THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS INC.

SIGNAL PROCESSING
ALGORITHMS, ARCHITECTURES, ARRANGEMENTS, AND APPLICATIONS
SPA 2024
September 25th - 27th, 2024, Poznań, POLAND

Multiple sampling with reduced resampling for particle filtering

Deepthi Kattula[*], P. Rajesh Kumar, Praveen B. Choppala[*]
Dept. of Electronics & Communication Engineering
Andhra University, India
{kdeepthi.rs, prof.prkumar, praveenchoppala}@andhrauniversity.edu.in

***Abstract*—The interest of this paper is the Bayesian particle filter for tracking applications. The filter is known to suffer from the degeneracy problem in which very few particles accumulate a large weight. This problem is overcome in the resampling step which randomly replaces particles with small weights by those which large weights without altering the particle representation of the posterior density. However, the resampling step is a sequential process and is thus computationally cumbersome when a large number of particles are used. Attempts to bypass resampling or make it computation-friendly have not yet yielded satisfactory outcomes in terms of the tracking accuracy or the alteration of the representation of the posterior. With this as the problem state, this paper aims to develop a particle filter scheme that avoids the resampling step but solves the degeneracy problem. This can be achieved by sampling multiple child particles at the current time step from each parent particle corresponding to the previous time step and then determining those that contribute to the posterior via the rule of unbiasedness. This idea is highly parallel-friendly because all the particles can be processed in batches, and sequentiality, if any, remains only within the replacement step. The proposed filter is, hence, suitable for hardware architectures. The merits of the proposed method are shown using a simulated study.**

***Index Terms*—particle filter, importance sampling, resampling, multiple sampling, root mean square error, parallel-friendly.**

I. INTRODUCTION

Bayesian state estimation entails inferring the state of a dynamic target from noisy data streams, often referred to as observations. [1]. The particle filter (PF) stands as the Bayesian solution for estimating nonlinear non-Gaussian systems. It employs a set of samples (or delta functions) with corresponding weights to model the posterior Probability Density Function (PDF). [2], [3] The PF operates through two key stages: sequential importance sampling (SIS) and sequential importance resampling (SIR). In SIS, particles are propagated from one-time step to the next and weighted to accurately depict the posterior PDF. However, the SIS method alone often leads to degeneracy, where one particle accumulates most of the weight while others become negligible contributors to the posterior distribution. To address this issue, the SIR stage comes into play. Here, particles with low weights are randomly replaced by those with higher weights, ensuring that the representation of the posterior PDF by the PF remains unchanged; this process is termed resampling [4], [5]. It's widely acknowledged that the accuracy of the PF's representation of the posterior improves asymptotically with an increase in the number of particles. This paper aims to devise a PF strategy capable of accurately representing the PDF with a reduced number of particles while also being highly compatible with parallel computing architectures.

Stochastic resampling methods are widely favoured in PFs. One of the initial techniques, known as multinomial selection [2], involves redistributing particles based on the cumulative sum of their normalized weights. This approach has seen improvements with the introduction of stratified resampling [6] and systematic resampling [3], both aimed at reducing the Monte Carlo variance in estimation. Utilizing random samples for resampling entails sequential search and extensive communication overhead among all particles [11]. This results in significant computational complexity, limiting the feasibility of employing more particles to represent the PDF accurately. In contrast, the residual resampler [7]addressed these challenges by stochastically duplicating particles using the principle of proportional allocation, which ensures unbiasedness. This criterion has been adopted in soft resampling techniques [8], [9]. Recently, a novel class of resamplers tailored for hardware architectures has emerged [15], [16]

One prominent approach to achieve precise tracking with fewer particles is the auxiliary particle filter (APF) [17]. The filter creates a batch of lookahead particles and evaluates their weights. Subsequently, it resamples these particles, employing the resampling indices to progress the existing particles to the subsequent time step. This methodology empowers particles to adeptly navigate the state space, diminishing the demand for a large number of particles. Recent advancements, such as the enhanced auxiliary particle filter (IAPF) [18], along with other lookahead techniques [19], [20], contribute to streamlining particle usage.

