The Unscented Kalman Filter for Nonlinear Estimation

Eric A. Wan and Rudolph van der Merwe
Oregon Graduate Institute of Science & Technology
20000 NW Walker Rd, Beaverton, Oregon 97006
ericwan@ece.ogi.edu, rvdmerwe@ece.ogi.edu

Abstract

The *Extended Kalman Filter* (EKF) has become a standard technique used in a number of nonlinear estimation and machine learning applications. These include estimating the state of a nonlinear dynamic system, estimating parameters for nonlinear system identification (*e.g.*, learning the weights of a neural network), and dual estimation (*e.g.*, the Expectation Maximization (EM) algorithm) where both states and parameters are estimated simultaneously.

This paper points out the flaws in using the EKF, and introduces an improvement, the *Unscented Kalman Filter* (UKF), proposed by Julier and Uhlman [5]. A central and vital operation performed in the Kalman Filter is the propagation of a Gaussian random variable (GRV) through the system dynamics. In the EKF, the state distribution is approximated by a GRV, which is then propagated analytically through the first-order linearization of the nonlinear system. This can introduce large errors in the true posterior mean and covariance of the transformed GRV, which may lead to sub-optimal performance and sometimes divergence of the filter. The UKF addresses this problem by using a deterministic sampling approach. The state distribution is again approximated by a GRV, but is now represented using a minimal set of carefully chosen sample points. These sample points completely capture the true mean and covariance of the GRV, and when propagated through the *true* nonlinear system, captures the posterior mean and covariance accurately to the 3rd order (Taylor series expansion) for *any* nonlinearity. The EKF, in contrast, only achieves first-order accuracy. Remarkably, the computational complexity of the UKF is the same order as that of the EKF.

Julier and Uhlman demonstrated the substantial performance gains of the UKF in the context of state-estimation for nonlinear control. Machine learning problems were not considered. We extend the use of the UKF to a broader class of nonlinear estimation problems, including nonlinear system identification, training of neural networks, and dual estimation problems. Our preliminary results were presented in [13]. In this paper, the algorithms are further developed and illustrated with a number of additional examples.

This work was sponsored by the NSF under grant grant IRI-9712346

1. Introduction

The EKF has been applied extensively to the field of nonlinear estimation. General application areas may be divided into *state-estimation* and *machine learning*. We further divide machine learning into *parameter estimation* and *dual estimation*. The framework for these areas are briefly reviewed next.

State-estimation

The basic framework for the EKF involves estimation of the state of a discrete-time nonlinear dynamic system,

$$\mathbf{x}_{k+1} = F(\mathbf{x}_k, \mathbf{v}_k) \quad (1)$$

$$\mathbf{y}_k = H(\mathbf{x}_k, \mathbf{n}_k), \quad (2)$$

where $\mathbf{x}_k$ represent the unobserved state of the system and $\mathbf{y}_k$ is the only observed signal. The *process* noise $\mathbf{v}_k$ drives the dynamic system, and the *observation* noise is given by $\mathbf{n}_k$. Note that we are not assuming additivity of the noise sources. The system dynamic model F and H are assumed known. In state-estimation, the EKF is the standard method of choice to achieve a recursive (approximate) maximum-likelihood estimation of the state $\mathbf{x}_k$. We will review the EKF itself in this context in Section 2 to help motivate the Unscented Kalman Filter (UKF).

Parameter Estimation

The classic machine learning problem involves determining a nonlinear mapping

$$\mathbf{y}_k = G(\mathbf{x}_k, \mathbf{w}) \quad (3)$$

where $\mathbf{x}_k$ is the input, $\mathbf{y}_k$ is the output, and the nonlinear map G is parameterized by the vector $\mathbf{w}$. The nonlinear map, for example, may be a feedforward or recurrent neural network ($\mathbf{w}$ are the weights), with numerous applications in regression, classification, and dynamic modeling. Learning corresponds to estimating the parameters $\mathbf{w}$. Typically, a training set is provided with sample pairs consisting of known input and desired outputs, $\{\mathbf{x}_k, \mathbf{d}_k\}$. The error of the machine is defined as $\mathbf{e}_k = \mathbf{d}_k - G(\mathbf{x}_k, \mathbf{w})$, and the goal of learning involves solving for the parameters $\mathbf{w}$ in order to minimize the expected squared error.

