A MODE PRESERVING PARTICLE FILTER

Mark R. Morelande and Alan M. Zhang
Melbourne Systems Laboratory
Dept. of Electrical and Electronic Engineering
The University of Melbourne, Australia
email: mrmore@unimelb.edu.au,alan.alanzhang@gmail.com

ABSTRACT
A new particle filtering technique is proposed which addresses the problem of mode preservation in the approximation of multi-modal posteriors. The key is a resampling step which attempts to distribute samples evenly over the state space. This idea is shown to produce enormous benefits compared to the optimal importance density in a test problem requiring the retention of several modes for an extended period.

Index Terms—Particle filtering; Bayesian estimation

1. INTRODUCTION
Sequential Bayesian estimation, or filtering, problems arise in many applications, including econometrics, robotics and object tracking. Particle filters (PFs) are a Monte Carlo method for performing approximately optimal sequential Bayesian estimation. In the most common PF implementation samples are drawn from an importance density which factorises in such a way that trajectory samples and their corresponding weights can be computed recursively. This procedure, referred to as sequential importance sampling, provides an asymptotically exact representation of the posterior distribution as the sample size increases [1]. In recent times PFs have gained in popularity because of the ready availability of computational power and the relative ease with which they can be applied to a wide variety of problems. Useful reviews are given in [2] and, more recently [3].

A key problem in particle filtering is that of sample degeneracy. In a multi-modal situation sample degeneracy is manifested as an inability to reliably retain modes for extended periods. An example of this occurs in multi-object tracking if objects move in close proximity and then separate [4]. Immediately after the objects separate the posterior distribution will have almost evenly weighted modes for each permutation of the object indices. Thus, for r objects there will be r! modes. In a PF all modes except one will eventually be lost. This is not particularly deleterious if it is desired only to know where objects are at the current time. Usually, however, sample degeneracy is undesirable and produces disastrous results in multi-modal situations. For instance, in mobile robot localisation it may be necessary to retain a multi-modal posterior distribution for quite a long time until unambiguous localisation information becomes available [5].

The aim of this paper is to develop a PF which is capable of preserving a multi-modal posterior for long periods of time. The basis of the proposed approach is a resampling step which, for a given sample, considers not only the suitability of that sample, but also the number of neighbours it has and their weights. This is done in such a way that the resampled samples are approximately uniformly distributed over the part of the parameter space for which the posterior is significantly non-zero. Formally, this is achieved by using an auxiliary variable sampler [6]. A simulation scenario which models the case of a robot receiving ambiguous self-location information is used to show the potential of this idea.

The paper is organised as follows. Section 2 is a brief review of sequential Bayesian estimation using PFs. The mode-preserving PF, referred to as the uniform sampler, is developed in Section 3 and simulation results demonstrating its performance are given in Section 4.

2. PARTICLE FILTERING REVIEW
The process of interest is characterised by the state, denoted as \( \theta_t \in \mathbb{R}^d \) at time t. The state is a Markov process which evolves as

\[
\theta_t | \theta_{t-1} \sim f(\cdot|\theta_{t-1}), \quad t = 1, 2, \ldots
\]

with \( \theta_0 \sim \pi_0(\cdot) \). The notation \( \sim \) means "is distributed as". Inference about the state is made through the observations \( y_i \in \mathbb{R}^m \), \( t = 1, 2, \ldots \). The \( t \)th observation satisfies

\[
y_t | (\theta_{1:t}, y_{1:t-1}) \sim g(\cdot|\theta_t), \quad t = 1, 2, \ldots
\]

where \( \theta_{1:t} \) denotes the sequence \( \theta_1, \ldots, \theta_t \) and similarly for \( y_{1:t-1} \).

The filtering problem involves recursive estimation of \( \theta_t \) from the observations \( y_{1:t}, t = 1, 2, \ldots \). Bayesian estimation of \( \theta_t \) is based on the posterior distribution, i.e., the distribution of \( \theta_t \) conditional on the observations \( y_{1:t} \). Let \( \pi_t(\cdot) \) denote the posterior probability density function (PDF) of \( \theta_t \). Then, given \( \pi_{t-1}(\cdot) \) and an observation \( y_t \), Bayes’ rule can be used to find the posterior PDF as

\[
\pi_t(\theta_t) \propto g(y_t|\theta_t)\pi_{t-1}(\theta_t)
\]

where the prior \( \pi_{t-1}(\cdot) \) is found as

\[
\pi_{t-1}(\theta_t) = \int f(\theta_t|\theta)\pi_{t-1}(\theta) \, d\theta
\]

The posterior PDF is used to calculate various estimators of \( \theta_t \) such as the minimum mean square error estimator, the posterior mean.

