Nonlinear/non-Gaussian sequential estimation applied to neural networks: Theory

Branimir Todorović[1], Claudio Moraga[2], Miomir Stanković[1], Branko Kovačević[3]

Abstract - **In this paper we consider a theoretical background of sequential Monte Carlo methods for nonlinear/non-Gaussian sequential estimation. Considered algorithms are applied to the state estimation of the nonlinear state space model in order to assess their performance quality and to stress the problem of degeneracy of the basic algorithm.**

Keywords - **Sequential Monte Carlo, neural networks, estimation, importance sampling, resampling**

I. INTRODUCTION

The problem of estimating the hidden state of a system using observations which arrive sequentially in time is very important in many fields of science, engineering and finance. The solution begins by modeling the state evolution and noise in the measurements. The resulting, so-called, state space model usually exhibits non-linear, non-Gaussian behavior.

The State Space Model (SSM) consists of two parts: The first one describes the evolution of the state $\{x_k, k = 1,2,...\}$:

$$x_k = f_k(x_{k-1}, d_k) \tag{1}$$

where $f_k : R^{n_x \times n_d} \to R^{n_x}$ is in general nonlinear function, and $\{d_k, k = 1,2,...\}$ is an i.i.d. process noise sequence, n_x, n_d are dimensions of the state and process noise vectors respectively. The second one is the measurement model:

$$y_k = h_k(x_k, v_k) \tag{2}$$

where $h_k : R^{n_x \times n_v} \to R^{n_y}$ is in general a nonlinear function, and $\{v_k, k = 1,2,...\}$ is an i.i.d. measurement noise sequence, n_y, n_v are dimensions of the measurement and measurement noise vectors respectively.

SSM of Neural Networks dynamic. Sequential estimation of neural networks (NN) using the state space model of networks dynamic and the extended Kalman filter (EKF) as an estimator has been thoroughly researched by several authors [9,2]. We have considered feed-forward [6] as well as recurrent network adaptation using EKF [7]. Due to the lack of space we give here as an example only the SSM model of a recurrent neural network. The transition equation describes the evolution of the network parameters and outputs while the measurement equation describes the influence of the measurement noise:

$$\begin{bmatrix} s_k \\ w_k \end{bmatrix} = \begin{bmatrix} f_k(s_{k-1}, w_{k-1}, u_k) \\ w_{k-1} \end{bmatrix} + \begin{bmatrix} d_{s,k} \\ d_{w,k} \end{bmatrix} \qquad y_k = s_k + v_k \tag{3}$$

where s_k is the n_s-dimensional network output; w_k is the n_w-dimensional vector of network parameters. $f_k(\cdot)$ is the nonlinear mapping realized by the recurrent network; $d_{s,k}$, $d_{w,k}$ are the process noises and v_k is the measurement noise. The hidden state of the network is defined as an augmented vector of network outputs and parameters.

Beside fast learning and built-in capability to deal with the time-varying parameters, we have demonstrated the possibility to unify the parameter and structure adaptation of the NN under the same algorithm based on EKF [6,7]. However ,this approach is based on the local linearization of the SSM, as well as Gaussian approximation of the relevant densities. In this paper we consider a theoretical background of sequential Monte Carlo algorithms as an alternative approach to the problem of neural network estimation. Considered algorithms are applied in the state estimation of nonlinear SSM in order to assess their performance quality and to stress the problem of degeneracy of the basic algorithm.

