JOURNAL OF GUIDANCE, CONTROL, AND DYNAMICS
Vol. 22, No. 2, March–April 1999

# Nonlinear Projection Filter Based on Galerkin Approximation

Randal Beard,* John Kenney,† Jacob Gunther,‡ Jonathan Lawton,† and Wynn Stirling§
*Brigham Young University, Provo, Utah 84602*

**The conditional probability density function of the state of a stochastic dynamic system represents the complete solution to the nonlinear filtering problem because, with the conditional density in hand, all estimates of the state, optimal or otherwise, can be computed. It is well known that, for systems with continuous dynamics, the conditional density evolves, between measurements, according to Kolmogorov's forward equation. At a measurement, it is updated according to Bayes formula. Therefore, these two equations can be viewed as the dynamic equations of the conditional density and, hence, the exact nonlinear filter. In this paper, Galerkin's method is used to approximate the nonlinear filter by solving for the entire conditional density. Using a discrete cosine transform to approximate the projections required in Galerkin's method leads to a computationally realizable nonlinear filter. The implementation details are given and performance is assessed through simulations.**

## I. Introduction

ESTIMATING the state of a stochastic dynamic system from noisy observations is an important problem in engineering. The extensive work on this problem for linear systems was initiated by Kalman and Bucy.[1,2] In the decades since this early work, many important theoretical results for the linear problem have emerged and the linear filter has found wide practical application. However, since most systems are not truly linear, linear filtering theory does not apply directly to most physical systems.

In the more general setting of nonlinear systems, filter theory is less developed but has received attention from a number of researchers since the 1960s.[3–6] For linear systems with Gaussian inputs, the probability density function of the state conditioned on the measurements is Gaussian. Hence, optimal linear filters need only propagate the conditional mean and covariance that completely describes the density function. However, for nonlinear systems, the conditional density may not have a finite parameterization. General nonlinear filters must therefore propagate the entire density function.

The extended Kalman filter (EKF), a heuristic filter based on the linearized dynamics of a system,[7–10] has become the standard for state estimation of nonlinear systems. The EKF assumes that 1) deviations from the reference state trajectory are small, 2) the mathematical description of the system dynamics and observations is accurate, and 3) the conditional density function of the state is Gaussian. For large deviations from the reference trajectory, the EKF performs poorly or becomes unstable. Furthermore, if constructed from an erroneous model, the EKF state estimates can diverge from the true state. In addition, because of the nonlinear nature of the EKF, its estimate may depend on the initial conditions of the filter. In the final section of this paper, we show an example where the EKF gives erroneous results if it is not correctly initialized. These statements are not meant to discredit the EKF, the practical importance of which cannot be overstated, only to point out that the EKF is based on a linearization argument that, for severely nonlinear systems, can cause it to perform poorly. It should be noted there has been extensive work on alleviating some of the preceding problems.[11–15]

In this paper, we construct a nonlinear filter that approximates the exact nonlinear filter for systems with continuous nonlinear dynamics and discrete nonlinear observations. Accordingly, the paper represents a departure from current research directions in nonlinear filtering. Rather than pursue enhancements or modifications of the EKF, or explore dual relationships to various nonoptimal control algorithms, we investigate a method of approximating the exact nonlinear filtering problem. Note that if the exact nonlinear filtering problem could be solved directly or approximated efficiently, it would be widely used today in various industrial applications. We hope that this paper is a step in that direction.

The exact nonlinear filter consists of two dynamic equations[3]:

1) A partial differential equation (Kolmogorov's forward equation) that describes how the conditional density evolves between measurements, and

2) A difference equation (Bayes formula) that describes how it is modified by information supplied by new measurements.

To solve these equations we employ Galerkin's method, a classic procedure for approximating solutions of partial differential equations (PDEs).[16,17] In the context of the forward Kolmogorov equation, Galerkin's method was suggested by Risken[18] as a possible way to approximate the solution. However, a detailed analysis was not pursued.

Galerkin's method assumes that the exact solution to a PDE can be expanded as an infinite sum of basis elements. An approximate solution is found by truncating this sum and projecting the resulting error onto the finite subspace spanned by the basis elements used to approximate the solution.[16] To distinguish the approximate filter from the exact filter, and for lack of a better name, we will refer to the resulting filter as the nonlinear projection filter (NPF). Using a complex exponential basis as approximating elements, we show that the nonlinear filter can be implemented efficiently (for low-order systems) using discrete cosine transforms (DCT) resulting in a fast nonlinear filter that could be implemented in real time on a digital signal processor. Sinusoidal bases have been used before to implement Galerkin-based algorithms, but in much different contexts.[19–21] This work is also a natural extension of the application of Galerkin's method to optimal and robust control.[22–25] While there have been numerous studies that have applied numerical methods for solving PDEs,[26] we are not aware of a careful study of Galerkin's spectral method to the nonlinear filtering problem.