Eq. (3) can rarely be evaluated exactly so numerical methods are usually required. PFs are a numerical method in which a discrete approximation to the posterior PDF is constructed using random sampling. The formulation of the PF given here closely follows that given in [6]. Assume the availability of samples \( \theta_{1:t-1}^1, \ldots, \theta_{1:t-1}^n \), where n is the sample size, with corresponding
weights $w_{i-1}, \ldots, w_{n-1}$ representing the posterior PDF $\pi_{t-1}(\cdot)$. The approximation formed by these samples and weights is

$$
\pi_{t-1}(\theta_{i-1}) \approx \sum_{i=1}^{n} w_{i-1} \delta(\theta_{i-1} - \theta_{i-1})
$$

(5)

where $\delta(\cdot)$ is Dirac’s delta. Substituting (5) into (4) and the result into (3) gives

$$
\pi_t(\theta_i) \propto g(y_i|\theta_i) \sum_{i=1}^{n} w_{i-1} f(\theta_i|\theta_{i-1})
$$

(6)

where the symbol $\propto$ means “is approximately proportional to”. The mixture (6) can be re-written as

$$
\pi_t(\theta_i, i) \propto w_{i-1} g(y_i|\theta_i) f(\theta_i|\theta_{i-1})
$$

(7)

where the mixture index $i$ is an index on the samples which form the approximation to $\pi_{t-1}(\cdot)$. This is called an auxiliary variable in [6] because it can be sampled and then discarded. Samples of the state and auxiliary variable are to be drawn from the importance density

$$
q_t(\theta_i, i) = \psi_{i} h_i(\theta_i|\theta_{i-1}, y_i)
$$

(8)

An importance sampling approximation to (7) can be obtained using $q_t(\cdot)$ in two steps. First, samples are drawn $(\theta_i, j') \sim q_t(\cdot)$ then the sample weights are calculated as, for $i = 1, \ldots, n$,

$$
w_{ij} \propto \pi_t(\theta_i, j') q_t(\theta_i, j') = \frac{w_{i-1} g(y_i|\theta_i) f(\theta_i|\theta_{i-1})}{\psi_{i} h_i(\theta_i|\theta_{i-1}, y_i)}
$$

(9)

The samples $\theta_{1}^i, \ldots, \theta_{n}^i$ and weights $w_{1}^i, \ldots, w_{n}^i$ form the desired approximation to $\pi_t(\cdot)$. For a large class of importance densities this sample-based approximation converges as the sample size $n \to \infty$. In practice, for a finite sample size, care must be taken when selecting $q_t(\cdot)$ if good performance is to be achieved.

3. MODE PRESERVING SAMPLING

The proposed procedure for mode preserving sampling is based on the idea that it is desirable to spread the samples evenly over the area of the state space for which the posterior is non-negligible. It is argued that this will reduce the speed with which sample degeneracy occurs because it will prevent samples concentrating around modes which are fortuitously favoured by the observations. Although this idea is not without drawbacks, a discussion of these is postponed until after the proposed technique is presented.

Let $A \subset \mathbb{R}^d$ be such that $\pi_t(\theta_i) \leq 1$. A uniform sample over $A$ can be obtained by drawing $\theta_i \sim U_A(\cdot)$, where $U_A(\cdot)$ is the uniform distribution over the set $A$, and computing the weights $w_{i}^t \propto \pi_t(\theta_i)$. This procedure is not implementable for several reasons, most obviously because the posterior PDF $\pi_t(\cdot)$ is unavailable. In Section 3.1 it is shown how uniform sampling can be approximately implemented in a PF framework. A discussion of several issues related to the proposed sampling procedure is given in Section 3.2.

3.1. Development

To simplify the exposition the following decomposition is used:

$$
g(y_i|\theta_i) f(\theta_i|\theta_{i-1}) = \psi_i(y_i) \psi_i(\theta_i|\theta_{i-1}, y_i)
$$

(10)

This is referred to as full adaption in [6]. The more general case where full adaption is not practicable is discussed afterwards. If (10) holds then (7) can then be written as

$$
\pi_t(\theta_i, i) \propto w_{i-1}^t \psi_i(y_i) \psi_i(\theta_i|\theta_{i-1}, y_i)
$$

(11)

If the sampling density $h_i(\cdot)$ for the state in (8) is set to $\psi_i(\cdot)$ then the weight calculation (9) becomes

$$
w_{ij} \propto w_{i-1}^t \psi_i(y_i)/\psi_i^t
$$

(12)