II. BAYESIAN ESTIMATION USING IMPORTANCE SAMPLING

Let us consider the problem of evaluating the hidden state process $x_{0:k} = \{x_k, k = 0,1,...\}$ using the observations $y_{0:k} = \{y_k, k = 0,1,...\}$. In a Bayesian framework all relevant information on $x_{0:k}$ is included in the posterior pdf $p(x_{0:k}/y_{0:k})$. The optimal (with respect to any criterion) estimation of the state and the measure of the accuracy of the estimate may be obtained from it. Assuming that the initial pdf, $p(x_0/y_0) = p(x_0)$, is available (y_0 being the set of no measurements) recursively in time, the density $p(x_{0:k}/y_{0:k})$ including the marginal filtering pdf $p(x_k/y_{0:k})$ and the expectation:

$$\mathrm{E}_{p(x_{0:k}/y_{1:k})}[f_k(x_{0:k})] = \int f_k(x_{0:k}) p(x_{0:k}/y_{0:k}) dx_{0:k} \tag{4}$$

for any $p(x_{0:k}/y_{1K})$-integrable $f_k : R^{(k+1)\cdot n_x} \to R$. Taking into account that the state corresponds to a Markov process: $p(x_{0:k}) = p(x_0)\prod_{j=1}^{k} p(x_j/x_{j-1})$, and the observations are conditionally independent given the states $p(y_{1:k}/x_{0:k}) = \prod_{j=1}^{k} p(y_j/x_j)$, the recursive formula for $p(x_{0:k}/y_{0:k})$ can be obtained:

[1]Faculty of Occupational Safety, University of Niš, 18000 Niš, Yugoslavia, E-mail: bssmtod@EUnet.yu
[2]Department of Artificial Intelligence, Polytechnical University of Madrid, Spain, and Department of Computer Science, University of Dortmund, Germany, E-mail: moraga@cs.uni-dortmund.de
[3]Faculty of Electrical Engineering, University of Belgrade, 11000 Belgrade, Yugoslavia, E-mail: kovacevic_b@kiklop.etf.bg.ac.yu

$$p(x_{0:k}/y_{0:k}) = \frac{p(y_k/x_k)\,p(x_k/x_{k-1})}{p(y_k/y_{0:k-1})}\,p(x_{0:k-1}/y_{0:k-1}) \qquad (5)$$

Unfortunately, in general the normalizing constant $p(y_k/y_{0:k-1})$ cannot be obtained analytically. A possible numerical solution consists of using a Monte Carlo integration method.

Perfect Monte Carlo simulations. Let us assume that we are able to simulate N i.i.d. random samples $\{x_{0:k}^{(i)}; i = 1,2,...N\}$ according to $p(x_{0:k}/y_{0:k})$. An empirical estimate of this pdf is given by: $\hat{p}(x_{0:k}/y_{0:k}) = \{\sum_{i=1}^{N}\delta_{x_{0:k}^{(i)}}(dx_{0:k})\}/N$ and the following estimate of expectation (4) is obtained:

$$\overline{\mathrm{E}_{p(x_{0:k}/y_{0:k})}[f_k(x_{0:k})]} = \frac{1}{N}\sum_{i=1}^{N} f_k(x_{0:k}^{(i)}) \qquad (6)$$

According to the strong law of the large numbers (SLLN) one obtains:

$$\overline{\mathrm{E}_{p(x_{0:k}/y_{0:k})}[f_k(x_{0:k})]} \xrightarrow[N\to\infty]{a.s.} \mathrm{E}_{p(x_{0:k}/y_{1:k})}[f_k(x_{0:k})] \qquad (7)$$

where a.s denotes "almost sure" convergence. However, it is often impossible to sample form the posterior distribution $p(x_{0:k}/y_{0:k})$. In our case, that pdf is what is unknown and it should be estimated.

Bayesian Importance Sampling. An alternative solution consists of using the importance sampling (IS) method based on choosing the so-called importance function, that is a pdf $\pi(x_{0:k}/y_{0:k})$ from which one can easily sample. A short description of the method follows. If $p(x_{0:k}/y_{1:k}) > 0$ implies $\pi(x_{0:k}/y_{0:k}) > 0$ then one can write:

$$\begin{aligned}\mathrm{E}_{p(\cdot/y_{1:k})}[f_k(x_{0:k})] &= \int f_k(x_{0:k})\frac{p(x_{0:k}/y_{1:k})}{\pi(x_{0:k}/y_{1:k})}\pi(x_{0:k}/y_{1:k})dx_{0:k} \\ &= \mathrm{E}_{\pi(\cdot/y_{1:k})}[f_k(x_{0:k})\omega^*(x_{0:k})]\end{aligned} \qquad (8)$$

where $\omega^*(x_{0:k})$ are referred to as importance weights:

$$\omega^*(x_{0:k}) = \frac{p(x_{0:k}/y_{0:k})}{\pi(x_{0:k}/y_{0:k})} \qquad (9)$$