Our contribution: In this paper, we propose to sample multiple children particles for each parent particle corresponding to the previous time step. This can be accomplished by generating

[*] Corresponding Author.

multiple target heading disturbances from the Markov state transition density. We then determine those particles that become important and span regions of the high probability density of the posterior using the rule of proportional allocation. The key merit of this proposal is that the resultant particle set represents the posterior accurately, and the variance of the weights is maintained high without having to actually perform random resampling. Moreover, the method is fully parallel friendly.

The rest of the paper is organised as follows. The Bayesian estimation methodology is first presented in section I followed by the particle filter in section II. Then we present the proposed multiple sampling approach in section III, followed by simulation study in section IV and concluding remarks in section V.

II. Particle filtering

In this section, we fix the notation and describe the PF operation. The latent target state $\mathbf{x}_t \in \mathbb{R}^{d_\mathbf{x}}$ at time instant $t \in \mathbb{N}$ is a hidden Markov process with initial distribution $p(\mathbf{x}_{t=0}, \Theta)$ and governed by the Markov state transition density

$$\mathbf{x}_t \sim p(\mathbf{x}_t|\mathbf{x}_{t-1}, \Theta), t = 1, ..., T \tag{1}$$

The sensor observations $\mathbf{y}_t \in \mathbb{R}^{d_\mathbf{y}}$ are conditionally independent given the state variable $\mathbf{x}_t$ and are governed by the observation density as

$$\mathbf{y}_t \sim p(\mathbf{y}_t|\mathbf{x}_t, \Theta), t = 1, ..., T \tag{2}$$

Here, Θ denotes the set of model parameters which are assumed to be known. The set of states and the observations are denoted as $\mathbf{x_{1:t}} = \{\mathbf{x}_1, \cdots, \mathbf{x}_t\}$ and $\mathbf{y_{1:t}} = \{\mathbf{y}_1, \cdots, \mathbf{y}_t\}$.

Bayesian filtering aims to estimate the target distribution of the target state $p(\mathbf{x}_t|\mathbf{y}_{1:t}), t = 1, ..., T$ according to the Bayesian recursion

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto \int p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})d(\mathbf{x}_{t-1}) \tag{3}$$

The PF aims to represent this posterior PDF using a set of particles and their associated weights $\{\mathbf{x}_t^i, w_t^i\}_{i=1}^N$ as

$$p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) \approx \sum_{i=1}^{N} w_{t-1}^i \delta(\mathbf{x}_{t-1} - \mathbf{x}_{t-1}^i) \tag{4}$$

If the weighted particles $\{\mathbf{x}_{t-1}^i, w_{t-1}^i\}_{i=1}^N$ corresponding to time $t-1$ are available, then to move to time t, the PF generates a new set of particles from the old ones using a importance sampling distribution as

$$\bar{\mathbf{x}}_t^i \sim q(\mathbf{x}_t|\mathbf{x}_{t-1}^i, \mathbf{y}_t), \ i = 1, \cdots, N \tag{5}$$

The new particles are then weighted as

$$\begin{aligned} \bar{w}_t^i &= w_{t-1}^i \frac{p(\mathbf{y}_t|\mathbf{x}_t^i)p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)}{q(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i, \mathbf{y}_t)} & (6) \\ &\propto w_{t-1}^i p(\mathbf{y}_t|\bar{\mathbf{x}}_t^i), \ i = 1, \cdots, N & (7) \end{aligned}$$