A uniform sampling can be accomplished if $\psi_i^t$ is such that $w_{ij} \propto \pi_t(\theta_i)$. This cannot be achieved exactly because $\pi_t(\cdot)$ is not known and because $\psi_i^t$ must be calculated without knowledge of $\theta_i$. However, $\pi_t(\theta_i)$ can be replaced with an approximation to obtain

$$
\psi_i^t \propto w_{i-1}^t \psi_i(y_i)/\psi_i^t
$$

(13)

with $\tilde{\psi}_i^t$ an approximation of the posterior PDF at the yet-to-be-sampled state $\theta_i^t$. The approximation used here is

$$
\tilde{\psi}_i^t = \sum_{j=1}^{n} w_{ij} k(\hat{\theta}_j^t - \hat{\theta}_i^t)
$$

(14)

where $w_{ij} \propto w_{i-1}^t \psi_i(y_i), \hat{\theta}_j^t \sim v(\cdot|\hat{\theta}_j^t, y_i)$ and $k(\cdot)$ is a suitable kernel density. With the index sampling probabilities of (13) the sample weights become

$$
w_{ij} \propto \tilde{\psi}_i^t
$$

(15)

If (10) can’t be used it is usually possible, for example, by linearisation, to define an approximate observation PDF $\tilde{g}(\cdot)$ such that

$$
\hat{g}(y_i|\theta_i) f(\theta_i|\theta_{i-1}) = \tilde{g}(y_i|\theta_i) \psi_i(\theta_i|\theta_{i-1}, y_i)
$$

(16)

Then, state samples can be drawn from $\tilde{\psi}(\cdot|\hat{\theta}_j^t, y_i)$ and the index sample probabilities set to $w_{ij} \propto w_{i-1}^t \tilde{\psi}_i^t(y_i)/\tilde{\psi}_i^t$. The resulting sample weights are $w_{ij} \propto \tilde{\psi}_i^t g(y_i|\theta_i)/\tilde{g}(y_i|\theta_i)$.

3.2. Discussion

Conventional PFs consider only the suitability of the individual samples when indices are sampled. Thus, the bootstrap filter [7] samples indices according to $w_{i-1}$ while the optimal importance density (OID) [8] samples according to the product $w_{i-1} \psi_i(y_i)$. In contrast, the uniform sampler takes into account not only the suitability of the individual sample, but also the nature of the estimated posterior PDF in its neighbourhood. Thus, if a certain region contains a large number of non-negligibly weighted samples, so that the estimated posterior PDF in this region is high, the uniform sampler will tend to select fewer samples from this region than the OID would. Likewise, samples in areas of low estimated posterior PDF are more likely to be selected than in the OID. It is expected that this biasing of the sampling to areas of low estimated posterior PDF, whether they are the tails of a distribution or an apparently improbable mode, will produce increased robustness.

An important property of conventional PFs is that they have $O(n)$ computational expense for a sample size $n$. The need to compute the estimated posterior PDF (14) for each sample means that the uniform sampler does not share this property. A straightforward implementation requires $O(n^2)$ computations. Several steps can be taken to reduce this expense. First we note that the summation need not extend over all samples. For the $r$th sample only those samples which lie in a neighbourhood of $\hat{\theta}_r$ need to be considered.
This is a range searching problem for which a large number of computationally efficient algorithms have been proposed. For instance, range searching using a k-D tree [9] requires $O(n \log n)$ operations to build the tree and $O(n^{1-1/d} + b)$ operations to find the $b$ neighbours of a sample. Significant reductions can be obtained using approximate range searching, as suggested in [10] where, for convex ranges, an algorithm with expense $O(n \log n + 1/\epsilon^{d-1})$, where $\epsilon$ is the proportional error permitted in the range query, is described. Note that approximate range searching may affect performance for a finite sample size but does not affect asymptotic convergence as $n \to \infty$. It is also important to consider that computation of the transition prior $f(\cdot | \cdot)$ is often considerably simpler than computation of the observation PDF $g(\cdot | \cdot)$ and the drawing of samples. In such cases, for moderate sample sizes the additional expense required for the uniform sampler will be small relative to the expense of the operations shared by the uniform sampler and conventional PFs. Our experience suggests that this additional computational expense is well worth the improved performance offered by the uniform sampler.

It is not clear that uniform sampling is desirable under all conditions. In a multi-modal posterior with well-separated modes the uniform sampler tends to place the same number of samples in each mode with the weights of the samples indicating the relevance of the mode. This may be inefficient if the observations favour certain modes not by chance, but rather because these modes are in fact more probable. Our argument is that it is more important to ensure that all modes are represented than that the number of samples assigned to a mode reflects its dominance. After all, the error involved in losing a mode is potentially much greater than the error which results from representing a probable mode with fewer samples.

4. A Numerical Example

The example used to demonstrate the performance of the proposed uniform sampler involves a robot moving around the edge of the region $[0,1]^2$. The aim is to localise the robot using observations from four sensors which provide proximity measurements in the directions $k\pi/2$, $k = 1, \ldots, 4$, relative to the robot orientation. The initial information is that the robot is located in one of the corners. The orientations for each corner are chosen so that the sensors provide equivalent observations in each corner. These orientations change in correspondence with the changes of the robot orientation. The situation is depicted in Fig. 1.