The estimate of expectation $\mathrm{E}_{p(x_{0:k}/y_{1:k})}[f_k(x_{0:k})]$ can be obtained by simulating N i.i.d. samples $\{x_{0:k}^{(i)}; i = 1,2,...N\}$ according to $\pi(x_{0:k}/y_{0:k})$:

$$\begin{aligned}\overline{\mathrm{E}_{p(\cdot/y_{1:k})}[f_k(x_{0:k})]} &= \overline{\mathrm{E}_{\pi(\cdot/y_{1:k})}[f_k(x_{0:k})\omega^*(x_{0:k})]} \\ &= \frac{1}{N}\sum_{i=1}^{N} f_k(x_{0:k}^{(i)})\omega_k^{*(i)}\end{aligned} \qquad (10)$$

where the importance weights $\{\omega_k^{*(i)} = \omega^*(x_{0:k}^{(i)}); i = 1,2,...N\}$ are given by:

$$\omega_k^{*(i)} = \frac{p(x_{0:k}^{(i)}/y_{0:k})}{\pi(x_{0:k}^{(i)}/y_{0:k})} = \frac{p(y_{0:k}/x_{0:k}^{(i)})p(x_{0:k}^{(i)})}{p(y_{0:k})\pi(x_{0:k}^{(i)}/y_{0:k})} \qquad (11)$$

The estimate (10) is unbiased and according to SLLN has a.s. convergence toward $\mathrm{E}_{p(x_{0:k}/y_{1:k})}[f_k(x_{0:k})]$ when $N \to \infty$. However, the estimate (10) requires the knowledge of the normalizing constant $p(y_{0:k})$ that in general cannot be expressed in the closed form. One can solve the problem by introducing unnormalized weights of the form:

$$\omega_k(x_{0:k}) = \frac{p(y_{0:k}/x_{0:k})p(x_{0:k})}{\pi(x_{0:k}/y_{0:k})} \qquad (12)$$

which are proportional to the "true" importance weights $\omega_k \propto \omega_k^*$. The normalizing constant $p(y_{0:k})$ can be rewritten in the following form:

$$p(y_{0:k}) = \int \omega_k(x_{0:k})\pi(y_{0:k}/x_{0:k})\,dx_{0:k} \qquad (13)$$

By substituting (13) and (12) in (8), we obtain:

$$\mathrm{E}_{p(\cdot/y_{1:k})}[f_k(x_{0:k})] = \frac{\mathrm{E}_{\pi(\cdot/y_{1:k})}[f_k(x_{0:k})\omega_k(x_{0:k})]}{\mathrm{E}_{\pi(\cdot/y_{1:k})}[\omega_k(x_{0:k})]} \qquad (14)$$

and the estimate of $\mathrm{E}_{p(\cdot/y_{1:k})}[f_k(x_{0:k})]$ is given by:

$$\overline{\mathrm{E}_{p(\cdot/y_{1:k})}[f_k(x_{0:k})]} = \frac{\frac{1}{N}\sum_{i=1}^{N} f_k x_{0:k}^{(i)})\omega_k^{(i)}}{\frac{1}{N}\sum_{j=1}^{N}\omega_k^{(j)}} = \sum_{i=1}^{N} f_k(x_{0:k}^{(i)})\tilde{\omega}_k^{(i)} \qquad (15)$$

where $\tilde{\omega}_k^{(i)}$ are the normalized importance weights:

$$\tilde{\omega}_k^{(i)} = \omega_k^{(i)}(\textstyle\sum_{j=1}^{N}\omega_k^{(i)})^{-1} \qquad (16)$$

The "true" importance weights $\omega_k^{*(i)}$ have been replaced by the following estimate $\hat{\omega}_k^{*(i)} = N\tilde{\omega}_k^{(i)}$.