where (7) follows by taking a convenient assumption that the particles are drawn from the Markov state transition density as $q(\mathbf{x}_t|\mathbf{x}_{t-1}^i, \mathbf{y}_t) = p(\mathbf{x}_t|\mathbf{x}_{t-1}^i)$. Once normalised as $\bar{w}_t^i = \bar{w}_t^i / \sum_{i=1}^N \bar{w}_t^j$, these weighted set of particles are representative of the posterior PDF at time t as in (4). As time progresses, this operation causes the discrepancy between the weights to increase, leading to degeneracy. The solution to this is the resampling step wherein those particles that have negligible weights are replaced by exact copies of other particles that have larger weights, i.e., for $i = 1, \cdots, N$, we sample an index $j(i)$ distributed according to the probability $\mathrm{P}(j(i) = m) = \bar{w}_t^m, m = 1, \cdots, N$ and set $\mathbf{x}_t^i = \bar{\mathbf{x}}_t^{j(i)}$ and set $w_t^i = 1/N$.

III. Proposed multiple sampling particle filter

In this section, we present the multiple sampling PF method that completely bypasses the need to resample within a sequential framework. Assume that the posterior PDF of the target state at time $t-1$ is represented by a set of weighted particles $\{\mathbf{x}_{t-1}^i, w_{t-1}^i\}_{i=1}^N$. Following (5) and (7), we predict a new set of particles and their updated normalised weights $\{\bar{\mathbf{x}}_t^i, \bar{w}_t^i\}_{i=1}^N$. As a general principle of unbiassedness, the resampling of this set should preserve the original particle distribution if there is no more information being considered in the process. That is, the expected number of times that each particle is resampled, n_t^i, is proportional to its weight as

$$n_t^i = \lfloor N\bar{w}_t^i \rceil, i = 1, \cdots, N \tag{8}$$

Since the weights lie in the interval $\bar{w}_t^i \in (0, 1)$, the number of replications for each particle will be in the interval $n_t^i \in (0, N)$. Also, if a weight is $\bar{w}_t^i = 1/N$, then $n_t^i = 1$. This implies that in a set of particles with equal weights, each particle is equally important in representing the posterior PDF. The process of eliminating small weight particles and replacing them with those having large weights, traditionally termed resampling, is governed by this principle, which is also called the principle of proportional allocation, that ensures the particle approximation of the posterior PDF is not altered (or unbiased). Using this principle, we propose the following PF approach.

The state space model described by the state transition density $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \Theta)$ defines the process of generating new particles (read, children) at time t from the previous particles (read, parents) at time $t-1$. In our proposed method, instead of generating one particle, i.e., child, from each previous particle, i.e., parent, as is the case in the traditional PF, we propose to generate multiple children at time t for each parent particle, say $M > 1$ at time step. This can be achieved in the following steps.

1) Generate multiple disturbances $\mathbf{a}_t^{i,j}, i = 1, \cdots, N, j = 1, \cdots, M$ as

$$\mathbf{a}_t^{i,j} \sim p(\mathbf{a}_t|\Theta) \tag{9}$$

i.e., M disturbances for each ith particle.

2) Generate one child particle at t according to importance sampling distribution (5) for each parent particle $\mathbf{x}_{t-1}^i$

3) Append the generated disturbances to each child particle as follows

$$\bar{\mathbf{x}}_t^{i,j} = f(\mathbf{x}_{t-1}^i, \mathbf{a}_t^{i,j},),\ i = 1, \cdots, N, j = 1, \cdots, M \tag{10}$$

$$\tag{11}$$

This will implicitly form a set of MN particles $\bar{\mathbf{x}}_t^{i,j}, i = 1, \cdots, N, j = 1, \cdots, M$ completely in a parallelised fashion. i.e., generate M disturbances for each of the N particles, generate N child particles and combine the two using (11).

4) Rearrange the particles. This can be written as $\{\bar{\mathbf{x}}_t^i\}_{i=1}^{MN}$.

5) Compute the weights of the generated particles according to (7) for $i = 1, \cdots, MN$ particles.

6) Do Normalisation process so that they sum to one.

7) The weighted particle set $\{\bar{\mathbf{x}}_t^i, \bar{w}_t\}_{i=1}^{MN}$ will now be representative of the posterior PDF at time t with MN particles. However, since we need only N particles to represent the posterior PDF, rather unbiassedly, we employ the subsequent approach used in [8].

