OPTIMISATION OF PARTICLE FILTERS USING SIMULTANEOUS PERTURBATION STOCHASTIC APPROXIMATION

†*Bao Ling Chan** - ‡*Arnaud Doucet* - †*Vladislav B.Tadic*

†The University of Melbourne, Dept of Electrical and Electronic Engineering,
Victoria 3010, Australia.
‡University of Cambridge, Dept of Engineering,
Cambridge, CB2 1PZ, UK.
Email:b.chan@ee.mu.oz.au - ad2@eng.cam.ac.uk - v.tadic@ee.mu.oz.au

ABSTRACT

This paper addresses the optimisation of particle filtering methods aka Sequential Monte Carlo (SMC) methods using stochastic approximation. First, the SMC algorithm is parameterised smoothly by a parameter. Second, optimisation of an average cost function is performed using Simultaneous Perturbation Stochastic Approximation (SPSA). Simulations demonstrate the efficiency of our algorithm.

1. INTRODUCTION

Many data analysis tasks revolve around estimating the state of a dynamic model where only inaccurate observations are available. As many real world models involve non-linear and non-Gaussian elements, optimal state estimation is a problem that does not typically admit a closed form solution. Recently, there have been a surge of interest in SMC methods to solve this estimation/filtering problem numerically. These methods utilise a large number say $N \gg 1$ of random samples (or particles) to represent the posterior probability distribution of interest. Numerous algorithms have been proposed in literature; see [1] for a book-length review. Although most algorithms converge asymptotically ($N \to \infty$) towards the optimal solution, their performance can vary by an order of magnitude for a fixed N. Current algorithms are designed to optimise certain local criteria such as the conditional variance on the importance weights or the conditional variance of the number of offspring. The effects of these local optimisations are unclear on standard performance measures of interest such as the average Mean Square Error (MSE).

In [2], a principled approach to optimise performance of the SMC methods is proposed. Assuming the SMC algorithm is parameterised "smoothly" by a parameter $\theta \in \Theta$ where Θ is an open subset of R^m. Under stability assumptions on the dynamic model of interest [3], the particles, their corresponding weights, the true state and the observation of the system form a homogenous and ergodic Markov chain. Performance measure can thus be defined as the expectation of a cost function with respect to the invariant distribution of this Markov chain which is parameterised by θ. The minimising θ^* for the cost function is obtained using the Robbins-Monro Stochastic Approximation (RMSA) technique. The RMSA technique requires one to be able to derive an estimate of the gradient; see [2] for details. However, this method suffers from several drawbacks. It involves a so-called score function whose variance increases over time and needs to be discounted. For some interesting parametrizations of the SMC algorithm, the computational complexity is of $O\left(N^2\right)$ which is prohibitive. Finally one would need to develop alternative gradient estimation techniques to incorporate non-differentiable stratified/systematic resampling steps or Metropolis-Hastings steps.

In this paper, we are proposing another stochastic approximation method namely the SPSA as an alternative means to optimise the SMC methods. SPSA is an efficient technique introduced by Spall [4] where the gradient is approximated using a randomized finite difference method. Contrary to standard finite difference method, one needs to compute only 2 estimates of the performance measure instead of $2m$ estimates; m being the dimension of θ. The use of the SPSA technique results in a very simple optimisation algorithm for the SMC methods.

The rest of the paper will be organized as follows: In section 2, a generic SMC algorithm is described and the performance measures of interest are introduced. In section 3, we describe the SPSA technique and show how it can be used to optimize the SMC algorithm. Finally, two examples are used in section 4 to demonstrate the efficiency of the optimisation procedure.

2. SMC METHODS FOR OPTIMAL FILTERING

2.1. State space model and a generic SMC method

Let $\{X_n\}_{n\geq 0}$ and $\{Y_n\}_{n\geq 0}$ be R^p and R^q-valued stochastic processes. The signals $\{X_n\}_{n\geq 0}$ is modelled as a Markov process of initial density $\mu(x_0)$ and transition density $f(x_n|x_{n-1})$. The observations $\{Y_n\}_{n\geq 0}$ are assumed to be conditionally independent given $\{X_n\}_{n\geq 0}$ and Y_n admits a marginal density $g(y_n|x_n)$. We denote for any process $\{Z_n\}_{n\geq 0}$, $Z^n \equiv (Z_0, Z_1, \cdots, Z_n)$. We are interested in estimating the signal X_n given the observation sequence Y^n. The filtering distribution $\Pr(X_n \in dx_n | Y^n) = p(x_n|Y^n)\,dx_n$, i.e. the conditional distribution of X_n given Y^n, satisfies the following recursion

