EXACT MOMENT MATCHING FOR EFFICIENT IMPORTANCE FUNCTIONS IN SMC METHODS

Saikat Saha, Pranab K. Mandal,
Department of Applied Mathematics
University of Twente
PO Box 217, 7500 NB Enschede, The Netherlands

Yvo Boers, Hans Driessen.
THALES Nederland BV
Haaksbergerstraat 49, 7554 PA
Hengelo, The Netherlands

ABSTRACT
In this article we introduce a new proposal distribution to be used in conjunction with the sequential Monte Carlo (SMC) method of solving non-linear filtering problem. The proposal distribution incorporates all the information about the to be estimated current state form both the available state and observation processes. This makes it more effective than the state transition density which is more commonly used but ignores the recent observation. Because of its Gaussian nature it is also very easy to implement. We show further that the introduced proposal performs better than other similar importance functions which also incorporate both state and observations.

1. INTRODUCTION
Consider a nonlinear dynamic system given by

\[ x_k = f(x_{k-1}) + w_k, \quad w_k \sim \mathcal{N}(0, Q) \]  

\[ y_k = h(x_k) + v_k, \quad v_k \sim \mathcal{N}(0, R), \quad k = 1, 2, \ldots \]  

where \((x_k)\) are the unobservable system values (the state) with initial (prior) \(p(x_0) \equiv p(x_0|x_{-1})\) and \((y_k)\) are the observed values (the measurements). Furthermore, the process noises \((w_k)\) are assumed to be independent of the measurement noises \((v_k)\). The main statistical problem related to this type of state-space model is to estimate the unobserved system value \(x_n\) from all the observations \(y_{1:n} \equiv \{y_1, y_2, \ldots, y_n\}\), up to time \(n\). This can be given by, for example, the conditional density or filtered density \(p(x_n|y_{1:n})\). For a point estimate one can consider the corresponding conditional mean. However, except in a few special cases such as when both the system and observation equation (1)–(2) are linear (Kalman filter), it is not possible to obtain an analytical solution. As a result, analytical approximations such as Extended Kalman filter and Gaussian sum filter are developed ([1, 2]). There are also other approximate methods available. For example, the use of numerical integration to arrive at solutions ([3]), the unscented Kalman filter ([4]) and the Gaussian quadrature Kalman filter ([5]).

The sequential Monte Carlo (SMC) methods, on the other hand, use simulation technique to reach a solution. With the advent of more and more powerful computers, the SMC methods have started receiving growing attention in recent times ([6, 7, 8, 9, 10, 11]). The biggest advantage of the SMC is that the method can easily adapt to the nonlinearity in the model and/or non-Gaussian noises. In the SMC methods, often referred to as Particle Filters (PF’s), probability distributions are represented by a cloud of particles. Particles are recursively generated via Monte Carlo simulation from a so called importance function, \(\pi(\cdot)\), also often referred to as proposal distribution. Furthermore, each particle receives an importance weight attached to it. Although the resulting distributions (represented by the particle clouds) do converge to the true filtered density as the Monte Carlo sample size tends to infinity, for finite sample size the efficiency of the SMC method depends heavily on the proposal density used.

Usually the ‘naïve’ proposal \(p(x_k|x_{k-1})\) is used as the importance function. The main reasons behind this choice are the ease of drawing samples from this Gaussian distribution and the simplicity of weight update equations. However, if the measurement is very informative, a lot of samples are wasted. To make the method more effective importance functions of the form \(\pi = p(x_k|x_{k-1}, y_k)\), i.e., the one which incorporates both the system and observation processes are suggested in [8, 12]. There are two major, in practice prohibitive, drawbacks for using this type of importance function. First, drawing samples according to \(p(x_k|x_{k-1}, y_k)\) is, in general, difficult. Secondly, it is also difficult to get an analytical expression for the proposal density needed to update the importance weights. To circumvent that, the authors in [12] approximate the observation model (2) with a linearised version so that \(p(x_k|x_{k-1}, y_k)\) becomes Gaussian and subsequently uses that as the importance function. In practice, however, there are many situations where linearizing the observation equation by way of differentiation is not possible, for reasons of nonsmoothness. For example, when observation equation involves indicator function or when inequality constraints are enforced by means of the likelihood.