While a number of optimization approaches exist (*e.g.*, gradient descent using backpropagation), the EKF may be used to estimate the parameters by writing a new state-space representation

$$\mathbf{w}_k = \mathbf{w}_{k-1} + \mathbf{u}_k \tag{4}$$
$$\mathbf{y}_k = G(\mathbf{x}_k, \mathbf{w}_k) + e_k. \tag{5}$$

where the parameters $\mathbf{w}_k$ correspond to a stationary process with identity state transition matrix, driven by process noise $\mathbf{u}_k$ (the choice of variance determines tracking performance). The output $\mathbf{y}_k$ corresponds to a nonlinear observation on $\mathbf{w}_k$. The EKF can then be applied directly as an efficient "second-order" technique for learning the parameters. In the linear case, the relationship between the Kalman Filter (KF) and Recursive Least Squares (RLS) is given in [3]. The use of the EKF for training neural networks has been developed by Singhal and Wu [9] and Puskorious and Feldkamp [8].

Dual Estimation

A special case of machine learning arises when the input $\mathbf{x}_k$ is unobserved, and requires coupling both state-estimation and parameter estimation. For these *dual estimation* problems, we again consider a discrete-time nonlinear dynamic system,

$$\mathbf{x}_{k+1} = F(\mathbf{x}_k, \mathbf{v}_k, \mathbf{w}) \tag{6}$$
$$\mathbf{y}_k = H(\mathbf{x}_k, \mathbf{n}_k, \mathbf{w}). \tag{7}$$

where both the system states $\mathbf{x}_k$ and the set of model parameters $\mathbf{w}$ for the dynamic system must be simultaneously estimated from only the observed noisy signal $\mathbf{y}_k$. Approaches to dual-estimation are discussed in Section 4.2.

In the next section we explain the basic assumptions and flaws with the using the EKF. In Section 3, we introduce the Unscented Kalman Filter (UKF) as a method to amend the flaws in the EKF. Finally, in Section 4, we present results of using the UKF for the different areas of nonlinear estimation.

2. The EKF and its Flaws

Consider the basic state-space estimation framework as in Equations 1 and 2. Given the noisy observation $\mathbf{y}_k$, a recursive estimation for $\mathbf{x}_k$ can be expressed in the form (see [6]),

$$\hat{\mathbf{x}}_k = (\text{prediction of } \mathbf{x}_k) + \mathcal{K}_k \cdot [\mathbf{y}_k - (\text{prediction of } \mathbf{y}_k)] \tag{8}$$

This recursion provides the optimal minimum mean-squared error (MMSE) estimate for $\mathbf{x}_k$ assuming the prior estimate $\hat{\mathbf{x}}_{k-1}$ and current observation $\mathbf{y}_k$ are Gaussian Random Variables (GRV). We need not assume linearity of the model. The optimal terms in this recursion are given by

$$\hat{\mathbf{x}}_k^- = E[F(\hat{\mathbf{x}}_{k-1}, \mathbf{v}_{k-1})] \tag{9}$$
$$\mathcal{K}_k = \mathbf{P}_{\mathbf{x}_k\mathbf{y}_k}\mathbf{P}_{\tilde{\mathbf{y}}_k\tilde{\mathbf{y}}_k}^{-1} \tag{10}$$
$$\hat{\mathbf{y}}_k^- = E[H(\hat{\mathbf{x}}_k^-, \mathbf{n}_k)], \tag{11}$$

where the optimal prediction of $\mathbf{x}_k$ is written as $\hat{\mathbf{x}}_k^-$, and corresponds to the expectation of a nonlinear function of the random variables $\hat{\mathbf{x}}_{k-1}$ and $\mathbf{v}_{k-1}$ (similar interpretation for the optimal prediction $\hat{\mathbf{y}}_k^-$). The optimal gain term $\mathcal{K}_k$ is expressed as a function of posterior covariance matrices (with $\tilde{\mathbf{y}}_k = \mathbf{y}_k - \hat{\mathbf{y}}_k^-$). Note these terms also require taking expectations of a nonlinear function of the prior state estimates.

The Kalman filter calculates these quantities exactly in the linear case, and can be viewed as an efficient method for analytically propagating a GRV through linear system dynamics. For nonlinear models, however, the EKF *approximates* the optimal terms as:

$$\hat{\mathbf{x}}_k^- \approx F(\hat{\mathbf{x}}_{k-1}, \bar{\mathbf{v}}) \tag{12}$$
$$\mathcal{K}_k \approx \hat{\mathbf{P}}_{\mathbf{x}_k\mathbf{y}_k}\hat{\mathbf{P}}_{\tilde{\mathbf{y}}_k\tilde{\mathbf{y}}_k}^{-1} \tag{13}$$
$$\hat{\mathbf{y}}_k^- \approx H(\hat{\mathbf{x}}_k^-, \bar{\mathbf{n}}), \tag{14}$$