$$p\left(x_n|Y^{n-1}\right) = \int f(x_n|x_{n-1})\, p\left(x_{n-1}|Y^{n-1}\right) dx_{n-1},$$

$$p(x_n|Y^n) = \frac{g(Y_n|x_n)\, p\left(x_n|Y^{n-1}\right)}{\int g(Y_n|x_n)\, p(x_n|Y^{n-1})\, dx_n}.$$

Except in very simple cases, this recursion does not admit a closed form solution and one needs to perform numerical approximations.

*Thanks to Cooperative Research Centre for Sensor Signal and Information Processing (CSSIP) for support.

SMC methods seek to approximate the true filtering distribution recursively with the weighted empirical distribution of a set of $N \gg 1$ samples $\hat{X}_n \equiv (\hat{X}_{n,1}, \hat{X}_{n,2}, \cdots, \hat{X}_{n,N})$, termed as particles with associated importance weights $\hat{W}_n \equiv (\hat{W}_{n,1}, \hat{W}_{n,2}, \cdots, \hat{W}_{n,N})$, $\hat{W}_{n,k} > 0$, $\sum_{k=1}^{N} \hat{W}_{n,k} = 1$

$$P_N(X_n \in dx_n | Y^n) = \sum_{k=1}^{N} \hat{W}_{n,k} \delta_{\hat{X}_{n,k}}(dx_n).$$

The particles are generated and weighted accordingly via a sequence of importance sampling and resampling steps. Assuming at time $n-1$, a set of particles $\hat{X}_{n-1}$ with weights $\hat{W}_{n-1}$ approximating $\Pr(X_{n-1} \in dx_{n-1} | Y^{n-1})$ is available. Let us introduce an importance sampling density, q where new particles $\tilde{X}_n \equiv (\tilde{X}_{n,1}, \tilde{X}_{n,2}, \cdots, \tilde{X}_{n,N})$ are sampled independently from

$$\tilde{X}_{n,k} \sim q\left(\hat{X}_{n-1,k}, Y_n, \bullet\right)$$

New normalized weights $\tilde{W}_n \equiv (\tilde{W}_{n,1}, \tilde{W}_{n,2}, \cdots, \tilde{W}_{n,N})$ are then evaluated to account for the discrepancy with the "target" distribution

$$\tilde{W}_{n,k} \propto \hat{W}_{n-1,k} \frac{g\left(Y_n | \tilde{X}_{n,k}\right) f\left(\tilde{X}_{n,k} \middle| \hat{X}_{n-1,k}\right)}{q\left(\hat{X}_{n-1,k}, Y_n, \tilde{X}_{n,k}\right)}.$$

In the resampling step, the particles $\tilde{X}_n$ are then multiplied/ eliminated accordingly to obtain $\hat{X}_n$ i.e.

$$\hat{X}_n = \left(\tilde{X}_{n,1}, \ldots, \tilde{X}_{n,1}, \ldots, \tilde{X}_{n,N}, \ldots, \tilde{X}_{n,N}\right)$$

where $\tilde{X}_{n,k}$ is copied $I_{n,k}$ times. The random variables $I_n \equiv (I_{n,1}, I_{n,2}, \cdots, I_{n,N})$ are sampled from a probability distribution $\Pr(I_n = i | \tilde{W}_n)$ where $i \equiv (i_1, i_2, \cdots, i_N)$. Several algorithms such as multinomial and systematic resampling have been proposed. These algorithms ensure that the number of particles is kept constant; i.e. $\sum_{k=1}^{N} I_{n,k} = N$. In the standard approaches, the weights $\hat{W}_{n,k}$ are then set to N^{-1}. However it is also possible to resample with weights proportional to say $\tilde{W}_{n,k}^{\alpha}$. In this case, the weights after resampling are proportional to $\tilde{W}_{n,k}^{1-\alpha}$.

This algorithm converges as $N \to \infty$ under very weak conditions [5]. However, for a fixed N, the performance is highly dependent on the choice of q and the resampling scheme. We assume here that one can parameterise smoothly the SMC algorithm by $\theta \in \Theta \subseteq R^m$. For example, this parameter can correspond to some parameters of the importance sampling density. The optimisation method described in this paper is based on the generic SMC algorithm outlined. However, one can easily extend the optimisation method to other complex algorithms such as the auxiliary particle filter or to algorithms including Markov chain Monte Carlo (MCMC) moves.