An important issue concerns the convergence of the Galerkin approximation. We shall prove that the approximation residual converges to zero as the dimension of the finite dimensional subspace used in the approximation tends to infinity. In other words, the NPF converges to the exact nonlinear filter. One limitation in our convergence result is that we do not obtain an explicit estimate of the approximation error. We only show that the approximation error can be made arbitrarily small by making the order of the approximation large enough.

Because the NPF is an approximation of the exact nonlinear filter, it can outperform the EKF for large variations in the state, for model

Received Dec. 8, 1997; revision received July 23, 1998; accepted for publication Oct. 5, 1998. 

*Assistant Professor, Department of Electrical and Computer Engineering.

†Research Assistant, Department of Electrical and Computer Engineering.

‡Research Assistant, Department of Electrical and Computer Engineering; currently Technical Staff, Merisoft, Provo, UT 84602.

§Professor, Department of Electrical and Computer Engineering.

mismatches, and for arbitrary initial conditions. In addition, because we propagate the conditional density, we can compute optimal state estimates for any criterion. In contrast, the EKF is only capable of approximating the conditional mean.

The remainder of the paper is organized as follows. The nonlinear filtering problem and the exact nonlinear filter is given in Sec. II. In Sec. III, Galerkin's method is used to approximate the exact nonlinear filter resulting in the NPF. In Sec. IV we show that the NPF converges to the exact nonlinear filter as the order of approximation increases. Practical implementation of the filter and the use of exponential basis elements is discussed in Sec. V. Finally, Sec. VI contains simulations comparing the NPF and the EKF.

## II. Exact Nonlinear Filter

Most physical systems evolve continuously in time while measurements may only be taken periodically at discrete time instants. Suppose the $n$-dimensional state $x_t$ of a continuous nonlinear stochastic dynamic system satisfies

$$\mathrm{d}x_t = f(t, x_t)\,\mathrm{d}t + G(t, x_t)\,\mathrm{d}\beta_t, \qquad t \geq t_0 \tag{1}$$

where $\{\beta_t, t \geq t_0\}$ is a $p$-dimensional Brownian motion with covariance matrix $Q(t)\,\mathrm{d}t$. Let $m$-dimensional noisy measurements be made at discrete times $t_k$

$$y_k = h\left(x_{t_k}, t_k\right) + v_k, \qquad k = 1, 2, \ldots \tag{2}$$

where $\{v_k, k \geq 1\}$ is an $m$-dimensional white Gaussian sequence independent of $\mathrm{d}\beta_t$ with covariance matrix $R_k$. Define the collection of measurements taken up to and including time $t$ as $Y_t = \{y_k : t_k \leq t\}$. We seek equations of evolution for the conditional density $p(t, x \mid Y_t)$ because it summarizes all the statistical information about the state contained in the measurements $Y_t$ and the initial condition $p(t_0, x)$. From $p(t, x \mid Y_t)$, the conditional mean and variance can be computed, which for nonlinear systems generally depend on all of the higher order moments.

Between observations at $t_k$ and $t_{k+1}$, $p \equiv p(t, x \mid Y_t)$ diffuses according to Kolmogorov's forward equation

$$\frac{\partial p}{\partial t} = -\sum_{i=1}^{n} \frac{\partial (p f_i)}{\partial x_i} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 \left[p (G Q G^T)_{ij}\right]}{\partial x_i \partial x_j} \tag{3}$$

where either $p(t_0, x)$ or $p(t_k, x \mid Y_{t_k})$, the measurement update at $t_k$, is used as the initial condition, and where $(A)_{ij}$ is the $(i, j)$th element of the matrix $A$. At an observation, $p$ satisfies the difference equation

$$p\left(t_{k+1}, x \mid Y_{t_{k+1}}\right) = \frac{p(y_{k+1} \mid x)\, p\left(t_{k+1}^-, x \,\middle|\, Y_{t_k}\right)}{\int p(y_{k+1} \mid \xi)\, p\left(\xi, t_{k+1}^- \,\middle|\, Y_{t_k}\right) \mathrm{d}\xi} \tag{4}$$

where

$$p(y_{k+1} \mid x) = \frac{\exp\left\{-\frac{1}{2}[y_k - h(x, t_k)]^T R_k^{-1} [y_k - h(x, t_k)]\right\}}{\sqrt{(2\pi)^m \mid R_k \mid}}$$

and $t_{k+1}^-$ is the instant in time right before the $(k+1)$ sample. For a detailed derivation and discussion of these results, see Ref. 3.

Equations (3) and (4) represent dynamic equations for the exact nonlinear filter. Equation (3) is used to compute predictions between measurements, while measurements are used to update the information about the state via Eq. (4). Closed-form solutions to these equations are generally not available. One exception is the linear-Gaussian case: For a linear dynamic system driven by Gaussian noise, this filter reduces to the standard Kalman filter.[3] In the next section we show that Galerkin's approximation method can be used to reduce Eq. (3) to a linear ordinary differential equation and Eq. (4) to an algebraic update equation.

