diagonal elements were sampled independently from $\mathrm{U}(0, 1)$. We used a bootstrap filter with 100 particles in the particle EM and the unscented transform [1], [2] with $\kappa = 0$, $\alpha = 1$, and $\beta = 0$ (see [26]) in the sigma-point EM. Both algorithms were ran for 100 iterations for each data trajectory.

For visualizing the convergence of the noise covariance matrices, we use the Kullback–Leibler (KL) divergence (see, *e.g.*, [19]), which is a non-symmetric measure of the difference between two probability distributions. We plot the KL divergence between the true and estimated process distributions both for the dynamic and measurement noise. The divergence $\mathrm{D}_{\mathrm{KL}}\left[\mathcal{N}_\mathrm{q} \| \mathcal{N}_{\tilde{\mathrm{q}}}\right]$ is thus a measure of the information lost when $\mathcal{N}_{\tilde{\mathrm{q}}}$ is used to approximate $\mathcal{N}_\mathrm{q}$. The evolution of the log-KL divergences using both EM methods are shown in Figure 4.

To demonstrate the initial convergence, we initialized the covariances to bear little resemblance to their true values. This demonstrates the initial convergence of the methods. The convergence of the particle EM as a function of iterations is faster. The noise covariance estimate converges close to the correct value, but the dynamic model noise covariance converges slowly. Furthermore, when continuing the algorithm for more iterations, we noticed that the KL divergences of the particle EM estimates of $\mathbf{Q}$ start to increase around 500 iterations. Investigating this phenomenon is beyond the scope of this paper, but this most likely stems from a positive feedback in EM combined with the stochasticity in the particle filtering scheme. Even though the sigma-point expectation–maximization is outperformed by the particle counterpart, the sigma-point scheme delivers similar results with small computational burden.

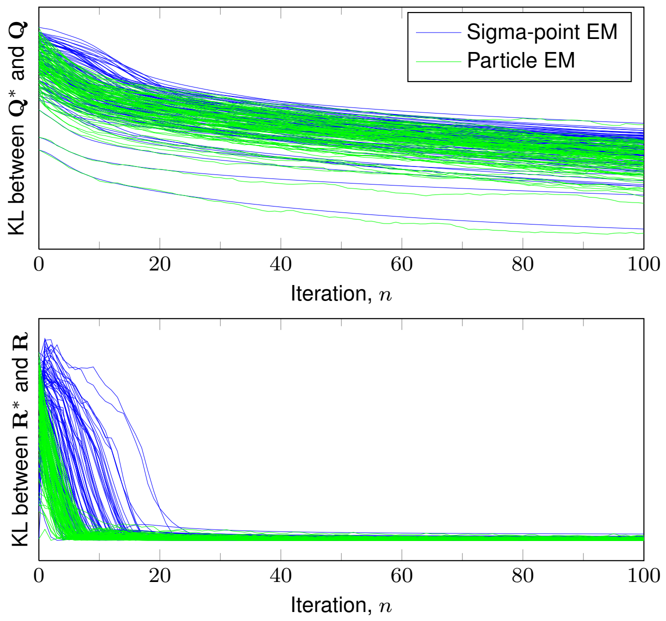

Fig. 4. Results for estimating the noise covariance matrices $\mathbf{Q}$ and $\mathbf{R}$ in the multi-dimensional coordinated turn example. The convergence is shown by comparing the estimates to the known Gaussian noise distributions in terms of the Kullback–Leibler divergence. The particle EM estimates appear to converge faster.

VIII. Conclusion and Discussion

In this paper, we considered the expectation–maximization algorithm for approximating maximum likelihood estimates of parameters in state space models. We focused on EM algorithms where the expectation step is approximated by sigma-point smoothers or particle smoothers, and compared them to direct likelihood maximization.

We presented a unifying view of the approximative EM methods employing either sigma-point or particle smoothers, as well as described the connection to direct ML estimation. We showed how the approximations to the pairwise smoothing distributions required in EM are obtained from the smoother algorithms during a single smoother run. Furthermore, for models that are linear-in-parameters, we have explicitly described how the EM bounds can be analytically maximized by generalizing the convenient matrix notations used for linear models in [11].

Using simulated data from a commonly encountered UNGM model, we compared the point estimates produced by the various maximum likelihood methods to the ground-truth values, as well as to posterior distributions sampled using particle MCMC. In this experiment, both sigma-point and particle EM methods produced good approximations to the direct ML estimates. However, in this one-dimensional case, direct likelihood maximization is feasible and thus the EM approximations are not required. On the other hand, in multiple dimensions, direct optimization of the likelihood may be difficult, especially when estimating matrix-valued parameters, such as noise covariance matrices. We implemented the EM algorithms for covariance matrix estimation in a five-dimensional coordinated turn model. In this example, the particle smoother EM converged faster in terms of iterations, while the sigma-point method produced similar results.

With the particle smoother used in this paper, the computational cost of the particle EM is quadratic in the number of particles. In addition, the number of particles required is much higher than the number of sigma-points, and therefore the particle method requires a considerably higher number of function evaluations. Furthermore, the particle filter scales poorly to higher-dimensional problems, whereas the computational cost of sigma-point filtering is linear in the state dimension.