2.2. Performance measure

In this subsection, we define the performance measure to optimize with respect to θ. The key remark is that the current state X_n, the observation Y_n, the particles $\tilde{X}_n$ and the corresponding weights $\tilde{W}_n$ form a homogenous and ergodic Markov chain under some stability assumptions on the dynamic model of interest [3]. By assuming that this holds for any $\theta \in \Theta$, we can define a meaningful time average cost function $J(\theta)$ for the system

$$J(\theta) = E_\theta\left[f\left(Y, X, \tilde{X}, \tilde{W}\right)\right],$$

where the expectation is with respect to the invariant distribution of the Markov chain $\left(Y_n, X_n, \tilde{X}_n, \tilde{W}_n\right)$. We are interested in estimating

$$\theta^* = \arg\min J(\theta).$$

We emphasize here that these cost functions are independent of the observations; the observation process being integrated out. This has several important practical consequences. In particular, one can optimize the SMC algorithm off-line by simulating the data and then use the resulting optimized algorithm on real data. We consider here the following two cost functions to minimize but others can be defined without modifying the algorithm.

• *Mean Square Error (MSE)*

$$f(Y_n, X_n, \tilde{X}_n, \tilde{W}_n) = \left(X_n - \sum_{k=1}^{N} \tilde{X}_{n,k} \tilde{W}_{n,k}\right)^2.$$

It is of interest to minimize the average MSE between the true state and the Monte Carlo estimate of $E[X_n | Y^n]$. As discussed previously, although the true state X_n is unknown, one can simulate data in a training phase to estimate θ^* and then use the optimised SMC filter on real data.

• *Effective Sample Size (ESS)*

$$f(Y_n, X_n, \tilde{X}_n, \tilde{W}_n) = -\left(\sum_{k=1}^{N} \tilde{W}_{n,k}^2\right)^{-1}.$$

An appealing measure for the accuracy of a particle filter is its "effective sample size"; i.e. a measure of the uniformity of the importance weights. The larger the ESS is, the more particles are concentrated in the region of interest and hence the better the chance of the algorithm to respond to fast changes. The maximum value for ESS is N and is maximised when $\tilde{W}_{n,k} = N^{-1}$ for all k. We are interested in maximizing the ESS, that is minimizing its opposite given by $-\left(\sum_{k=1}^{N} \tilde{W}_{n,k}^2\right)^{-1}$.

3. OPTIMISATION OF SMC USING SPSA

3.1. Simultaneous Perturbation Stochastic Approximation

The problem of minimising a differentiable cost function $J(\theta)$, where $\theta \in \Theta \subseteq R^m$ can be translated into finding the zeros of the gradient $\nabla J(\theta)$. A recursion procedure to estimate θ^* such that $\nabla J(\theta) = 0$ proceeds as follows

$$\theta_n = \theta_{n-1} - \gamma_n \widehat{\nabla J_n} \tag{1}$$

where $\widehat{\nabla J_n}$ is the "noise corrupted" estimate of gradient $\nabla J(\theta)$ estimated at the point θ_{n-1} and $\{\gamma_n\}$ denotes a sequence of positive scalars such that $\gamma_n \to 0$ and $\sum_{n=1}^{\infty} \gamma_n = \infty$. Under appropriate conditions, the iteration in (1) will converge to θ^* in some stochastic sense (usually "almost surely"); see [4].

The essential part of (1) is how to obtain the gradient estimate $\widehat{\nabla J_n}$. The gradient estimation method in [2] can be computationally very intensive, as discussed in the introduction. We propose here to use the SPSA technique where the gradient is approximated via a finite difference method using only the estimates of the cost function of interest. The SPSA technique is a proven success among other finite difference methods with reduced number of estimates required for convergence; see [4]. The SPSA technique requires all elements of θ_{n-1} to be varied randomly simultaneously

to obtain two estimates of the cost function. Only two estimates are required regardless of the dimension m of the parameter. The two estimates required are of the form $J(\theta_{n-1} \pm perturbation)$ for a two-sided gradient approximation. In this case, the gradient estimate $\widehat{\nabla J_n} = \left(\widehat{\nabla J_{n,1}}, \ldots, \widehat{\nabla J_{n,m}}\right)^{\mathrm{T}}$ is given by