Other nonlinear filters have also been derived. In particular Zakai has developed a nonlinear filter that has been studied by a number of researchers.[27,28] The main reason that we have chosen to study the Kolmogorov equation instead of the Zakai equation is that the Zakai equation propagates a density that does not have unit mass. Therefore the solution must be normalized at each point in time before useful data can be extracted. Also related to this paper is the work of Daum,[29] which derives a finite-dimensional nonlinear filter for continuous processes with discrete measurements. However, Daum makes several restrictive assumptions, most notably the use of a linear observer. In addition, Daum does not discuss the solution of the forward Kolmogorov equation, rather he assumes that the solution is known. The global Galerkin approximation method outlined in this paper could possibly be used in conjunction with Daum's filter. For continuous processes with continuous measurements, Brigo et al.[30] used a method related to our's, where the Kushner–Stratonovich equation is projected onto the tangent space of a finite dimensional manifold of probability densities to produce a finite dimensional approximation of the full nonlinear filter.

## III. Approximate Filter

### A. Prediction Equation

In this section we use the global or spectral Galerkin approximation method to reduce Eq. (3) from a function of space and time to a function of time only. The result is a linear ordinary differential equation that is easy to solve numerically. Galerkin's method is discussed in most textbooks on PDEs. Particularly good introductions to the topic can be found in Refs. 16, 17, and 31.

To apply Galerkin's method, we first assume that the solution of Eq. (3) satisfies

$$p(t, x \mid Y_t) = \sum_{\ell=0}^{\infty} b_\ell(t) \phi_\ell(x)$$

where equality is in the sense of the $L_2$ norm, and where $\{\phi_\ell\}_{\ell=0}^{\infty}$ is a complete set of basis functions for $L_2$. We approximate $p$ by truncating the sum:

$$p_N(t, x \mid Y_n) \triangleq \sum_{\ell=0}^{N-1} c_\ell(t) \phi_\ell(x) \tag{5}$$

where the coefficients $c_\ell$ satisfy the projection equation

$$\int_\Omega \left\{ \frac{\partial p_N}{\partial t} + \sum_{i=1}^{n} \frac{\partial (p_N f_i)}{\partial x_i} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 \left[p_N (G Q G^T)_{ij}\right]}{\partial x_i \partial x_j} \right\} \phi_q \,\mathrm{d}x = 0$$

for $q = 0, \ldots, N-1$. Because $p$ does not have compact support, the set $\Omega$ should be $\mathbb{R}^n$. However, $p$ will be almost zero over most of $\mathbb{R}^n$. Practically we must select $\Omega$ to be a closed and bounded subset of $\mathbb{R}^n$. The convergence results in Sec. IV assume that $\Omega$ is a sufficiently large compact set. Interchanging summation and differentiation, we obtain

$$\sum_{\ell=0}^{N-1} \dot{c}_\ell \int_\Omega \phi_\ell \phi_q \,\mathrm{d}x = \sum_{\ell=0}^{N-1} c_\ell \left\{ -\sum_{i=1}^{n} \int_\Omega \frac{\partial [\phi_\ell f_i]}{\partial x_i} \phi_q \,\mathrm{d}x + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \int_\Omega \frac{\partial^2 \left[\phi_\ell (G Q G^T)_{ij}\right]}{\partial x_i \partial x_j} \phi_q \,\mathrm{d}x \right\} \tag{6}$$

for $q = 0, \ldots, N-1$. Equation (6) is a system of $N$ linear ordinary differential equations that may be written in matrix notation by defining

$$\boldsymbol{c} = [c_0, \ldots, c_{N-1}]^T, \qquad [\boldsymbol{M}]_{i,j} = \int_\Omega \phi_i \phi_j \,\mathrm{d}x \tag{7}$$

$$[\boldsymbol{A}_1(t)]_{i,j} = -\sum_{k=1}^{n} \int_\Omega \frac{\partial [\phi_j f_k]}{\partial x_k} \phi_i \,\mathrm{d}x \tag{8}$$

$$[\boldsymbol{A}_2(t)]_{i,j} = \frac{1}{2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \int_\Omega \frac{\partial^2 \left[\phi_j (G Q G^T)_{k\ell}\right]}{\partial x_k \partial x_\ell} \phi_i \,\mathrm{d}x \tag{9}$$

With these definitions, Eq. (6) can be rewritten as

$$\boldsymbol{M}\dot{\boldsymbol{c}} = [\boldsymbol{A}_1(t) + \boldsymbol{A}_2(t)]\boldsymbol{c}$$

Letting $\boldsymbol{A}_N(t) = \boldsymbol{M}^{-1}[\boldsymbol{A}_1(t) + \boldsymbol{A}_2(t)]$, we have

$$\dot{\boldsymbol{c}}(t) = \boldsymbol{A}_N(t)\boldsymbol{c}(t)$$

The initial condition for this ordinary differential equation can be obtained from either the initial density function $p(t_0, x)$ or from the measurement update $p(t_k, x \mid Y_{t_k})$. In Sec. III.B we will show how to obtain the initial coefficients in the case of a measurement update. At time $t = t_0$, the initial coefficients are obtained by choosing $\boldsymbol{c}(t_0)$ such that

$$\int_\Omega p(t_0, x)\phi_q \, \mathrm{d}x = \int_\Omega p_N\left(x, t_0 \,\middle|\, Y_{t_0}\right)\phi_q \, \mathrm{d}x$$

$$= \sum_{\ell=0}^{N-1} c_\ell(t_0) \int_\Omega \phi_\ell \phi_q \, \mathrm{d}x$$

for $q = 0, \ldots, N-1$. To write this in matrix notation define

$$\boldsymbol{s} \triangleq \left[\int_\Omega p(t_0, x)\phi_0 \, \mathrm{d}x, \ldots, \int_\Omega p(t_0, x)\phi_{N-1} \, \mathrm{d}x\right]^T$$