8) Sort the particles and the weights in accordance to the weights and obtain a new set $\{\bar{\mathbf{x}}_t^i, \bar{w}_t\}_{i=1}^{MN}$ such that $\bar{w}_t^1 \geq \bar{w}_t^2 \geq \cdots \geq \bar{w}_t^N$.

9) Compute the number of replications for each particle according to (8).

This implies that $1/N$ is the minimum weight for a particle to have at least one replication. Each $\bar{\mathbf{x}}_t^i$ is replicated n_t^i times to obtain a new particle set $\{\mathbf{x}_t^i\}_{i=1}^N$ and each of its replication is reweighted as

$$w_t^i = \bar{w}_t^i / n_t^i \tag{12}$$

Since only N weighted particles are sufficient, we terminate the algorithm once N particles are created in the set $\{\mathbf{x}_t^i\}_{i=1}^N$. This is the proportional allocation scheme which can be performed completely deterministically while keeping the particle representation of the posterior PDF unaltered. The weight of the discarded particles is reallocated to the retained weight as

$$w_t^i = w_t^i + a_t^{\text{spare}} \tag{13}$$

where the discarded weight at each time step is given by $a_t^{\text{spare}} = 1 - w_t^i$.

The advantage of this proposal is that it bypasses the need for the computationally expensive resampling step and works solely on sampling a fixed number of multiple children particle for each parent particle. This approach is highly suitable for parallel architectures as sampling can be performed batch-wise on the entire particle set. Moreover, the proposed method is expected to yield fairly reliable results with very few particles as each particle is drawn with conformity to unbiassedness property. This substantially accelerates the PF process for real-time tracking applications. An inherent challenge, however, is to vet the theoretical properties of the proposed method in converging to the posterior, as the method relies on disturbances rather than the samples themselves to align the samples within the regions of the posterior.

IV. Simulation Study

In this section we present the simulation results for the proposed method. All the results of this paper are averaged over 1000 Monte Carlo iterations.

A. Linear state space model

We first test the proposed method using a linear Gaussian auto regressive AR(1) model to verify its theoretical conformity with the optimal Kalman filter. The model is given by

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\phi \mathbf{x}_{t-1}, \tau^2) \tag{14}$$

$$p(\mathbf{y}_t|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t, \sigma^2) \tag{15}$$

for $t = 1, \cdots, T$, where the model parameters $\Theta = (\phi, \tau^2, \sigma^2) = (0.99, 1, 0.1)$, the initial target state is $\mathbf{x}_{t=0} = 0$ and the total number of time steps is $T = 100$. We compare the proposed method using $M = 3$ and $M = 5$ against the systematic resampler [3], the residual resampler [7], the APF [17], the soft resampler [8] and the recently random network resampler [16]. Firstly, we test the faithfulness of our proposed method in representing the posterior PDF in accordance to the Kolomogorov Smirnov (KS) statistic disagreement with the optimal Kalman filter. KS testing has been used previously for other PF problems and provides a reliable measure of the accuracy of the estimate of the posterior. Here, after we obtain the final set of weighted particles, we de-mean and de-correlate them according to the theoretically optimal Kalman filter distribution and then take the maximum KS deviation of the resultant with the Gaussian error function. This approach was first proposed in [9] and is a well-accepted measure to test the representation accuracy of the PF. A small value of the KS statistic indicates a faithful representation of the posterior PDF of the Kalman filter. As the number of particles increases, any PF method converges to the posterior. How it behaves with fewer particles is of key interest. Figure 1 shows the KS statistic versus the number of particles for the given model, and it can be observed that our proposed method achieves the lowest KS statistic value with fewer particles and just $M = 3$ child particles per parent particle. The main reason for this faithfulness is that we sample multiple particles in a way that obeys the unbiassed criterion of resampling. That being said, it has also been found that generating more children particles per parent particle (see, $M = 5$ case) can lead to higher weight particles being replicated more times, causing loss of information contained at the tails of the posterior.