![Fig. 1. Scenario for numerical example. The solid circle is initial robot position with the arrow indicating the orientation. The dotted line is the robot trajectory. The dashed circles and arrows are additional possible locations and orientations.](image-url)

The example considered here is not intended to accurately model any particular real-world scenario. Rather, the intention is to create a scenario in which the mode preserving properties of the proposed uniform sample could be examined. In the scenario of Fig. 1 there are four modes which must be retained indefinitely since no information is provided to select between the modes.

The state $\theta_t$ is the robot position in Cartesian coordinates. The assumed transition prior is

$$f(\theta_t | \theta_{t-1}) = N(\theta_t; \theta_{t-1}, \kappa^2 I_2)$$

where $I_2$ is the $2 \times 2$ identity matrix. Let $\rho(\theta, \phi)$ denote the distance between the boundary and the position $\theta$ in a direction $\phi$. The observation vector $y_t = [y_{t,1}, \ldots, y_{t,4}]$ at time $t$ is generated as

$$g(y_t; \theta_t, \phi_t) = \prod_{j=1}^{4} N(y_{t,j}; \min\{0.3, \rho(\theta_t, \phi_t + j\pi/2)\}, \sigma^2)$$

Note that the orientation $\phi_t$ is assumed known and that meaningful observations are only produced in directions in which the boundary is closer than 0.3. The OID for the given transition prior and observation PDF is a mixture of truncated Gaussian distributions. Thus full adaption is possible and the uniform sampler can be implemented using (13) for the index sampling probabilities. The uniform sampler is compared below with the OID.

We first show an exemplary run for a case in which the robot moves at a rate of 0.02 per time step. The process noise standard deviation is $\kappa = 0.03$ and the observation noise standard deviation is $\tau = 0.02$. Both algorithms are applied with a sample size of $n = 500$. Figs. 2 and 3 show the sample constellations of the OID and uniform sampler, respectively, at times $t = 20, 40, 60$ and 80. At times $t = 20$ and 60 the samples should be spread along each side of the region between 0.3 and 0.7. At times $t = 40$ and 80 the samples should be concentrated in the four corners of the region. This desired behaviour is displayed by the uniform sampler but not the OID. Even at $t = 20$ it is clear that the OID cannot adequately represent the four elongated modes. As a result by the time the robot reaches the first corner, at $t = 40$, two modes have been lost. One of the remaining two modes is lost by the second corner.

Monte Carlo simulations are now used to compare the performances of the OID and the uniform sampler over a number of observation realisations. Two performance metrics are used. The first
is the mean number of modes retained by the filter. The second is the mode imbalance as measured by the mean distance between the centre of the square $[0, 1]^2$ and the filter’s posterior mean approximation. Note that for perfectly balanced modes this distance is zero. Both of these statistics are computed at the end of a surveillance period which runs for $t = 400$ time steps. The robot and noise parameters are as described above. The robot passes through 10 corners during the surveillance period. Fig. 4 shows the mean number of modes and balance error at the end of the surveillance period plotted against sample size. The results were obtained for the OID and the uniform sampler by averaging over 100 realisations. The uniform sampler is markedly superior to the OID. A sample size of 500 is sufficient to retain appreciable probability mass at all four modes. In contrast, the mean number of modes retained by the OID is below two even with a sample size of 2000. Although the uniform sampler successfully retains particles in the correct locations, the significant balance errors indicate that their weights do not reflect the distribution of mass between the modes. This is an unavoidable consequence of the approximations accumulated at each observation update. All sequential Monte Carlo methods suffer from this problem. Even so, compared to the OID the uniform sampler provides a much more balanced representation of the posterior which improves quite rapidly with sample size. It is notable that, even for a moderate sample size, the imbalance is sufficiently small that, should observations capable of distinguishing between the modes become available, the uniform sampler will produce correct estimates of the robot position.

5. CONCLUSIONS

The principal idea of this paper has been that a particle filter which seeks to sample uniformly over the part of the state space in which the posterior is non-negligible will be robust in multi-modal situations. The proposed uniform sampler achieves an approximately uniform distribution of samples within the usual sequential importance sampling framework through a resampling step which selects samples based not only on their individual weighting, but also considers the accumulated weights of samples in their neighbourhood. This is conceptually quite different to the conventional approach to particle filtering.

The performance of the proposed uniform sampler was demonstrated on an example which required the retention of several equally probable modes for an extended period of time. In this example it outperformed the optimal importance density by a large margin. It would be of interest to see if similar performance improvements can be achieved more generally. For instance, the uniform sampler may be particularly useful for approximation of heavy-tailed posteriors.

6. REFERENCES