The initial conditions then become $\boldsymbol{M}\boldsymbol{c}(t_0) = \boldsymbol{s}$ or $\boldsymbol{c}(t_0) = \boldsymbol{M}^{-1}\boldsymbol{s}$.

We have therefore converted Eq. (3), describing the evolution of $p$ into a system of ordinary differential equations

$$\dot{\boldsymbol{c}}_N(t) = \boldsymbol{A}_N(t)\boldsymbol{c}_N(t), \qquad \boldsymbol{c}_N(t_0) = \boldsymbol{M}^{-1}\boldsymbol{s} \tag{10}$$

which describes the evolution of the coordinates $\boldsymbol{c}$ of $p_N$ in $\mathrm{span}\{\phi_\ell\}_0^{N-1}$. Furthermore, if $f$, $G$, and $Q$ are independent of time, then $\boldsymbol{A}_N(t) = \boldsymbol{A}_N$ and Eqs. (10) have the following simple solution that describes how $\boldsymbol{c}$ changes between measurements: $\boldsymbol{c}(t) = e^{\boldsymbol{A}_N(t-t_k)}\boldsymbol{c}(t_k)$, $t \in [t_k, t_{k+1})$.

### B. Measurement Update

At an observation, $p$ satisfies Eq. (4). Again we apply Galerkin's method and replace $p$ by $p_N$ in Eq. (4) and define the updated approximate conditional density

$$p_N\left(x, t_{k+1} \,\middle|\, Y_{t_{k+1}}\right) = \frac{p(y_{k+1} \mid x)\, p_N\left(x, t_{k+1}^- \,\middle|\, Y_{t_k}\right)}{\int_\Omega p(y_{k+1} \mid \xi)\, p_N\left(\xi, t_{k+1}^- \,\middle|\, Y_{t_k}\right) \mathrm{d}\xi}$$

Projecting this equation onto the space $\mathrm{span}\{\phi_\ell\}_0^{N-1}$, we have

$$\sum_{\ell=0}^{N-1} c_\ell(t_{k+1}) \int_\Omega \phi_\ell \phi_q \, \mathrm{d}x$$

$$= \frac{\sum_{\ell=0}^{N-1} c_\ell\left(t_{k+1}^-\right) \int_\Omega p(y_{k+1} \mid x)\phi_\ell \phi_q \, \mathrm{d}x}{\sum_{\ell=0}^{N-1} c_\ell\left(t_{k+1}^-\right) \int_\Omega p(y_{k+1} \mid x)\phi_\ell \, \mathrm{d}x}$$

for $q = 0, \ldots, N-1$. Or, in matrix notation,

$$\boldsymbol{c}(t_{k+1}) = \frac{\boldsymbol{M}^{-1}\boldsymbol{\Upsilon}(y_{k+1})\boldsymbol{c}\left(t_{k+1}^-\right)}{\boldsymbol{\upsilon}(y_{k+1})^T \boldsymbol{c}\left(t_{k+1}^-\right)} \tag{11}$$

where

$$[\boldsymbol{\Upsilon}(y_{k+1})]_{q,\ell} = \int_\Omega p(y_{k+1} \mid x)\phi_\ell \phi_q \, \mathrm{d}x \tag{12}$$

$$[\boldsymbol{\upsilon}(y_{k+1})]_\ell = \int_\Omega p(y_{k+1} \mid x)\phi_\ell \, \mathrm{d}x \tag{13}$$

Therefore, at a measurement update, the solution to Eq. (10) is restarted with initial condition $\boldsymbol{c}(t_k)$ obtained from Eq. (11). It is important to note that the matrices $\boldsymbol{\Upsilon}(y_{k+1})$ and $\boldsymbol{\upsilon}(y_{k+1})$ depend on the measurement at time $t_{k+1}$, and hence cannot be computed off-line like the matrix $\boldsymbol{A}_N$ in Eq. (10). To make this filter practical, we need to be able to compute the elements of $\boldsymbol{\Upsilon}(y_k)$ and $\boldsymbol{\upsilon}(y_k)$ quickly. In Sec. V, we will use complex exponential basis functions to reduce implementation of the projections to computing DCTs and inverse discrete cosine transforms (IDCTs). Before doing so, we address the convergence of the NPF.