In this article we propose another Gaussian importance function that is built by first approximating the conditional distribution of \((x_k, y_k)\) given \(x_{k-1}\), which, according to the
model, can be of any nature, by a Gaussian distribution whose moments are matched exactly to the theoretical moments obtained from the dynamical system equations (1)–(2). While both the importance functions in [12] and in this article are essentially deduced from Gaussian approximation of \( p(x_k, y_k|x_{k-1}) \), the difference lies in the way the moments of the distributions are calculated. In [12] they are based on linearisation of \( h(\cdot) \) in (2), whereas we use exact moments. Recently, other importance functions are proposed in [13] based on similar Gaussian approximations of \( p(x_k, y_k|x_{k-1}) \) where the authors further approximate the moments by different numerical methods. We show that although these methods also perform better than the linearisation method proposed in [12] the improvement is at comparable level with our proposed method of exact moment matching. However, our method is computationally less demanding than that in [13].

The rest of the article is organized as follows. In section 2 the general SMC method, and the construction of importance function proposed in [12] are briefly reviewed. We describe our proposed importance function in section 3 and compare it with the one in [13]. The numerical comparison results of these methods are presented in section 4. Finally section 5 concludes the article with a discussion.

2. PRELIMINARIES

2.1. General SMC method

Suppose the system dynamics are given by (1)–(2). The sequential Monte Carlo method which is based on importance sampling allows us to estimate recursively in time the distribution \( p(x_{0:n}|y_{1:n}) \) and expectations of the form

\[
I(g_n) = \int g_n(x_{0:n}) p(x_{0:n}|y_{1:n}) \, dx_{0:n}.
\]

The basic idea is as follows. Draw \( N \) independent samples \( \{x_{0:n}^{(i)}\}, i = 1, \ldots, N \) from a normalized importance function (proposal distribution) \( \pi(x_{0:n}|y_{1:n}) \), whose support include that of the true posterior. By associating (unnormalized) importance weight \( w_n^{(i)} = p(x_{0:n}^{(i)}|y_{1:n})/\pi(x_{0:n}^{(i)}|y_{1:n}) \) to the \( i \)-th sample \( x_{0:n}^{(i)} \), one can estimate \( I(g_n) \) by

\[
\hat{I}_N(g_n) = \sum_{i=1}^N f_n(x_{0:n}^{(i)}) \tilde{w}_n^{(i)}, \quad \tilde{w}_n^{(i)} = \frac{w_n^{(i)}}{\sum_{j=1}^N w_n^{(j)}}.
\]

Furthermore, the (weighted) particle cloud \( \{x_{0:n}^{(i)}, \tilde{w}_n^{(i)}\}, i = 1, \ldots, N \} \) can be considered to be representing the conditional distribution \( p(x_{0:n}|y_{1:n}) \). To make the method recursive one can choose the importance function of the form

\[
\pi(x_{0:n}|y_{1:n}) = \pi(x_0) \prod_{k=1}^n \pi(x_k|x_{0:k-1}, y_{1:k}).
\]

Then the recursive evaluation of the importance weights can be done as follows with successive observations \( y_k \) arrive.

\[
w_n^{(i)} = \frac{p(x_{0:n}^{(i)}|y_{1:n})}{\pi(x_{0:n}^{(i)}|y_{1:n})} \propto w_{n-1}^{(i)} p(y_n|x_n^{(i)}) p(x_n^{(i)}|x_{n-1}^{(i)}) p(x_{n-1}^{(i)}|x_{0:n-1}, y_{1:n}).
\]

Then the sequential importance sampling procedure can be summarized as follows. Recursively over time \( k = 0, 1, 2, \ldots \)

- sample \( x_k^{(i)} \sim \pi(x_k|x_{0:k-1}, y_{1:k}) \) and set \( x_{0:k}^{(i)} \triangleq (x_{0:k-1}^{(i)}, x_k^{(i)}) \)
- evaluate the importance weights (up to a normalizing constant) according to (3)

To avoid carrying the trajectories with small normalized importance weights and to concentrate upon the ones with large weights, the effective sample size \( N_{eff} = 1/\sum_{i=1}^N (\tilde{w}_n^{(i)})^2 \) (see [14]) can be used. The above algorithm is then further augmented by the following step

- If \( N_{eff} \) is below a specified threshold, resample \( \{x_{0:k}^{(i)}\}_{i=1}^N \) with probabilities \( \{\tilde{w}_n^{(i)}\}_{i=1}^N \) keeping the sample size still to be \( N \) and assign equal weights \( 1/N \).