III. MONTE CARLO FILTER USING SEQUENTIAL IMPORTANCE SAMPLING

The aim of sequential Monte Carlo estimation is to obtain an estimate of $p(x_{0:k}/y_{0:k})$, and to be able to propagate it in time without modifying subsequently the past simulated trajectories $\{x_{0:k}^{(i)}; i = 1,...,N\}$. The following form of the importance function makes such scenario possible:

$$\begin{aligned}\pi(x_{0:k}/y_{0:k}) &= \pi(x_{0:k-1}/y_{0:k-1})\pi(x_k/x_{0:k-1}, y_{0:k}) \\ &= \pi(x_0/y_0)\prod_{j=1}^{k}\pi(x_j/x_{0:j-1}, y_{0:j})\end{aligned} \qquad (17)$$

By substituting (17) in (12) we obtain an expression for computing the importance weights recursively in time:

$$\omega_k(x_{0:k}) = \omega_{k-1}(x_{0:k-1})\frac{p(y_k/x_k)p(x_k/x_{k-1})}{\pi(x_k/x_{0:k}, y_{1:k})} \qquad (18)$$

If $\pi(x_k/x_{0:k-1}, y_{1:k}) = \pi(x_k/x_{k-1}, y_k)$ and if only a filtering pdf $p(x_k/y_{1:k})$ is regarded, the path $x_{0:k-1}^{(i)}$ can be discarded, as well as the history of observations $y_{0:k-1}$. The unnormalized importance weights are given by:

$$\omega_k^{(i)} \propto \omega_{k-1}^{(i)} \frac{p(y_k/x_k^{(i)})p(x_k^{(i)}/x_{k-1}^{(i)})}{\pi(x_k^{(i)}/x_{k-1}^{(i)}, y_k)} \tag{19}$$

Using sampled particles $x_k^{(i)} \sim \pi(x_k/x_{k-1}^{(i)}, y_k)$ and corresponding normalized weights $\tilde{\omega}_k^{(i)} = \omega_k^{(i)}(\sum_{j=1}^{N}\omega_k^{(i)})^{-1}$, $i = 1,...N$, the posterior filtering density is approximated by:

$$\hat{p}(x_k/y_{0:k}) = \sum_{i=1}^{N} \tilde{\omega}_k^{(i)} \delta(x_k - x_k^{(i)}) \tag{20}$$

and the estimate of $E_{p(\cdot/y_{0:k})}[f_k(x_k)]$ is obtained as:

$$\overline{E_{p(\cdot/y_{0:k})}[f_k(x_k)]} = \sum_{i=1}^{N} \tilde{\omega}_k^{(i)} f_k(x_k^{(i)}) \tag{21}$$

IV. Degeneracy of the SIS Algorithm

The best possible choice for importance function $\pi(x_{0:k}/y_{0:k})$ is the posterior density of interest $p(x_{0:k}/y_{0:k})$. In that case the mean and the variance of importance weights are respectively $\mathrm{E}_{\pi(\cdot/y_{0:k})}[\omega_k^*] = 1$ and $\mathrm{var}_{\pi(\cdot/y_{0:k})}(\omega_k^*) = 1$. However, the unconditional variance of the importance weights, obtained when the importance function is defined by (17), increases over time [3]. The proof for this statement is obtained by extending the Kong-Liu-Wang theorem [4] to the case of the importance function of the form (17) [3].

Practically this increase of the variance means that one of the importance weights will tend to one while other will tend to zero. Thus, the effective particle size reduces from N to almost 1 and the large portion of the computational power will be wasted on updating particles whose contribution to the approximation to $p(x_k/y_{0:k})$ is zero. The most common strategies to deal with this problem are the proper selection of importance function and resampling.