## IV. Convergence

In this section, we state conditions that guarantee that the NPF given by Eqs. (10) and (11) converges (in the $L_2$ norm) to the exact nonlinear filter given by Eqs. (3) and (4). The essence of the proof is to show that the bound on the estimation error is given by a graph that looks like Fig. 1. The evolution of the error between the Galerkin approximation and the actual nonlinear filter is bounded by the curve, where $\epsilon$ can be made arbitrarily small by making the number of approximating terms $N$ large. The bound on the error grows exponentially in between each sampling period and then jumps a finite amount at each measurement update. The bound derived in this section and depicted in Fig. 1 is conservative.

To make the convergence arguments transparent, we introduce some special notation that will only be used in this section. Assume that $\Omega \subset \mathbb{R}^n$ is a closed and bounded subset of $\mathbb{R}^n$ and let $\hat{p}_\Omega$ be the truncation of $p$ to $\Omega$, i.e.,

$$\hat{p}_\Omega(t, x) = \begin{cases} p(t, x) & \text{if} \quad x \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

We assume that $p$ is integrable, which implies that for every $\epsilon > 0$, there exists a closed and bounded set $\Omega \subset \mathbb{R}^n$ such that $\|p - \hat{p}_\Omega\|_{L_2} < \epsilon$, for all $t \in [t_0, t_1)$.

We write the Kolmogorov equation as

$$\frac{\partial p}{\partial t} + \mathcal{L}(p) = 0$$

where

$$\mathcal{L}(p) \triangleq \sum_{i=1}^n \frac{\partial(pf_i)}{\partial x_i} - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \frac{\partial^2\left[p(GQG^T)_{ij}\right]}{\partial x_i \partial x_j}$$

We assume that the basis functions $\{\phi_\ell\}_{\ell=0}^\infty$ form a complete basis for $L_2(\Omega)$ and that $\int_\Omega \mathcal{L}(\phi_j)\phi_k \, \mathrm{d}x < \infty$ for all $(j, k)$. Using arguments similar to those outlined in Ref. 24, we can assume, without loss of generality, that the basis functions $\{\phi_\ell\}_{\ell=0}^\infty$ are orthonormal on the set $\Omega$. This assumption is made to simplify the convergence proof, but does not impose a practical limitation on the selection of the basis functions. Practically, the basis functions are only required to be linearly independent.

Because $\hat{p}_\Omega \in L_1 \cap L_2$ it has a Fourier expansion, which we denote as

$$\hat{p}_\Omega(t, x \mid Y_t) = \sum_{\ell=0}^\infty b_\ell(t)\phi_\ell(x) \tag{14}$$

Because for each $t$ in the interval of interest, $\hat{p}_\Omega$ satisfies Kolmogorov's equation on $\Omega$, we get

$$\int_\Omega \left[\frac{\partial(\hat{p}_\Omega - p_N)}{\partial t} + \mathcal{L}(\hat{p}_\Omega - p_N)\right]\phi_q \, \mathrm{d}x = 0$$

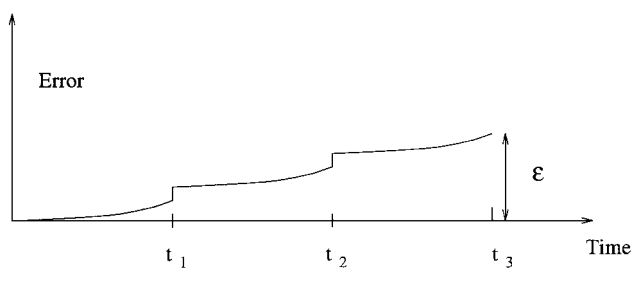


**Fig. 1 Bound on the evolution of the approximation error. $\epsilon$ can be made arbitrarily small by making the order of the approximation $N$ large enough.**

$q = 0, \ldots, N-1$. Using Eqs. (14) and (5) we get

$$\sum_{\ell=0}^{N-1}(\dot{b}_\ell - \dot{c}_\ell)\int_\Omega \phi_\ell\phi_q\,\mathrm{d}x + \sum_{\ell=0}^{N-1}(b_\ell - c_\ell)\int_\Omega \mathcal{L}(\phi_\ell)\phi_q\,\mathrm{d}x$$

$$= -\sum_{\ell=N}^{\infty}\dot{b}_\ell\int_\Omega \phi_\ell\phi_q\,\mathrm{d}x - \sum_{\ell=N}^{\infty} b_\ell\int_\Omega \mathcal{L}(\phi_\ell)\phi_q\,\mathrm{d}x$$

$q = 0, \ldots, N-1$. Define $\boldsymbol{b}_N = (b_0, \ldots, b_{N-1})^T$ and $\Phi_N = (\phi_0, \ldots, \phi_{N-1})^T$. Using the orthonormality of the basis functions and using the previously defined quantity $\boldsymbol{A}_N$, we obtain

$$(\dot{\boldsymbol{b}}_N - \dot{\boldsymbol{c}}_N) + \boldsymbol{A}_N(\boldsymbol{b}_N - \boldsymbol{c}_N) = -\sum_{\ell=N}^{\infty} b_\ell\int_\Omega \mathcal{L}(\phi_\ell)^{M^{-1}}\Phi_N\,\mathrm{d}x \quad (15)$$

Define

$$\mu_N = \left\| \sum_{\ell=N}^{\infty} b_\ell\int_\Omega \mathcal{L}(\phi_\ell)^{M^{-1}}\Phi_N\,\mathrm{d}x \right\|$$

The following theorem states the main result of this section.