Usually in practice, the importance function is taken to be the transition density, i.e., \( \pi(x_n|x_{0:n-1}, y_{1:n}) = p(x_n|x_{n-1}) \). Since the transition density is Gaussian, it is easy to draw sample from and to perform the weight update. It is known that this algorithm suffers from the degeneracy problem, that is to say, the variance of the importance weights can only increase over time. However, It has been shown in [12] that an importance function of the form

\[
\pi(x_n|x_{0:n-1}, y_{1:n}) = p(x_n|x_{n-1}, y_n)
\]

addresses this issue by minimizing the variance of the (unnormalized) importance weight \( w_k^{(i)} \) conditional upon \( x_{0:k-1} \) and \( y_{1:k} \). The advantage can be seen intuitively as well. When generating samples of \( x_k \) one should use not only the previous state (estimates) but also the available current observation \( y_k \), for it also contains information about \( x_k \). See [14, 8].

In general though this choice of importance function is difficult to implement because generating samples from it is not at all easy. Furthermore one needs an analytical expression of the importance function to be used in weight update equation, which is also in general difficult with this choice.

2.2. Importance Function by Doucet et al

Consider the system dynamics given by (1)–(2). If \( h(\cdot) \) in (2) is linear, it can be shown that \( p(x_k, y_k|x_{k-1}) \) is Gaussian and subsequently \( p(x_k|x_{k-1}, y_k) \) is also Gaussian. Using this fact, Doucet et. al. ([12]) linearise the observation equation to obtain

\[
y_k \approx h(f(x_{k-1})) + C_k(x_k - f(x_{k-1})) + w_k,
\]
where \( C_k = \frac{\partial h}{\partial x_k} (f(x_{k-1})) \) and subsequently use the corresponding Gaussian distribution as the importance function, i.e., \( \pi(x_k|x_{0:k-1}, y_{1:k}) \sim \mathcal{N}(m_k, \Sigma_k) \), where
\[
\Sigma_k^{-1} = Q^{-1} + C_k^T R^{-1} C_k
\]
\[
m_k = \Sigma_k Q^{-1} f(x_{k-1}) + \Sigma_k C_k^T R^{-1} (y_k - h(f(x_{k-1})) + C_k f(x_{k-1})).
\]

3. IMPORTANCE FUNCTION BASED ON MOMENT MATCHING

3.1. Exact Moment Matching (EMM)

A close look at the importance function used in [12] reveals that the authors have essentially approximated the conditional distribution of \((x_k, y_k)\) given \(x_{k-1}\) by the Gaussian distribution with mean \(\mu^*\) and covariance matrix \(\Sigma^*\) given by
\[
\mu^* = \left( \frac{f(x_{k-1})}{h(f(x_{k-1}))} \right) \quad \text{and} \quad \Sigma^* = \begin{pmatrix} Q & QC_k^T \\ C_k Q & C_k QC_k^T + R \end{pmatrix}.
\]

In this article we propose to approximate the conditional distribution of \((x_k, y_k)\) by the Gaussian distribution whose moments are exactly matched. In other words, with mean and covariance matrix given by
\[
\mu = \left( \frac{f(x_{k-1})}{\mu_h} \right) \quad \text{and} \quad \Sigma = \begin{pmatrix} Q & \Sigma_h \\ \Sigma_h^T & \Sigma_h + R \end{pmatrix},
\]
where \(\mu_h, \Sigma_h, \text{and } \Sigma_{hh}\) all depend on \(x_{k-1}\) and are given by
\[
\mu_h(x_{k-1}) = E(h(X_k)|x_{k-1})
\]
\[
\Sigma_h(x_{k-1}) = E(X_k [h(X_k)]^T |x_{k-1}) - f(x_{k-1})\mu_h^T
\]
\[
\Sigma_{hh}(x_{k-1}) = E(h(X_k)[h(X_k)]^T |x_{k-1}) - \mu_h\mu_h^T.
\]