$$\widehat{\nabla J_{n,i}} = \frac{\widehat{J}(\theta_{n-1} + c_n\Delta_n) - \widehat{J}(\theta_{n-1} - c_n\Delta_n)}{2c_n\Delta_{n,i}}$$

where $\{c_n\}$ denotes a sequence of positive scalars such that $c_n \to 0$ and $\Delta_n = (\Delta_{n,1}, \Delta_{n,2}, \cdots, \Delta_{n,m})$ is a m-dimensional random perturbation vector. Careful selection of algorithm parameters γ_n, c_n and Δ_n is required to ensure convergence. The γ_n and c_n sequence generally take the form of $\gamma_n = a/(A+n)^\alpha$ and $c_n = c/n^\beta$ respectively with non-negative coefficients a, c, A, α and β. The practically effective values for α and β are 0.602 and 0.101. Each components of Δ_n is usually generated from a symmetric Bernoulli ± 1 distribution. See [6] for guidelines on coefficient selection.

3.2. Optimisation Algorithm using SPSA

We present here how to incorporate an optimisation algorithm using two-sided SPSA within a SMC framework. To optimise the parameters of a parametrised importance density, the algorithm proceeds as follows at time n:

Step 1: *Sequential importance sampling*
- For $n = 1$ to N, sample $\tilde{X}_{n,k} \sim q_{\theta_{n-1}}(\hat{X}_{n-1,k}, Y_n, \bullet)$.
- Compute the normalized importance weights as

$$\tilde{W}_{n,k} \propto \hat{W}_{n-1,k} \frac{g\left(Y_n | \tilde{X}_{n,k}\right) f\left(\tilde{X}_{n,k} \middle| \hat{X}_{n-1,k}\right)}{q_{\theta_{n-1}}\left(\hat{X}_{n-1,k}, Y_n, \tilde{X}_{n,k}\right)}$$

Step 2: *Cost function evaluation*
- Generate a m-dimensional simultaneous perturbation vector Δ_n.
- Compute $(\theta_{n-1} + c_n\Delta_n)$ and $(\theta_{n-1} - c_n\Delta_n)$.
- For $k = 1$ to N, sample $\tilde{X}_k^+ \sim q_{\theta_{n-1}+c_n\Delta_n}(\hat{X}_{n-1,k}, Y_n, \bullet)$.
- Compute the normalized importance weights as

$$\tilde{W}_k^+ \propto \hat{W}_{n-1,k} \frac{g\left(Y_n | \tilde{X}_k^+\right) f\left(\tilde{X}_k^+ \middle| \hat{X}_{n-1,k}\right)}{q_{\theta_{n-1}+c_n\Delta_n}\left(\hat{X}_{n-1,k}, Y_n, \tilde{X}_k^+\right)}$$

- For $k = 1$ to N, sample $\tilde{X}_k^- \sim q_{\theta_{n-1}-c_n\Delta_n}(\hat{X}_{n-1,k}, Y_n, \bullet)$
- Compute the normalized importance weights as

$$\tilde{W}_k^- \propto \hat{W}_{n-1,k} \frac{g\left(Y_n | \tilde{X}_k^-\right) f\left(\tilde{X}_k^- \middle| \hat{X}_{n-1,k}\right)}{q_{\theta_{n-1}-c_n\Delta_n}\left(\hat{X}_{n-1,k}, Y_n, \tilde{X}_k^-\right)}$$

- Evaluate cost function $J(\theta_{n-1}+c_n\Delta_n)$ and $J(\theta_{n-1}-c_n\Delta_n)$ from $\left\{\tilde{X}^+, \tilde{W}^+\right\}$ and $\left\{\tilde{X}^-, \tilde{W}^-\right\}$ respectively.

Step 3: *Gradient approximation*
- For $i = 1$ to m, evaluate gradient components as

$$\widehat{\nabla J_{n,i}} = \frac{\widehat{J}(\theta_{n-1} + c_n\Delta_n) - \widehat{J}(\theta_{n-1} - c_n\Delta_n)}{2c_n\Delta_{n,i}}$$

Step 4: *Parameter update*
- Update θ_{n-1} to new value θ_n as

$$\theta_n = \theta_{n-1} - \gamma_n\widehat{\nabla J_n}$$

Step 5: *Resampling*
- Multiply/discard particles $\tilde{X}_n$ with respect to high/low importance weights $\tilde{W}_n$ to obtain N particle $\hat{X}_n$.