*Theorem 1.* If $\Omega$ is compact, $p$ is integrable, and the basis functions $\{\phi_\ell\}_{\ell=0}^{\infty}$ form a complete orthonormal basis for $\mathcal{L}_2(\Omega)$, and if in addition

$$\frac{\mu_N}{\|\boldsymbol{A}_N\|}e^{\|\boldsymbol{A}_N\|} \to 0 \qquad \text{as} \qquad N \to 0$$

then given a fixed time $\hat{t} > 0$, and any $\epsilon > 0$, there exists a $K$ such that for $N > K$, $\|p(t, x \mid Y_t) - p_N(t, x \mid Y_t)\|_{L_2} < \epsilon$ for all $t < \hat{t}$.

The proof to this theorem is made transparent by the following three lemmas.

*Lemma 1.* Under the hypothesis of Theorem 1, for any $\epsilon > 0$, there exists a positive integer $K$ and a $\delta_k > 0$ such that if $N > K$ and $\|\boldsymbol{b}_N(t_k) - \boldsymbol{c}(t_k)\| < \delta_k$, then $\|\boldsymbol{b}_N(t) - \boldsymbol{c}(t)\| < \epsilon$, for all $t \in [t_k, t_{k+1}^-)$.

*Proof.* Since the differential equation $\dot{\xi} = -\boldsymbol{A}_N\xi$ is globally Lipschitz with Lipschitz constant $\|\boldsymbol{A}_N\|$, the theorem on the continuous dependence of ordinary differential equations on initial conditions and bounded disturbances,[32] and Eq. (15), imply that

$$\|\boldsymbol{b}_N(t) - \boldsymbol{c}_N(t)\| \le \delta_k e^{\|\boldsymbol{A}_N\|(t-t_k)} + \frac{\mu_N}{\|\boldsymbol{A}_N\|}\left(e^{\|\boldsymbol{A}_N\|(t-t_k)} - 1\right) \quad (16)$$

The first term can be made less than $\epsilon/2$ by making

$$\delta_k < (\epsilon/2)\exp\left[-\|\boldsymbol{A}_N\|\left(t_{k+1}^- - t_k\right)\right]$$

The hypothesis of Theorem 1 implies that there exists a $K$ such that $N > K$ implies that the second term is less than $\epsilon/2$. QED

*Lemma 2.* Under the hypothesis of Theorem 1, for any $\epsilon > 0$, there exists a positive integer $K$ and a $\delta_{k+1}^- > 0$ such that if $N > K$ and $\|\boldsymbol{b}_N(t_{k+1}^-) - \boldsymbol{c}_N(t_{k+1}^-)\| < \delta_{k+1}^-$, then $\|\boldsymbol{b}_N(t_{k+1}) - \boldsymbol{c}_N(t_{k+1})\| < \epsilon$.

*Proof.* Equation (11) shows that

$$\boldsymbol{c}(t_{k+1}) = \frac{\boldsymbol{M}^{-1}\mathbf{Y}(y_{k+1})\boldsymbol{c}\left(t_{k+1}^-\right)}{\boldsymbol{v}(y_{k+1})^T\boldsymbol{c}\left(t_{k+1}^-\right)} \quad (17)$$

If we truncate Eq. (4) to $\Omega$ and integrate, we obtain

$$\boldsymbol{b}(t_{k+1}) = \frac{\boldsymbol{M}^{-1}\mathbf{Y}(y_{k+1})\boldsymbol{b}\left(t_{k+1}^-\right) + \rho_{1N}}{\boldsymbol{v}(y_{k+1})^T\boldsymbol{b}\left(t_{k+1}^-\right) + \rho_{2N}} + \rho_{3N} \quad (18)$$

where

$$\rho_{1N} = \int_\Omega \sum_{\ell=N}^{\infty} b_\ell\, p(y_{k+1} \mid \xi)\phi_\ell(\xi)\Phi_N(\xi)\,\mathrm{d}\xi$$

$$\rho_{2N} = \int_\Omega \sum_{\ell=N}^{\infty} b_\ell\, p(y_{k+1} \mid \xi)\phi_\ell(\xi)\,\mathrm{d}\xi$$

$$\rho_{3N} = \int_\Omega \sum_{\ell=N}^{\infty} b_\ell\,\Phi(\xi)\,\mathrm{d}\xi$$

Since the Fourier series converges, $\|\rho_{1N}\|$, $|\rho_{2N}|$, and $\|\rho_{3N}\|$ can be made less than any arbitrary positive number for $N$ sufficiently large.