Selection of the importance function. The degeneracy of the SIS algorithm could be limited by selecting the importance function which will minimize the variance of the importance weights conditional upon the simulated trajectory $x_{0:k-1}^{(i)}$ and the observations $y_{0:k}$. Doucet [3] proved that importance function $p(x_k/x_{k-1}^{(i)}, y_k)$, introduced by Zaritski et all. [9] minimizes the variance of the importance weights $\omega_k^{*(i)}$ conditional upon $x_{0:k-1}^{(i)}$ and $y_{0:k}$. For this distribution unnormalized importance weights (12) are given by:

$$\omega_k^{(i)} = \omega_{k-1}^{(i)} \frac{p(y_k/x_k^{(i)})\, p(x_k^{(i)}/x_{k-1}^{(i)})}{p(x_k^{(i)}/x_{k-1}^{(i)}, y_k)} = \omega_{k-1}^{(i)} p(y_k/x_{k-1}^{(i)}) \tag{22}$$

Application of the optimal importance function requires sampling form $p(x_k/x_{k-1}^{(i)}, y_k)$ and evaluation, at least up to the proportionality, of $p(y_k/x_{k-1}^{(i)})$:

$$p(y_k/x_{k-1}^{(i)}) = \int p(y_k/x_k)\, p(x_k/x_{k-1}^{(i)})\, dx_k \tag{23}$$

Unfortunately, the analytical solution of the integral (23) in the general case cannot be obtained. However, as it was pointed out in [3], there is an important class of the state space models for which an analytical solution of (23) exists. As an example, the following state space model is considered.

$$\begin{aligned} x_k &= f_k(x_{k-1}) + d_k,\ d_k \sim N(0_{n_d \times 1}, Q_k) \\ y_k &= H_k x_k + v_k,\ v_k \sim N(0_{n_v \times 1}, R_k) \end{aligned} \tag{24}$$

The optimal importance function for SSM (25) is the Gaussian density:

$$p(x_k/x_{k-1}, y_k) = N(\bar{x}_k, \Sigma_k) \tag{25}$$

where

$$\begin{aligned} \Sigma_k &= (Q_k^{-1} + H_k^T R_k^{-1} H_k)^{-1} \\ \bar{x}_k &= \Sigma_k (Q_k^{-1} f(x_{k-1}) + H_k^T R_k^{-1} y_k) \end{aligned} \tag{26}$$

Density $p(y_k/x_{k-1})$ is obtained as:

$$\begin{aligned} p(y_k/x_{k-1}) \propto \exp(-0.5 \cdot (y_k - H_k f_k(x_{k-1}))^{\mathrm{T}} \cdot \\ (H_k Q_k H_k^T + R_k)^{-1} \cdot (y_k - H_k f_k(x_{k-1})) \end{aligned} \tag{27}$$

In case that SSM is given as:

$$\begin{aligned} x_k &= f_k(x_{k-1}) + d_k,\ d_k \sim N(0, Q_k) \\ y_k &= h_k(x_k) + v_k,\ v_k \sim N(0, R_k) \end{aligned} \tag{28}$$

by linearizing observation equation in $f_k(x_{k-1})$ we obtain:

$$\begin{aligned} x_k &= f_k(x_{k-1}) + d_k,\ d_k \sim N(0, Q_k) \\ y_k^* &= H_k(x_{k-1}) x_k + v_k,\ v_k \sim N(0, R_k) \end{aligned} \tag{29}$$

where $y_k^* = y_k - h_k(f_k(x_{k-1}))$ and $H_k = \partial h_k(x_k)/\partial x_k\big|_{x_k = f_k(x_{k-1})}$.

After linearization, the suboptimal importance function for SSM (29) can be obtained in form of (25).

Conventional particle filters use the transition prior as the importance density $\pi(x_k/x_{k-1}^{(i)}, y_k) = p(x_k/x_{k-1}^{(i)})$, which yields importance weights $\omega_k^{(i)} = \omega_{k-1}^{(i)} p(y_k/x_k^{(i)})$. This method is easy to implement but inefficient because the state space is explored without knowledge of the observations (the recent observation is not included in $p(x_k/x_{k-1}^{(i)})$). The inefficiency is especially significant in case of low observation noise, when the likelihood is peaked and the predicted state is near the likelihood's tail (in case of sudden changes in state dynamics). In that case the large number of particles will have low likelihood $p(y_k/x_k^{(i)})$ and consequently small importance, thus they will be wasted. Examples of state estimation in the last section of this paper will illustrate such behavior.