where predictions are approximated as simply the function of the prior *mean* value for estimates (no expectation taken)[1] The covariance are determined by linearizing the dynamic equations ($\mathbf{x}_{k+1} \approx \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{v}_k$, $\mathbf{y}_k \approx \mathbf{C}\mathbf{x}_\mathbf{k} + \mathbf{D}\mathbf{n}_k$), and then determining the posterior covariance matrices analytically for the linear system. In other words, in the EKF the state distribution is approximated by a GRV which is then propagated analytically through the "first-order" linearization of the nonlinear system. The readers are referred to [6] for the explicit equations. As such, the EKF can be viewed as providing "first-order" approximations to the optimal terms[2]. These approximations, however, can introduce large errors in the true posterior mean and covariance of the transformed (Gaussian) random variable, which may lead to sub-optimal performance and sometimes divergence of the filter. It is these "flaws" which will be amended in the next section using the UKF.

3. The Unscented Kalman Filter

The UKF addresses the approximation issues of the EKF. The state distribution is again represented by a GRV, but is now specified using a minimal set of carefully chosen sample points. These sample points completely capture the true mean and covariance of the GRV, and when propagated through the *true* non-linear system, captures the posterior mean and covariance accurately to the 3rd order (Taylor series expansion) for *any* nonlinearity. To elaborate on this,

[1] The noise means are denoted by $\bar{\mathbf{n}} = E[\mathbf{n}]$ and $\bar{\mathbf{v}} = E[\mathbf{v}]$, and are usually assumed to equal to zero.

[2] While "second-order" versions of the EKF exist, their increased implementation and computational complexity tend to prohibit their use.

we start by first explaining the *unscented transformation*.

The unscented transformation (UT) is a method for calculating the statistics of a random variable which undergoes a nonlinear transformation [5]. Consider propagating a random variable $\mathbf{x}$ (dimension L) through a nonlinear function, $\mathbf{y} = g(\mathbf{x})$. Assume $\mathbf{x}$ has mean $\bar{\mathbf{x}}$ and covariance $\mathbf{P}_{\mathbf{x}}$. To calculate the statistics of $\mathbf{y}$, we form a matrix $\boldsymbol{\mathcal{X}}$ of $2L+1$ *sigma* vectors $\mathcal{X}_i$ (with corresponding weights W_i), according to the following:

$$\begin{aligned}
\mathcal{X}_0 &= \bar{\mathbf{x}} \\
\mathcal{X}_i &= \bar{\mathbf{x}} + \left(\sqrt{(L+\lambda)\mathbf{P}_{\mathbf{x}}}\right)_i \qquad i = 1,\ldots,L \\
\mathcal{X}_i &= \bar{\mathbf{x}} - \left(\sqrt{(L+\lambda)\mathbf{P}_{\mathbf{x}}}\right)_{i-L} \qquad i = L+1,\ldots,2L \\
W_0^{(m)} &= \lambda/(L+\lambda) \\
W_0^{(c)} &= \lambda/(L+\lambda) + (1-\alpha^2+\beta) \\
W_i^{(m)} &= W_i^{(c)} = 1/\{2(L+\lambda)\} \qquad i = 1,\ldots,2L
\end{aligned} \tag{15}$$

where $\lambda = \alpha^2(L+\kappa) - L$ is a scaling parameter. α determines the spread of the sigma points around $\bar{\mathbf{x}}$ and is usually set to a small positive value (*e.g.*, 1e-3). κ is a secondary scaling parameter which is usually set to 0, and β is used to incorporate prior knowledge of the distribution of $\mathbf{x}$ (for Gaussian distributions, $\beta = 2$ is optimal). $(\sqrt{(L+\lambda)\mathbf{P}_{\mathbf{x}}})_i$ is the ith row of the matrix square root. These sigma vectors are propagated through the nonlinear function,

$$\mathcal{Y}_i = g(\mathcal{X}_i) \quad i = 0,\ldots,2L \quad , \tag{16}$$

and the mean and covariance for $\mathbf{y}$ are approximated using a weighted sample mean and covariance of the posterior sigma points,

$$\bar{\mathbf{y}} \approx \sum_{i=0}^{2L} W_i^{(m)} \mathcal{Y}_i \tag{17}$$

$$\mathbf{P}_{\mathbf{y}} \approx \sum_{i=0}^{2L} W_i^{(c)} \{\mathcal{Y}_i - \bar{\mathbf{y}}\}\{\mathcal{Y}_i - \bar{\mathbf{y}}\}^T \tag{18}$$