Subtracting Eq. (18) from Eq. (17) we obtain

$$\|\boldsymbol{b}_N(t_{k+1}) - \boldsymbol{c}_N(t_{k+1})\|$$

$$\le \left\| \frac{\boldsymbol{M}^{-1}\mathbf{Y}\boldsymbol{b}\left(t_{k+1}^-\right)}{\boldsymbol{v}^T\boldsymbol{b}\left(t_{k+1}^-\right) + \rho_{2N}} - \frac{\boldsymbol{M}^{-1}\mathbf{Y}\boldsymbol{c}\left(t_{k+1}^-\right)}{\boldsymbol{v}^T\boldsymbol{c}\left(t_{k+1}^-\right)} \right\|$$

$$+ \left\| \frac{\rho_{1N}}{\boldsymbol{v}^T\boldsymbol{b}\left(t_{k+1}^-\right) + \rho_{2N}} \right\| + \|\rho_{3N}\|$$

Define

$$\alpha \triangleq \frac{1}{\boldsymbol{v}^T\boldsymbol{b}\left(t_{k+1}^-\right) + \rho_{2N}}, \qquad \alpha_N \triangleq \frac{1}{\boldsymbol{v}^T\boldsymbol{c}\left(t_{k+1}^-\right)}$$

Using this notation we have the following sequence of inequalities:

$$\left\| \frac{\boldsymbol{M}^{-1}\mathbf{Y}\boldsymbol{b}\left(t_{k+1}^-\right)}{\boldsymbol{v}^T\boldsymbol{b}\left(t_{k+1}^-\right) + \rho_{2N}} - \frac{\boldsymbol{M}^{-1}\mathbf{Y}\boldsymbol{c}\left(t_{k+1}^-\right)}{\boldsymbol{v}^T\boldsymbol{c}\left(t_{k+1}^-\right)} \right\|$$

$$= \left\|\alpha\boldsymbol{M}^{-1}\mathbf{Y}\boldsymbol{b}\left(t_{k+1}^-\right) - \alpha_N\boldsymbol{M}^{-1}\mathbf{Y}\boldsymbol{c}\left(t_{k+1}^-\right)\right\|$$

$$\le \|\boldsymbol{M}^{-1}\mathbf{Y}\|\left\|\alpha\boldsymbol{b}\left(t_{k+1}^-\right) - \alpha_N\boldsymbol{c}\left(t_{k+1}^-\right)\right\|$$

$$= \|\boldsymbol{M}^{-1}\mathbf{Y}\|\left\|\alpha\left(\boldsymbol{b}\left(t_{k+1}^-\right) - \boldsymbol{c}\left(t_{k+1}^-\right)\right) + (\alpha - \alpha_N)\boldsymbol{c}\left(t_{k+1}^-\right)\right\|$$

$$\le \|\boldsymbol{M}^{-1}\mathbf{Y}\|\left[|\alpha|\left\|\boldsymbol{b}\left(t_{k+1}^-\right) - \boldsymbol{c}\left(t_{k+1}^-\right)\right\| + |\alpha - \alpha_N|\left\|\boldsymbol{c}\left(t_{k+1}^-\right)\right\|\right]$$

Since $p(y_{k+1} \mid x)$ is a Gaussian density we know that $\|\boldsymbol{M}^{-1}\mathbf{Y}(y_{k+1})\| = B_1 < \infty$ and $|\alpha| = B_2 < \infty$. We also know that $\|\boldsymbol{c}(t_{k+1}^-)\| = B_3 < \infty$, where $B_1$, $B_2$, and $B_3$ are finite numbers.

Also note that

$$|\alpha - \alpha_N| = \left| \frac{1}{\boldsymbol{v}^T\boldsymbol{b}\left(t_{k+1}^-\right) + \rho_{2N}} - \frac{1}{\boldsymbol{v}^T\boldsymbol{c}\left(t_{k+1}^-\right)} \right|$$

$$\le \left\| \frac{\boldsymbol{v}}{\left[\boldsymbol{v}^T\boldsymbol{b}\left(t_{k+1}^-\right) + \rho_{2N}\right]\left[\boldsymbol{v}^T\boldsymbol{c}\left(t_{k+1}^-\right)\right]} \right\|$$

$$\times \left\|\boldsymbol{b}_N\left(t_{k+1}^-\right) - \boldsymbol{c}_N\left(t_{k+1}^-\right)\right\|$$

$$+ \left| \frac{\rho_{2N}}{\left[\boldsymbol{v}^T\boldsymbol{b}\left(t_{k+1}^-\right) + \rho_{2N}\right]\left[\boldsymbol{v}^T\boldsymbol{c}\left(t_{k+1}^-\right)\right]} \right|$$

The first quantity is bounded by a finite number $B_4$, and the second quantity is a sequence, say $\rho_{4N}$ that converges in $N$. Therefore

$$|\alpha - \alpha_N| \le B_2\left\|\boldsymbol{b}_N\left(t_{k+1}^-\right) - \boldsymbol{c}_N\left(t_{k+1}^-\right)\right\| + \rho_{4N}$$