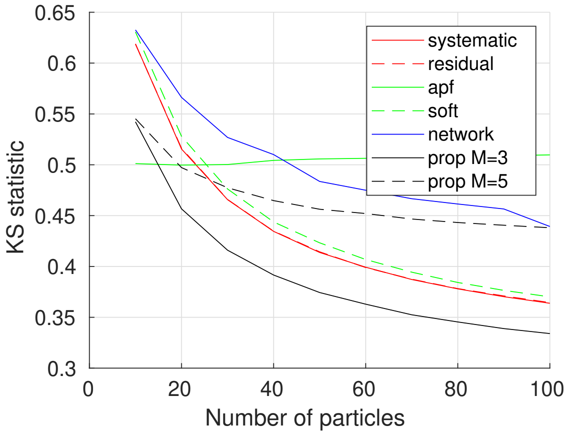

Fig. 1. The KS statistic versus the number of particles.

B. Nonlinear state space model

The nonlinear state space model is given by

$$\mathbf{t}_t = \frac{\mathbf{x}_{t-1}}{2} + \frac{25\mathbf{x}_{t-1}}{1+\mathbf{x}_{t-1}^2} + 8\cos(1.2t) + a_t \tag{16}$$

$$\mathbf{y}_t = \frac{\mathbf{x}_t^2}{20} + e_t \tag{17}$$

for $t = 1, \cdots, T$, where the heading disturbance is $a_t \sim \mathcal{N}(0, \tau^2)$ and the observation noise is $e_t \sim \mathcal{N}(0, \sigma^2)$ and the total number of time steps is $T = 100$. The model parameters are $\Theta = (\tau^2 = 10, \sigma^2 = 1)$. Figure 2 shows the root mean square error (RMSE) versus the number of particles at observation noise variance $\sigma^2 = 1$. It can be observed that the proposed method exhibits high tracking accuracy with fewer particles in both noise scenarios. The reason for this is that the sampled particles are made to explore regions of high probability density within the posterior PDF by virtue of sampling multiple copies to conform to the unbiassed criterion in (8).

Finally, we evaluate the performance of the proposed method is reducing the variance of the weights, as the key motivation for the method is to bypass the computationally cumbersome resampling step while being able to avoid degeneracy. For this, we measure the estimated effective sample size given by

$$\hat{N}_t^{\text{eff}} = \frac{1}{\sum_{i=1}^{N}(w_t^i)^2} \tag{18}$$

We compare the well-established systematic resampler with the proposed method. The distinctiveness between the two is that the former generates one child particle at time t for each parent particle corresponding to time $t-1$ and then randomly replaces the low-weight particles with those having large weights. This process is sequential in nature for all the N particles and involves particle replacement. In the latter, we propose to sample multiple child particles for each parent, precisely MN child particles, and then discard those over and beyond the Nth sample index once the unbiassedness criterion in (8) is reached. Figure 3 shows the effective sample size versus the time index for all $T = 100$ time steps for 20 and 100 particles. It can be observed that, in both cases, by virtue of sampling from regions that contribute to the posterior PDF, the proposed method is able to totally overcome degeneracy without having to resample particles. Sampling from regions of high importance of the posterior is possible because of sampling multiple child particles for each parent particle and selecting those that would remain important when propagated.

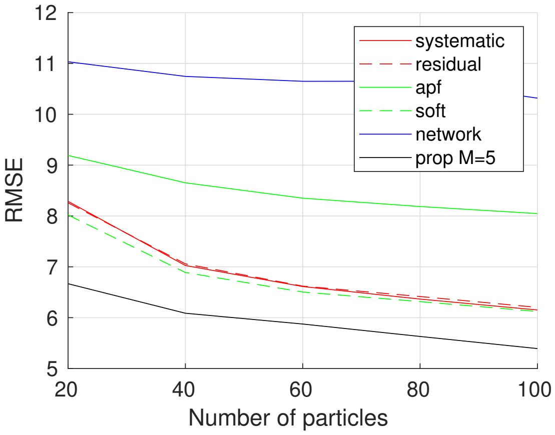

Fig. 2. The RMSE versus the number of particles at $\sigma^2 = 1$.