Note that this method differs substantially from general "sampling" methods (*e.g.*, Monte-Carlo methods such as particle filters [1]) which require orders of magnitude more sample points in an attempt to propagate an accurate (possibly non-Gaussian) distribution of the state. The deceptively simple approach taken with the UT results in approximations that are accurate to the third order for Gaussian inputs for all nonlinearities. For non-Gaussian inputs, approximations are accurate to at least the second-order, with the accuracy of third and higher order moments determined by the choice of α and β (See [4] for a detailed discussion of the UT). A simple example is shown in Figure 1 for a 2-dimensional system: the left plot shows the true mean and covariance propagation using Monte-Carlo sampling; the center plots

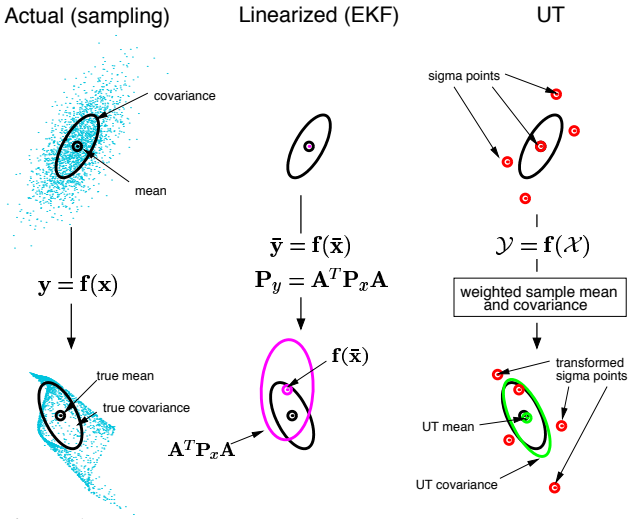

Figure 1: Example of the UT for mean and covariance propagation. a) actual, b) first-order linearization (EKF), c) UT.

show the results using a linearization approach as would be done in the EKF; the right plots show the performance of the UT (note only 5 sigma points are required). The superior performance of the UT is clear.

The *Unscented Kalman Filter* (UKF) is a straightforward extension of the UT to the recursive estimation in Equation 8, where the state RV is redefined as the concatenation of the original state and noise variables: $\mathbf{x}_k^a = [\mathbf{x}_k^T\ \mathbf{v}_k^T\ \mathbf{n}_k^T]^T$. The UT sigma point selection scheme (Equation 15) is applied to this new augmented state RV to calculate the corresponding sigma matrix, $\boldsymbol{\mathcal{X}}_k^a$. The UKF equations are given in Algorithm 3. Note that no explicit calculation of Jacobians or Hessians are necessary to implement this algorithm. Furthermore, the overall number of computations are the same order as the EKF.

4. Applications and Results

The UKF was originally designed for the state-estimation problem, and has been applied in nonlinear control applications requiring full-state feedback [5]. In these applications, the dynamic model represents a physically based parametric model, and is assumed known. In this section, we extend the use of the UKF to a broader class of nonlinear estimation problems, with results presented below.

4.1. UKF State Estimation

In order to illustrate the UKF for state-estimation, we provide a new application example corresponding to noisy time-series estimation.

In this example, the UKF is used to estimate an underlying clean time-series corrupted by additive Gaussian white noise. The time-series used is the Mackey-Glass-30 chaotic

Initialize with:

$$\hat{\mathbf{x}}_0 = E[\mathbf{x}_0]$$
$$\mathbf{P}_0 = E[(\mathbf{x}_0 - \hat{\mathbf{x}}_0)(\mathbf{x}_0 - \hat{\mathbf{x}}_0)^T]$$
$$\hat{\mathbf{x}}_0^a = E[\mathbf{x}^a] = [\hat{\mathbf{x}}_0^T \; \mathbf{0} \; \mathbf{0}]^T$$

$$\mathbf{P}_0^a = E[(\mathbf{x}_0^a - \hat{\mathbf{x}}_0^a)(\mathbf{x}_0^a - \hat{\mathbf{x}}_0^a)^T] = \begin{bmatrix} \mathbf{P}_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P_v} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{P_n} \end{bmatrix}$$

For $k \in \{1, \ldots, \infty\}$,

Calculate sigma points:

$$\boldsymbol{\mathcal{X}}_{k-1}^a = \left[\hat{\mathbf{x}}_{k-1}^a \;\; \hat{\mathbf{x}}_{k-1}^a \pm \sqrt{(L+\lambda)\mathbf{P}_{k-1}^a}\right]$$