Subsequently, we take the importance function to be given by \(\pi(x_k|x_{0:k-1}, y_{0:k}) \sim \mathcal{N}(m_k, \Sigma_k)\), where
\[
m_k = f(x_{k-1}) + \Sigma_h \Sigma_{hh} + R)^{-1} (y_k - \mu_h)
\]
\[
\Sigma_k = Q - \Sigma_h \Sigma_{hh} R)^{-1} \Sigma_h.
\]

Further, the weight update follows according to (3).

3.2. Comparison with the method by Guo et al

Recently, Guo et al. ([13]) also used similar techniques to determine the importance functions but in stead of the exact values they approximate the conditional moment in (6)-(8) by numerical techniques such as Gauss-Hermite quadrature (GHQ) rule and Julien-Uhlmann quadrature (JUQ) rule. For example, according to GHQ
\[
\int_{-\infty}^{\infty} g(x) \frac{1}{(2\pi)^{\frac{3}{2}}} \exp(-x^2) dx = \sum_{i=1}^{m} \omega_i g(x_i),
\]
where the number \(m\) and the location of the abscissas \(x_i\)’s and corresponding optimal weights \(\omega_i\)’s can be chosen beforehand ([15]). According to JUQ (see e.g., [16]) an \(n\)-dimensional standard Gaussian distribution is approximated by a discrete distribution taking values in \{\(z_1, \ldots, z_{2n+1}\}\) with corresponding probabilities \(P(z_k)\) given by
\[
z_k = \sqrt{n + \kappa \epsilon_k} \quad P(z_k) = \frac{1}{2(n + \kappa)} \quad \text{if } 1 \leq k \leq n,
\]
\[
z_k = - \sqrt{n - \kappa \epsilon_k} \quad P(z_k) = \frac{1}{2(n + \kappa)} \quad \text{if } n + 1 \leq k \leq 2n,
\]
\[
z_k = 0 \quad P(z_k) = \frac{2\kappa}{2(n + \kappa)} \quad \text{if } k = 2n + 1.
\]

Then
\[
\int \frac{g(x)}{(2\pi)^{\frac{3}{2}}} e^{-\frac{x^2}{2}} dx = \sum_{k=1}^{2n+1} g(z_k) P(z_k).
\]

Clearly, EMM can be implemented if the function \(h(\cdot)\) in the observation equation (2) is a polynomial function, because the moments of all order are known for a Gaussian random variable. In this case, as we shall see in the next section, EMM performs better than the method in [13].

4. NUMERICAL SIMULATION RESULTS

As in [12] we consider the system dynamics to be given by (1)-(2) with
\[
f(x_{k-1}) = \frac{x_{k-1}}{2} + \frac{25x_{k-1}}{1 + x_{k-1}^2} + 8 \cos(1.2k)
\]
\[
h(x_k) = \frac{x_k^2}{2\kappa}.
\]

We can then calculate the exact values of (6)-(8) as follows:
\[
\mu_h = \frac{f^2(x_{k-1}) + Q}{2\kappa}, \quad \Sigma_h = \frac{f(x_{k-1})Q}{10}, \quad \Sigma_{hh} = \frac{f^2(x_{k-1})Q}{100} + \frac{Q^2}{200}.
\]

The differences between (4) and (5) then appear in
\[
\mu_h = h(f(x_{k-1})) + \frac{Q}{2\kappa}, \quad \text{and } \Sigma_{hh} = C_k QC_k^T + \frac{Q^2}{200}.
\]

For GHQ, we use 5 point quadrature rule with the following abscissas and corresponding optimal weights ([15]).
\[
\begin{array}{cc}
x_i & 0 \\ \omega_i & 0.945309 ± 0.058572 ± 0.202018 \\
0 & 0.393619 ± 0.0199532
\end{array}
\]

For JUQ, setting \(n = 1\) and taking \(\kappa = 2\), we get \(z_1 = -z_2 = \sqrt{3}\), \(z_3 = 0\) with \(P(z_1) = P(z_2) = 1/6\), \(P(z_3) = 2/3\).