Now that we have shown that the proposed method faithfully represents the posterior PDF and also tracks accurately with fewer particles and also does not suffer from degeneracy, it is critical to understand its computational gain when implemented in parallel architectures. Note that the proposed method only employs sampling multiple child particles per parent particle, which can be performed completely in parallel. Moreover, the replication process does not involve any random search along the cumulative sum of the weights as do the conventional resamplers, hence we again gain in computation. While parallel implementation of the proposed method is out of the scope of this paper, it may be noted that the order of computation is much smaller for the same when compared to conventional PFs. Moreover, the method is expected to scale well against the recently proposed network resampler, as the latter involves random resampling in only $\log_2(N)$ particles, while the former does not involve even so much.

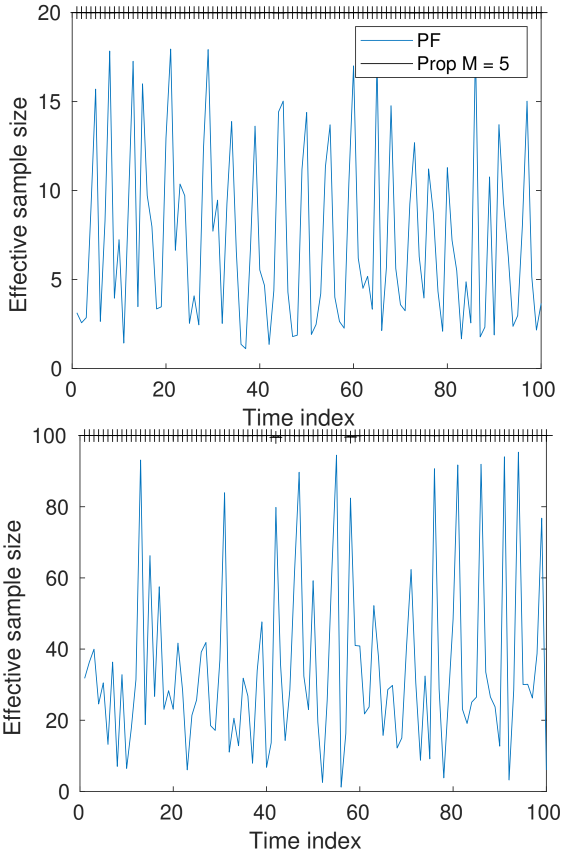

Fig. 3. The effective sample size versus the time index for $N = 20$ and $N = 100$ particles. The legend of the above panel applies to the bottom also.

V. Conclusion

This paper proposed a novel scheme of sampling multiple child particles per parent particle within the PF framework and then determining the most important ones using the proportional allocation rule. This scheme totally bypasses the computationally cumbersome resampling step while maintaining low degeneracy throughout the simulation. The benefit of the proposed method, shown using simulations, is that it is able to faithfully represent the posterior PDF with fewer particles when compared with the existing methods. Its execution is completely parallel friendly as it processes the particles in batches making computation easier and consumes less time. The merit of this method is that it helps enhance the system performance in real-time localization and tracking applications. In the future, we aim to use the principles of network theory to reduce the communication within particles and make the method more parallel-friendly.

References

[1] Ronald P.S. Mahler, “Multitarget Bayes filtering via first-order multitarget moments,” IEEE Transactions on Aerospace and Electronic systems, Vol. 39, No. 4, pp. 1152–1178, 2003.

[2] N. Gordon, D.J. Salmond, and A.F.M. Smith, “Novel approach to nonlinear/non-Gaussian Bayesian state estimation,” In IEE proceedings F (radar and signal processing), vol. 140, no. 2, pp. 107–113. 1993.