Resampling. The basic idea of resampling methods is to eliminate particles with small importance weights and multiply particles with large importance weights in order to limit the degeneracy of the sequential importance sampling algorithm. A new set of equally weighted samples $\{x_k^{(i)}, N^{-1}\}_{i=1}^{N}$ is obtained by resampling (with replacement) N times from an approximate discrete representation of $p(x_k/y_{0:k})$ given by (20), so that $\Pr(x_k^{(i)*} = x_k^{(j)}) = \tilde{\omega}_k^{(j)}$.

Some authors advocate the idea that resampling should be used only if the effective particle size is below a fixed threshold [4]. Otherwise, if the importance weights are nearly equal, resampling reduces the number of distinctive trajectories. Ef-

fective particle size is often used as a measure of the degeneracy of the algorithm. It is defined in [5] as:

$$N_{eff} = \frac{1}{1+\mathrm{var}_{\pi(\cdot/y_{0:k})}(\omega^*(x_{0:k}))} = \frac{N}{E_{\pi(\cdot/y_{0:k})}[\omega^*(x_{0:k})]} \le N \qquad (30)$$

Instead of N_{eff} which cannot be evaluated, in practice an estimate $\hat{N}_{eff}$ is used. This estimate is given by

$$\hat{N}_{eff} = \left(\sum_{i=1}^{N} (\tilde{\omega}_k^{(i)})^2\right)^{-1} \qquad (31)$$

When $\hat{N}_{eff}$ is below a fixed threshold N_{th}, it indicates the degeneracy case and a resampling step should be applied.

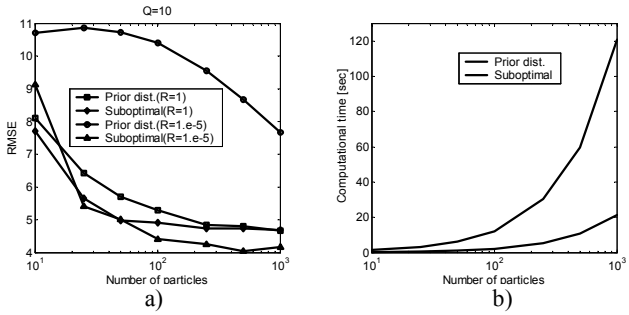
We have implemented and used in our experiments the stratified/systematic sampling described in [1]. A set of N points is sampled from a uniform distribution in the interval [0,1] each of the points a distance N^{-1} apart. The number of "children" of particle $x_k^{(i)}$ is the number of points that lie between $\sum_{j=1}^{i-1} \tilde{\omega}_k^{(j)}$ and $\sum_{j=1}^{i} \tilde{\omega}_k^{(j)}$.

V. STATE ESTIMATION EXAMPLES

In our preliminary experiments we have considered the problem of nonlinear state estimation. Our aim was to compare algorithms that use different importance functions, namely transition prior and suboptimal importance function. The considered nonlinear state space model is given as:

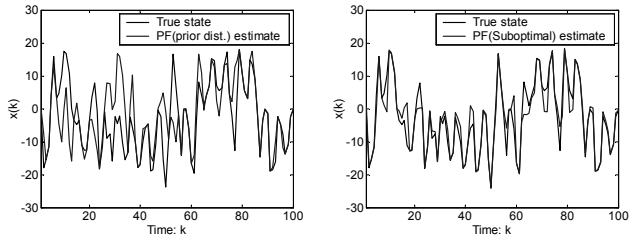
$$\begin{aligned} x_k &= f_k(x_{k-1}) + d_k \\ y_k &= x_k^2/20 + v_k \end{aligned} \qquad (32)$$