A drawback of the sigma-point EM approach is that the sigma-point smoother does not perform well in highly non-linear problems, for example, the one-dimensional model used in this paper when the measurement model is quadratic. Augmenting the noise to the state has been shown to improve the performance of sigma-point smoothers [4]. However, our experiments with the augmented-noise version of the sigma-point smoother showed that the performance of the sigma-point EM in highly non-linear problems is still weak even if the smoother state estimates improve.

Our findings suggest that sigma-point EM is a computationally feasible alternative to particle smoothing based EM, as well as direct likelihood maximization in high-dimensional problems with moderate non-linearities.

Acknowledgment

This work was supported by grants from the Academy of Finland (266940, 273475) and the Finnish Funding Agency for Technology and Innovation (40304/09). We acknowledge the computational resources provided by the Aalto Science-IT project.

References

[1] S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte, "A new approach for filtering nonlinear systems," in *Proceedings of the 1995 American Control, Conference, Seattle, Washington*, 1995, pp. 1628–1632.

[2] ——, "A new method for the nonlinear transformation of means and covariances in filters and estimators," *IEEE Trans. Autom. Control*, vol. 45, no. 3, pp. 477–482, March 2000.

[3] K. Ito and K. Xiong, "Gaussian filters for nonlinear filtering problems," *IEEE Trans. Autom. Control*, vol. 45, no. 5, pp. 910–927, May 2000.

[4] Y. Wu, D. Hu, M. Wu, and X. Hu, "Unscented Kalman filtering for additive noise case: Augmented vs. non-augmented," in *Proceedings of the American Control Conference*, 2005, pp. 4051–4055.

[5] ——, "A numerical-integration perspective on Gaussian filters," *IEEE Transactions on Signal Processing*, vol. 54, no. 8, pp. 2910–2921, August 2006.

[6] M. Šimandl and J. Duník, "Derivative-free estimation methods: New results and performance analysis," *Automatica*, vol. 45, no. 7, pp. 1749–1757, 2009.

[7] S. Särkkä and J. Hartikainen, "On Gaussian optimal smoothing of non-linear state space models," *IEEE Trans. Autom. Control*, vol. 55, no. 8, pp. 1938–1941, 2010.

[8] M. Hürzeler and H. R. Künsch, "Monte Carlo approximations for general state-space models," *Journal of Computational and Graphical Statistics*, vol. 7, no. 2, pp. 175–193, 1998.

[9] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.

[10] C. Andrieu, A. Doucet, S. Singh, and V. Tadic, "Particle methods for change detection, system identification, and control," *Proc. IEEE*, vol. 92, no. 3, pp. 423–438, March 2004.

[11] S. Särkkä, *Bayesian Filtering and Smoothing*, ser. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2013, vol. 3.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 39, no. 1, pp. 1–38, 1977.

[13] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of Time Series Analysis*, vol. 3, no. 4, pp. 253–264, 1982.

[14] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*. Springer, 1998, pp. 355–368.

[15] S. Roweis and Z. Ghahramani, "Learning nonlinear dynamical systems using the expectation–maximization algorithm," in *Kalman Filtering and Neural Networks*, S. Haykin, Ed. Wiley-Interscience, 2001, ch. 6, pp. 175–220.

[16] T. B. Schön, A. Wills, and B. Ninness, "System identification of nonlinear state-space models," *Automatica*, vol. 47, no. 1, pp. 39–49, 2011.

[17] V. Väänänen, "Gaussian filtering and smoothing based parameter estimation in nonlinear models for sequential data," Master's thesis, School of Electrical Engineering, Aalto University, Finland, 2012.

[18] M. Gašperin and Đ. Juričić, "Application of unscented transformation in nonlinear system identification," in *Proceedings of the 18th IFAC World Congress*, vol. 18, no. 1, 2011, pp. 4428–4433.

[19] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. Chapman & Hall/CRC Press, 2013.

[20] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 3, pp. 269–342, 2010.

[21] R. Cools, "Constructing cubature formulae: The science behind the art," in *Acta Numerica*, 1997, vol. 6, pp. 1–54.

[22] J. S. Liu and R. Chen, "Blind deconvolution via sequential imputations," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 567–576, 1995.

[23] I. S. Mbalawata, S. Särkkä, and H. Haario, "Parameter estimation in stochastic differential equations with Markov chain Monte Carlo and non-linear Kalman filtering," *Computational Statistics*, vol. 28, no. 3, pp. 1195–1223, 2013.

[24] M. L. Andrade Netto, L. Gimeno, and M. J. Mendes, "A new spline algorithm for non-linear filtering of discrete time systems," in *A Link Between Science and Application of Automatic Control*, Helsinki, Finland, 1978, pp. 2123–2130.

[25] G. Kitagawa, "Non-Gaussian state-space modeling of nonstationary time series: Rejoinder," *Journal of the American Statistical Association*, vol. 82, no. 400, pp. 1060–1063, 1987.

[26] I. Arasaratnam and S. Haykin, "Cubature Kalman filters," *IEEE Trans. Autom. Control*, vol. 54, no. 6, pp. 1254–1269, June 2009.

[27] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*. Wiley Interscience, 2001.

[28] A. Solin, "Cubature integration methods in non-linear Kalman filtering and smoothing," Bachelor's thesis, Faculty of Information and Natural Sciences, Aalto University, Finland, 2010.