It is possible to improve the algorithm in many ways for example by using iterates averaging or common random numbers. The idea behind common random numbers is to introduce a strong correlation between our estimates of $J(\theta_{n-1} - c_n\Delta_n)$ and $J(\theta_{n-1} + c_n\Delta_n)$ so as to reduce the variance of the gradient estimate; see [7] for details.

4. APPLICATION

We present two examples to illustrate the performance improvement brought by optimisation. The performance of the optimised filter is then compared to its un-optimised counterpart using the same signal and observation sequence. The following standard highly non-linear model [8] is used

$$X_n = \frac{1}{2}X_{n-1} + 25\frac{X_{n-1}}{1+(X_{n-1})^2} + 8\cos(1.2n) + V_n,$$
$$Y_n = \frac{X_n^2}{20} + W_n,$$

where $X_0 \sim \mathcal{N}(0,5)$, $V_n \overset{i.i.d}{\sim} \mathcal{N}(0,10)$ and $W_n \overset{i.i.d}{\sim} \mathcal{N}(0,1)$. For this model, we use the importance function obtained from local linearisation to incorporate the information from the observation

$$q_\theta(\hat{X}_{n-1,k}, Y_n, \tilde{X}_{n,k}) = \mathcal{N}\left(\tilde{X}_{n,k} : M_{n,k}(\theta), \Sigma_{n,k}(\theta)\right)$$

where $\Theta = R^2$ and

$$M_{n,k}(\theta) = \Sigma_{n,k}(\theta) f(\hat{X}_{n-1,k}) \left[\frac{1}{\theta_2} + \frac{Y_n + \frac{f^2(\hat{X}_{n-1,k})}{20}}{10}\right],$$
$$\Sigma_{n,k}(\theta) = \left[\frac{1}{\theta_2} + \frac{f^2(\hat{X}_{n-1,k})}{100}\right]^{-1},$$
$$f(\hat{X}_{n-1,k}) = \frac{1}{2}\hat{X}_{n-1,k} + \theta_1\frac{\hat{X}_{n-1,k}}{1+(\hat{X}_{n-1,k})^2} + 8\cos(1.2n).$$

The parameter $\theta = (\theta_1, \theta_2)$ forms part of the mean and the variance of the importance function. First, the SMC filter is optimised with respect to the ESS performance measure and then with respect to the MSE performance measure. Common random numbers method and systematic resampling scheme are employed in all simulations.

4.1. ESS optimisation

In Fig. 1, the optimum values of (θ_1, θ_2) are considerably different from the "un-optimised" values of $(25, 10)$ although θ has been initialised to $(25, 10)$. There is substantial improvement in terms of ESS, see Fig. 2, as θ converges to θ^*.

4.2. MSE optimisation

In Fig. 3, the optimum value of θ_1 is slightly larger than the initial value of 25. But, the optimum value of θ_2 is significantly different from the initial value of 10. Improvement in terms of MSE is observed in Fig. 4. However, the optimum values for (θ_1, θ_2) in term of MSE are considerably different from the values observed in section 4.1, ESS optimisation.