Time update:

$$\boldsymbol{\mathcal{X}}_{k|k-1}^x = \mathbf{F}[\boldsymbol{\mathcal{X}}_{k-1}^x, \boldsymbol{\mathcal{X}}_{k-1}^v]$$
$$\hat{\mathbf{x}}_k^- = \sum_{i=0}^{2L} W_i^{(m)} \mathcal{X}_{i,k|k-1}^x$$
$$\mathbf{P}_k^- = \sum_{i=0}^{2L} W_i^{(c)} [\mathcal{X}_{i,k|k-1}^x - \hat{\mathbf{x}}_k^-][\mathcal{X}_{i,k|k-1}^x - \hat{\mathbf{x}}_k^-]^T$$
$$\boldsymbol{\mathcal{Y}}_{k|k-1} = \mathbf{H}[\boldsymbol{\mathcal{X}}_{k|k-1}^x, \boldsymbol{\mathcal{X}}_{k-1}^n]$$
$$\hat{\mathbf{y}}_k^- = \sum_{i=0}^{2L} W_i^{(m)} \mathcal{Y}_{i,k|k-1}$$

Measurement update equations:

$$\mathbf{P}_{\tilde{\mathbf{y}}_k \tilde{\mathbf{y}}_k} = \sum_{i=0}^{2L} W_i^{(c)} [\mathcal{Y}_{i,k|k-1} - \hat{\mathbf{y}}_k^-][\mathcal{Y}_{i,k|k-1} - \hat{\mathbf{y}}_k^-]^T$$
$$\mathbf{P}_{\mathbf{x}_k \mathbf{y}_k} = \sum_{i=0}^{2L} W_i^{(c)} [\mathcal{X}_{i,k|k-1} - \hat{\mathbf{x}}_k^-][\mathcal{Y}_{i,k|k-1} - \hat{\mathbf{y}}_k^-]^T$$
$$\mathcal{K} = \mathbf{P}_{\mathbf{x}_k \mathbf{y}_k} \mathbf{P}_{\tilde{\mathbf{y}}_k \tilde{\mathbf{y}}_k}^{-1}$$
$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathcal{K}(\mathbf{y}_k - \hat{\mathbf{y}}_k^-)$$
$$\mathbf{P}_k = \mathbf{P}_k^- - \mathcal{K} \mathbf{P}_{\tilde{\mathbf{y}}_k \tilde{\mathbf{y}}_k} \mathcal{K}^T$$

where, $\mathbf{x}^a = [\mathbf{x}^T \; \mathbf{v}^T \; \mathbf{n}^T]^T$, $\boldsymbol{\mathcal{X}}^a = [(\boldsymbol{\mathcal{X}}^x)^T \; (\boldsymbol{\mathcal{X}}^v)^T \; (\boldsymbol{\mathcal{X}}^n)^T]^T$, λ=composite scaling parameter, L=dimension of augmented state, $\mathbf{P_v}$=process noise cov., $\mathbf{P_v}$=measurement noise cov., W_i=weights as calculated in Eqn. 15.

Algorithm 3.1: Unscented Kalman Filter (UKF) equations

series. The clean times-series is first modeled as a nonlinear autoregression

$$x_k = f\big(x_{k-1}, ...x_{k-M}, \mathbf{w}\big) + v_k \tag{19}$$

where the model f (parameterized by $\mathbf{w}$) was approximated by training a feedforward neural network on the clean sequence. The residual error after convergence was taken to be the process noise variance.

Next, white Gaussian noise was added to the clean Mackey-Glass series to generate a noisy time-series $y_k = x_k + n_k$. The corresponding state-space representation is given by:

$$\mathbf{x}_k = F(\mathbf{x}_{k-1}, \mathbf{w}) + B \cdot v_{k-1}$$

$$\begin{bmatrix} x_k \\ x_{k-1} \\ \vdots \\ x_{k-M+1} \end{bmatrix} = \begin{bmatrix} f(x_{k-1}, \ldots, x_{k-M}, \mathbf{w}) \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \ddots & 0 & \vdots \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_{k-1} \\ \vdots \\ x_{k-M} \end{bmatrix} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \cdot v_{k-1}$$

$$y_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \cdot \mathbf{x}_k + n_k \tag{20}$$

In the estimation problem, the noisy-time series y_k is the only observed input to either the EKF or UKF algorithms (both utilize the known neural network model). Note that for this state-space formulation both the EKF and UKF are order L^2 complexity. Figure 2 shows a sub-segment of the estimates generated by both the EKF and the UKF (the original noisy time-series has a 3dB SNR). The superior performance of the UKF is clearly visible.