[3] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, “A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking,” IEEE Trans. Signal Proc., vol 50, no. 2, pp. 174–188, 2002.

[4] R. Douc, and O Cappe, “Comparison of resampling schemes for particle filtering,” In Proc. IEEE Symp. on Image and Signal Processing and Analysis, pp. 64–69, 2005.

[5] J. D. Hol, T. B. Schon, and F. Gustafsson, “On resampling algorithms for particle filters,” Proc. 2006 IEEE Workshop on Nonlinear Statistical Signal Proc., pp. 79–82. 2006.

[6] G. Kitagawa, “Monte Carlo filter and smoother for non-Gaussian nonlinear state space models,” J. of Computational and Graphical Statistics, pp. 1–25,

[7] J. Liu, and R. Chen, ”Sequential Monte Carlo methods for dynamic systems,” J. American Statistical Association, Vol. 93, No. 443, pp. 1032–1044, 1998.

[8] P. B. Choppala, P. D. Teal, and M. R. Frean, “Soft resampling for improved information retention in particle filtering,” In Proc. IEEE IEEE International Conference on Acoustics, Speech and Signal Processing, Canada, 2013.

[9] P. B. Choppala, P. D. Teal, and M. R. Frean, “Soft systematic resampling for accurate posterior approximation and increased information retention in particle filtering,” In Proc. IEEE Workshop on Statistical Signal Processing, pp. 260-263, 2014.

[10] Tiancheng Li, Tariq P. Sattar, and Dedong Tang, “A fast resampling scheme for particle filters,” in Proc. IEEE Constantinides International Workshop on Signal Processing, London, 2013.

[11] P. B. Choppala, P. D. Teal, M. R. Frean, “Particle filter parallelisation using random network resampling,” in Proc. IEEE Information Fusion, Spain, 2014.

[12] T. Li, M. Bolic, and Petar M. Djuric, “Resampling methods for particle filtering: classification, implementation and strategies” in IEEE Signal processing magazine, Vol. 32, No. 3, pp. 70–86, 2015.

[13] L. M. Murray, Anthony Lee, and Pierre E. Jacob, “Parallel resampling in the particle filter, ” J. of Computational and Graphical Statistics, vol 25, no. 3, pp.789–805, 2016.

[14] Mehdi Chitchian, Andrea Simonetto, Alexander S. van Amesfoort, and T. Keviczky, “Distributed Computation Particle Filters on GPU Architectures for Real-Time Control Applications,” IEEE Trans. Control Systems Technology, vol 21, no. 6, pp. 2224–2238, 2013.

[15] A. Varsi, J Taylor, L Kekempanos, E. Knapp, and S. Maskell, “A Fast Parallel Particle Filter for Shared Memory Systems, ” IEEE. Signal Proc. Letters, Vol. 27, pp. 1570–1574, 2020.

[16] P. B. Choppala, P. D. Teal, and M. R. Frean, “Resampling and Network Theory,” IEEE Trans. Signal and Information Processing over Networks, vol 08, pp. 106–119, 2022.

[17] M. Pitt, and N. Shephard, “Filtering via simulation: Auxiliary particle filters,” J. American statistical Association, vol 94, no. 446, pp. 590–599, 1999.

[18] V. Elvira, L. Martino, M. Bugallo, and P. M. Djuric, “Elucidating the auxiliary particle filter via multiple importance sampling, ” IEEE Signal Processing Magazine, vol 36, no. 6, pp. 145–152, 2019.

[19] J.P. Norton, and G. V. Veres, “Improvement of the particle filter by better choice of the predicted sample set,” Proc. of the IFAC, vol 35, no. 1, pp. 365–370, 2002.

[20] M. Lin, R. Chen, and J. S. Liu, “Lookahead strategies for sequential Monte Carlo,” J. Statistical Science, vol 28, no. 1, pp.69–94, 2013.