where $f_k(x_{k-1}) = x_{k-1}/2 + 25x_{k-1}/(1+x_{k-1}^2) + 8\cos(1.2k)$, $x_0 \sim N(0,5)$, d_k and v_k are mutually independent white Gaussian noises. We have conducted two experiments. In both of them the pdf of the process noise was $p(d_k) = N(0,10)$. As for the observation noise in first experiment it was $p(v_k) = N(0,1)$, while the second experiment was conducted for $p(v_k) = N(0,1.\mathrm{e}-5)$. The quality of the algorithms performance was measured using Root Mean Squared Error (RMSE). For both examples 100 simulations of length n=200 were considered. Both filters were tested for N=10, 25, 50, 100, 250, 500, 1000 particles. The resampling step was applied whenever $\hat{N}_{eff} < N/3$. Fig. 1 illustrates the results obtained in both experiments. We can see that algorithm with suboptimal importance function had lower RMSE at the price of much higher computational cost.



a) b)

Fig. 1. a) RMSE vs. number of particles b) Computational time vs. number of particles



a) Transition prior as import. fun. b) Suboptimal import. fun.

Fig. 2 State estimate in case of low measurement noise (R=1.e-5)

The algorithm with transition prior as importance function gave especially poor results in case of low measurement noise (R=1.e-5) as it was predicted in theory.

VI. CONCLUDING REMARKS

We have considered a theoretical background of sequential Monte Carlo methods for nonlinear/non-Gaussian sequential estimation applied to neural networks. Experiment in nonlinear state estimation showed that the choice of the importance function is crucial for the performance. Based on theoretical and experimental results given in this paper, the second paper on this subject [8] will introduce algorithms for nonlinear\non-Gaussian estimation applied to neural networks.

REFERENCES

[1] Carpenter, J., Clifford, P. and Fearnhead, P.: "An improved particle filter for nonlinear problems," Technical Report, Department of Statistics, Oxford University, England 1997

[2] de Freitas, J. F. G., Niranjan, M. and Gee, A.H.: "Hierarchical Bayesian-Kalman models for regularization and ARD in sequential learning," Technical Report CUED/F-INFENG/TR 307, Cambridge University., 1997.

[3] Doucet, A.: "On sequential simulation-based methods for Bayesian filtering, "Technical Report CUED/F-INFENG/TR 310, Cambridge University, 1998.

[4] Liu, J. S. and Chen, R.: "Sequential Monte Carlo methods for dynamic systems", *Jour. of Amer. Stat. Assoc.* Vol. 93, pp 1031-1041, 1998

[5] Kong, A., Liu, J. S. and Wong, W. H.: "Sequential imputations and Bayesian missing data problems," *Jour. of Amer. Stat. Assoc.*Vol. 89 (425), pp 278-288, 1994.

[6] Todorović, B., Stanković, M, Todorović-Zarkula, S.: Structurally adaptive RBF network in non-stationary time series prediction, In *Proc. IEEE AS-SPCC*, Lake Louise, Alberta, Canada, Oct. 1-4 (2000) pp. 224-229

[7] Todorović, B., Stanković, M., Moraga, C.: "Extended Kalman Filter trained Recurrent Radial Basis Function Network in Nonlinear System Identification," *Proc. of ICANN 2002*, Spain, August 2002

[8] Todorović, B., Moraga C., Stanković, M., Kovačević, B.: "Nonlinear/non-Gaussian sequential estimation applied to neural networks: Algorithms" *Proc. of ICEST 2002*, Niš, October 2002.

[9] Williams, R.J.: "Some observations on the use of the extended Kalman filter as a recurrent network learning algorithm," Technical Report NU_CCS_92-1. Boston: Northeastern University, College of Computer Science, 1992

[10] Zaritskii, V. S., Svetnik, V. B., Shimelevich, L.I.: "Monte Carlo Technique in Problems of Optimal Data Processing," Auto. Reo. Cont., vol. 12, 1975, pp. 95-103.