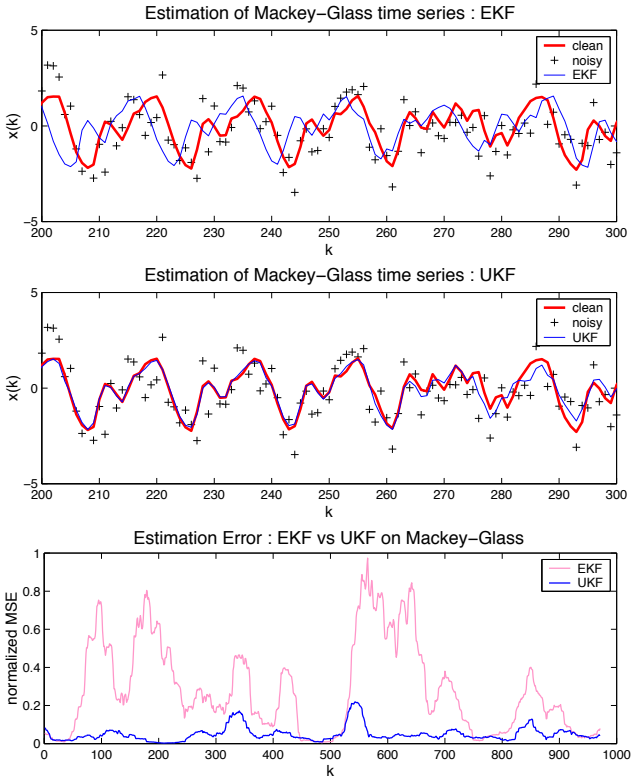

Figure 2: Estimation of Mackey-Glass time-series with the EKF and UKF using a known model. Bottom graph shows comparison of estimation errors for complete sequence.

4.2. UKF dual estimation

Recall that the dual estimation problem consists of simultaneously estimating the clean state $\mathbf{x}_k$ and the model pa-

rameters $\mathbf{w}$ from the noisy data y_k (see Equation 7). As expressed earlier, a number of algorithmic approaches exist for this problem. We present results for the Dual UKF and Joint UKF. Development of a Unscented Smoother for an EM approach [2] was presented in [13]. As in the prior state-estimation example, we utilize a noisy time-series application modeled with neural networks for illustration of the approaches.

In the the *dual extended Kalman filter* [11], a separate state-space representation is used for the signal and the weights. The state-space representation for the state $\mathbf{x_k}$ is the same as in Equation 20. In the context of a time-series, the state-space representation for the weights is given by

$$\mathbf{w}_k = \mathbf{w}_{k-1} + \mathbf{u}_k \tag{21}$$

$$y_k = f(\mathbf{x}_{k-1}, \mathbf{w}_k) + v_k + n_k. \tag{22}$$

where we set the innovations covariance $\mathbf{P_u}$ equal to $\mu\mathbf{P_w}$[3]. Two EKFs can now be run simultaneously for signal and weight estimation. At every time-step, the current estimate of the weights is used in the signal-filter, and the current estimate of the signal-state is used in the weight-filter. In the new *dual UKF* algorithm, both state- and weight-estimation are done with the UKF. Note that the state-transition is linear in the weight filter, so the nonlinearity is restricted to the measurement equation.

In the *joint extended Kalman filter* [7], the signal-state and weight vectors are concatenated into a single, *joint* state vector: $[\mathbf{x}_k^T\ \mathbf{w}_k^T]^T$. Estimation is done recursively by writing the state-space equations for the joint state as:

$$\begin{bmatrix}\mathbf{x}_k \\ \mathbf{w}_k\end{bmatrix} = \begin{bmatrix}F(\mathbf{x}_{k-1}, \mathbf{w}_{k-1}) \\ \mathbf{I}\cdot\mathbf{w}_{k-1}\end{bmatrix} + \begin{bmatrix}B\cdot v_k \\ \mathbf{u}_k\end{bmatrix} \tag{23}$$

$$y_k = \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix}\begin{bmatrix}\mathbf{x}_k \\ \mathbf{w}_k\end{bmatrix} + n_k, \tag{24}$$

and running an EKF on the joint state-space[4] to produce simultaneous estimates of the states $\mathbf{x}_k$ and $\mathbf{w}$. Again, our approach is to use the UKF instead of the EKF.