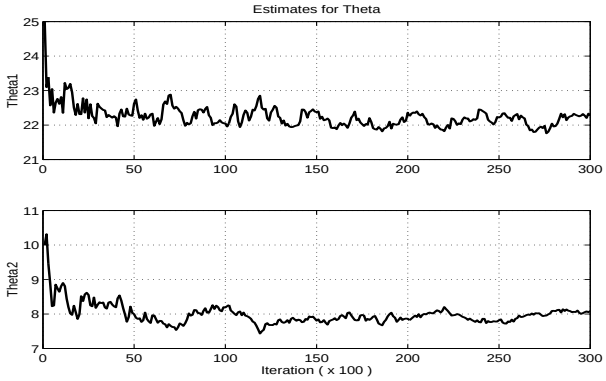

Fig. 1. Sequence of θ estimates over time

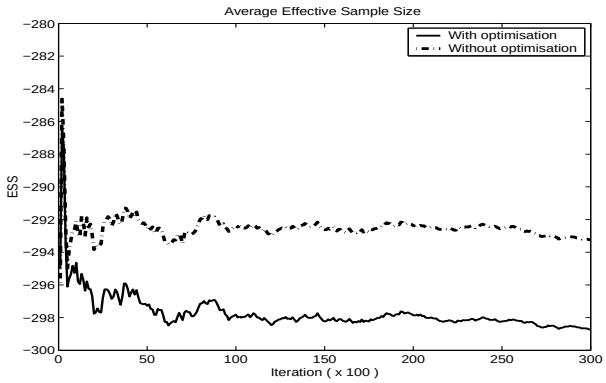

Fig. 2. Sequence of Average ESS estimates over time

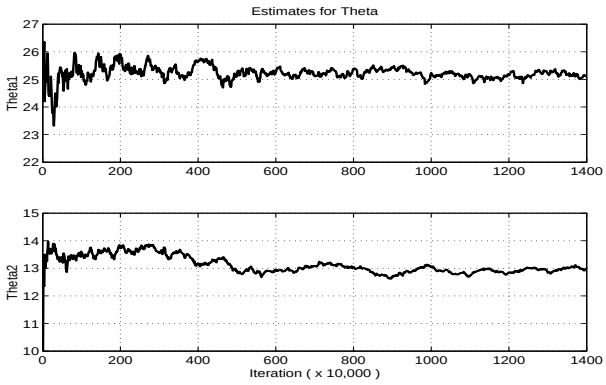

Fig. 3. Sequence of θ estimates over time

5. DISCUSSION

In this paper, we have demonstrated how to optimise in a principled way SMC methods. The minimising parameter for a particular performance measure can be easily obtained online by incorporating SPSA technique. No direct calculation of gradient is required. Advantages of SPSA over standard gradient estimation techniques can be summarised as such: relative ease of implementation and reduction in computational burden.

There are several potential extensions to this work. From an algorithmic perspective, it is of interest to speed up convergence by developing variance reduction methods for our gradient estimate. From a methodological perspective, the next logical step is to develop a self-adaptive algorithm where the parameter is not fixed but dependent on the current states of the particles. This is currently under study.

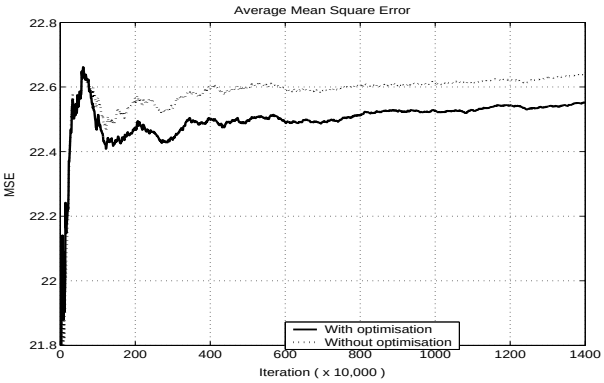

Fig. 4. Sequence of average MSE estimates over time

6. REFERENCES

[1] A. Doucet, J.F.G. de Freitas, and N.J. Gordon, *Sequential Monte Carlo Methods in Practice*, New York: Springer-Verlag, 2001.

[2] A. Doucet, and V.B. Tadić, "On-line optimization of sequential Monte Carlo methods using stochastic approximation", *Proc. American Control Conference*, May 2002.

[3] V.B. Tadić, and A. Doucet, "Exponential forgetting and geometric ergodicity in general state-space models", *Proc. IEEE Conference Decision and Control*, December 2002.

[4] J.C. Spall, "An overview of the simultaneous perturbation method for efficient optimisation," *John Hopkin Technical Digest*, vol. 19, no. 4, pp. 482-492, 1998.

[5] D. Crisan, and A. Doucet, "A survey of convergence results on particle filtering for practitioners", *IEEE Trans. Signal Processing*, vol. 50, no. 3, pp. 736-746, 2002.

[6] J.C. Spall, "Implementation of the simultaneous perturbation algorithm for stochastic optimisation", *IEEE Trans. Aerospace and Electronic Systems*, vol. 34, no. 3, 1998.

[7] N.L. Kleinman, J.C. Spall and D.Q. Naiman, "Simulation-based optimisation with stochastic approximation using common random numbers", *Management Science*, vol. 45, no. 11, pp. 1570-1578, 1999.

[8] G. Kitagawa, "Monte Carlo filter and smoother for non-Gaussian nonlinear state space models", *J. Comput. Graph. Statist.*, vol . 5, pp. 1-25, 1996.