In our analysis \(Q = 10\) and \(R = 1\). The initial distribution \(p(x_0)\) is taken to be \(\mathcal{N}(0, 5)\). Resampling was done when the effective sample size became less than one-third of the original sample size \(N\).

We have considered a time series data of length 100, that is to say, \((x_1, y_1), \ldots, (x_{100}, y_{100})\). We have applied different particle filter techniques (importance functions) on these data – naive \(p(x_k|x_{k-1})\) (Naïve); proposal by Doucet et al. (Lin); proposal by Guo et al. with GHQ approximation (GHQ) and JUQ approximation (JUQ) and the one proposed in section 3 (EMM). For each method we have calculated the root mean
squared error (RMSE) given by \( \sum_{k=1}^{100} (\hat{x}_k - x_k)^2 / 100 \), where \( \hat{x}_k = E(x_k|y_{1:k}) \) is the estimated state. Each of these methods are implemented with different Monte Carlo sample sizes \( N = 100, 250, 500 \) and 1000. Further, for each method and \( N \), the process is repeated 100 times (runs). In the following table the average of these 100 RMSE’s are presented. Below the RMSE values and within brackets are shown the average (over the 100 runs) CPU time (in seconds) taken by that particular method.

<table>
<thead>
<tr>
<th>( N )</th>
<th>Naive</th>
<th>Lin</th>
<th>GHQ</th>
<th>JUQ</th>
<th>EMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>4.7201</td>
<td>4.8271</td>
<td>4.4011</td>
<td>4.4680</td>
<td>4.5208</td>
</tr>
<tr>
<td></td>
<td>(0.3422)</td>
<td>(0.2283)</td>
<td>(0.2908)</td>
<td>(0.2884)</td>
<td>(0.2456)</td>
</tr>
<tr>
<td>250</td>
<td>4.5619</td>
<td>4.4142</td>
<td>4.3290</td>
<td>4.3417</td>
<td>4.3239</td>
</tr>
<tr>
<td></td>
<td>(0.4417)</td>
<td>(0.2777)</td>
<td>(0.4452)</td>
<td>(0.4456)</td>
<td>(0.3206)</td>
</tr>
<tr>
<td>500</td>
<td>4.4767</td>
<td>4.4019</td>
<td>4.3420</td>
<td>4.3510</td>
<td>4.3596</td>
</tr>
<tr>
<td></td>
<td>(0.6634)</td>
<td>(0.4172)</td>
<td>(0.9300)</td>
<td>(0.8809)</td>
<td>(0.4797)</td>
</tr>
<tr>
<td>1000</td>
<td>4.4452</td>
<td>4.3548</td>
<td>4.3183</td>
<td>4.2996</td>
<td>4.3331</td>
</tr>
<tr>
<td></td>
<td>(1.4736)</td>
<td>(0.8134)</td>
<td>(2.4937)</td>
<td>(2.2792)</td>
<td>(0.9378)</td>
</tr>
</tbody>
</table>

First of all, we see from the table that as expected, the performance (as measured by RMSE) of all the methods become similar as sample size \( N \) increases. After all, as \( N \to \infty \) the particle filter converges to the true posterior for any proposal distribution. For moderate sample size the RMSE performance of Lin is better than Naive. Performances of GHQ, JUQ and EMM are more or less similar (which is better than Lin) but time taken to arrive at the estimate is smaller in EMM than that by GHQ and JUQ.

## 5. CONCLUSION

In this article a new importance density is proposed which is based on Gaussian approximation of the conditional distribution of \( (x_k, y_k) \) given \( x_{k-1} \) with matching first two moments of the true conditional distribution. It has been shown to perform better than the other similar proposals either in the sense of RMSE and/or in the sense of time taken to perform the estimation. To use the proposed method one needs to know the moments of the system dynamics up to the second order. When the exact moments are not known one can approximate the \( h(\cdot) \) in (2) by a polynomial function and subsequently, as noted in section 3.2, EMM can be applied. Results in this direction will appear elsewhere.

## 6. REFERENCES