Dual Estimation Experiments

We present results on two time-series to provide a clear illustration of the use of the UKF over the EKF. The first series is again the Mackey-Glass-30 chaotic series with additive noise (SNR $\approx$ 3dB). The second time series (also chaotic) comes from an autoregressive neural network with random weights driven by Gaussian process noise and also corrupted by additive white Gaussian noise (SNR $\approx$ 3dB). A standard 6-4-1 MLP with $tanh$ hidden activation functions and a linear output layer was used for all the filters in the Mackey-Glass problem. A 5-3-1 MLP was used for the second problem. The process and measurement noise variances were assumed to be known. Note that in contrast to the state-estimation example in the previous section, only the noisy time-series is observed. A clean reference is never provided for training.

Example training curves for the different dual and joint Kalman based estimation methods are shown in Figure 3. A final estimate for the Mackey-Glass series is also shown for the Dual UKF. The superior performance of the UKF based algorithms are clear. These improvements have been found to be consistent and statistically significant on a number of additional experiments.

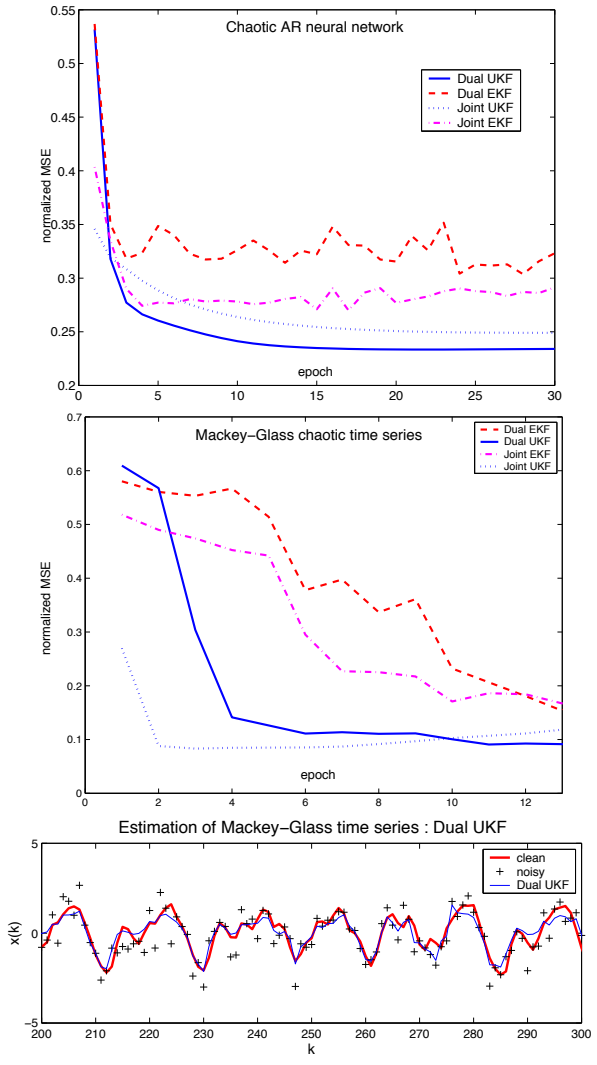

Figure 3: Comparative learning curves and results for the dual estimation experiments.

[3] μ is usually set to a small constant which can be related to the time-constant for RLS weight decay [3]. For a data length of 1000, $\mu \approx 1e-4$ was used.

[4] The covariance of $\mathbf{u}$ is again adapted using the RLS-weight-decay method.

4.3. UKF parameter estimation

As part of the dual UKF algorithm, we implemented the UKF for weight estimation. This represents a new parameter estimation technique that can be applied to such problems as training feedforward neural networks for either regression or classification problems.

Recall that in this case we write a state-space representation for the unknown weight parameters $\mathbf{w}$ as given in Equation 5. Note that in this case both the UKF and EKF are order L^2 (L is the number of weights). The advantage of the UKF over the EKF in this case is also not as obvious, as the state-transition function is linear. However, as pointed out earlier, the observation is nonlinear. Effectively, the EKF builds up an approximation to the expected Hessian by taking outer products of the gradient. The UKF, however, may provide a more accurate estimate through direct approximation of the expectation of the Hessian. Note another distinct advantage of the UKF occurs when either the architecture or error metric is such that differentiation with respect to the parameters is not easily derived as necessary in the EKF. The UKF effectively evaluates both the Jacobian and Hessian precisely through its sigma point propagation, without the need to perform any analytic differentiation.

We have performed a number of experiments applied to training neural networks on standard benchmark data. Figure 4 illustrates the differences in learning curves (averaged over 100 experiments with different initial weights) for the *Mackay-Robot-Arm* dataset and the *Ikeda* chaotic time series. Note the slightly faster convergence and lower final MSE performance of the UKF weight training. While these results are clearly encouraging, further study is still necessary to fully contrast differences between UKF and EKF weight training.

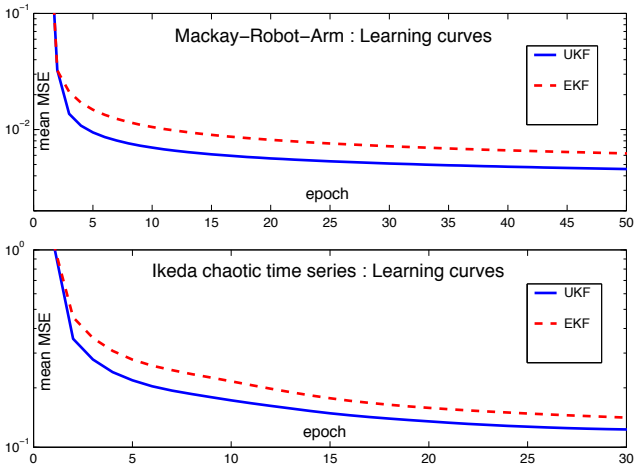

Figure 4: Comparison of learning curves for the EKF and UKF training. a) Mackay-Robot-Arm, 2-12-2 MLP, b) Ikeda time series, 10-7-1 MLP.

5. Conclusions and future work

The EKF has been widely accepted as a standard tool in the machine learning community. In this paper we have presented an alternative to the EKF using the unscented filter. The UKF consistently achieves a better level of accuracy than the EKF at a comparable level of complexity. We have demonstrated this performance gain in a number of application domains, including state-estimation, dual estimation, and parameter estimation. Future work includes additional characterization of performance benefits, extensions to batch learning and non-MSE cost functions, as well as application to other neural and non-neural (*e.g.*, parametric) architectures. In addition, we are also exploring the use of the UKF as a method to improve Particle Filters [10], as well as an extension of the UKF itself that avoids the linear update assumption by using a direct Bayesian update [12].

6. References

[1] J. de Freitas, M. Niranjan, A. Gee, and A. Doucet. Sequential monte carlo methods for optimisation of neural network models. Technical Report CUES/F-INFENG/TR-328, Dept. of Engineering, University of Cambridge, Nov 1998.

[2] A. Dempster, N. M. Laird, and D. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B39:1–38, 1977.

[3] S. Haykin. *Adaptive Filter Theory*. Prentice-Hall, Inc, 3 edition, 1996.

[4] S. J. Julier. The Scaled Unscented Transformation. *To appear in Automatica*, February 2000.

[5] S. J. Julier and J. K. Uhlmann. A New Extension of the Kalman Filter to Nonlinear Systems. In *Proc. of AeroSense: The 11th Int. Symp. on Aerospace/Defence Sensing, Simulation and Controls.*, 1997.

[6] F. L. Lewis. *Optimal Estimation*. John Wiley & Sons, Inc., New York, 1986.

[7] M. B. Matthews. A state-space approach to adaptive nonlinear filtering using recurrent neural networks. In *Proceedings IASTED Internat. Symp. Artificial Intelligence Application and Neural Networks*, pages 197–200, 1990.

[8] G. Puskorius and L. Feldkamp. Decoupled Extended Kalman Filter Training of Feedforward Layered Networks. In *IJCNN*, volume 1, pages 771–777, 1991.

[9] S. Singhal and L. Wu. Training multilayer perceptrons with the extended Kalman filter. In *Advances in Neural Information Processing Systems 1*, pages 133–140, San Mateo, CA, 1989. Morgan Kauffman.

[10] R. van der Merwe, J. F. G. de Freitas, A. Doucet, and E. A. Wan. The Unscented Particle Filter. Technical report, Dept. of Engineering, University of Cambridge, 2000. In preparation.

[11] E. A. Wan and A. T. Nelson. Neural dual extended Kalman filtering: applications in speech enhancement and monaural blind signal separation. In *Proc. Neural Networks for Signal Processing Workshop*. IEEE, 1997.

[12] E. A. Wan and R. van der Merwe. The Unscented Bayes Filter. Technical report, CSLU, Oregon Graduate Institute of Science and Technology, 2000. In preparation (http://cslu.cse.ogi.edu/nsel).

[13] E. A. Wan, R. van der Merwe, and A. T. Nelson. Dual Estimation and the Unscented Transformation. In S. Solla, T. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 666–672. MIT Press, 2000.