Collecting the preceding results we get

$$\|\boldsymbol{b}_N(t_{k+1}) - \boldsymbol{c}_N(t_{k+1})\| \le B_1\left[B_2\left\|\boldsymbol{b}_N\left(t_{k+1}^-\right) - \boldsymbol{c}_N\left(t_{k+1}^-\right)\right\|\right.$$

$$\left. + \left(B_4\left\|\boldsymbol{b}_N\left(t_{k+1}^-\right) - \boldsymbol{c}_N\left(t_{k+1}^-\right)\right\| + \rho_{4N}\right)B_2\right] + \rho_{3N} + \rho_{5N}$$

where

$$\rho_{5N} = \left\| \frac{\rho_{1N}}{\boldsymbol{v}^T\boldsymbol{b}\left(t_{k+1}^-\right) + \rho_{2N}} \right\|$$

The lemma follows from this formula. QED

*Lemma 3.* If at a particular instant of time $t$,

$$\|\boldsymbol{b}_N(t) - \boldsymbol{c}_N(t)\| \to 0$$

as $N \to \infty$, then

$$\|\hat{p}_\Omega(t, x \mid Y_t) - p_N(t, x \mid Y_t)\|_{L_2} \to 0$$

*Proof.*

$$\|\hat{p}_\Omega - p_N\|_{L_2}^2 = \int_\Omega |\hat{p}_\Omega - p_N|^2 \, dx \le \int_\Omega |(\boldsymbol{b}_N - \boldsymbol{c}_N)\Phi_N|^2 \, dx + \int_\Omega \left| \sum_{\ell=N}^{\infty} b_\ell \phi_\ell \right|^2 dx = (\boldsymbol{b}_N - \boldsymbol{c}_N)^T \boldsymbol{M} (\boldsymbol{b}_N - \boldsymbol{c}_N) + \int_\Omega \left| \sum_{\ell=N}^{\infty} b_\ell \phi_\ell \right|^2 dx$$

By the mean value theorem of calculus, there exists a $\xi \in \Omega$ such that

$$\|\hat{p}_\Omega - p_N\|_{L_2}^2 \le \|\boldsymbol{b}_N - \boldsymbol{c}_N\|^2 + \lambda(\Omega) \left| \sum_{\ell=N}^{\infty} b_\ell \phi_\ell(\xi) \right|^2$$

where $\lambda(\Omega)$ is the Lebesgue measure of $\Omega$. Because the second term on the right-hand side is the tail of a convergent Fourier series, it converges to zero. The lemma then follows from the hypothesis on $\|\boldsymbol{b}_N - \boldsymbol{c}_N\|$. QED

*Proof of Theorem 1.* First note that

$$\|p - p_N\|_{L_2} \le \|p - \hat{p}_\Omega\|_{L_2} + \|\hat{p}_\Omega - p_N\|_{L_2}$$

Because an appropriate choice of $\Omega$ guarantees that $\|p - \hat{p}_\Omega\|_{L_2} < \epsilon/2$ for all $t < \hat{t}$, we need to show that $\|\hat{p}_\Omega - p_N\|_{L_2} < \epsilon/2$ for all $t < \hat{t}$.

Given $t_k < \hat{t} \le t_{k+1}$ and $\epsilon$, Lemma 3 indicates that there exists a $\hat{\delta}$ such that if $\|\boldsymbol{b}_N(\hat{t}) - \boldsymbol{c}_N(\hat{t})\| < \hat{\delta}$, then $\|p(\hat{t}, x \mid Y_{\hat{t}}) - p_N(\hat{t}, x \mid Y_{\hat{t}})\|_{L_2} < \epsilon/2$. From Lemma 1, there exists a $K_k$ and a $\delta_k$ such that if $N > K_k$ and $\|\boldsymbol{b}_N(t_k) - \boldsymbol{c}(t_k)\| < \delta_k$, then $\|\boldsymbol{b}_N(\hat{t}) - \boldsymbol{c}_N(\hat{t})\| < \hat{\delta}$. From Lemma 2, there exists a $K_k^-$ and a $\delta_k^-$ such that if $N > K_k^-$ and $\|\boldsymbol{b}_N(t_k^-) - \boldsymbol{c}_N(t_k^-)\| < \delta_k^-$, then $\|\boldsymbol{b}_N(t_k) - \boldsymbol{c}(t_k)\| < \delta_k$. Another application of Lemma 1 implies that there exists a $K_{k-1}$ and a $\delta_{k-1}$ such that if $N > K_{k-1}$ and if $\|\boldsymbol{b}_N(t_{k-1}) - \boldsymbol{c}(t_{k-1})\| < \delta_{k-1}$, then $\|\boldsymbol{b}_N(t_k^-) - \boldsymbol{c}_N(t_k^-)\| < \delta_k^-$.

The preceding argument is repeated $k$ times, resulting in the requirement that $\|\boldsymbol{b}_N(0) - \boldsymbol{c}_N(0)\| < \delta_0$. If we assume that the basis functions are orthonormal, then at time $t = 0$ there is zero approximation error because

$$\hat{p}_\Omega(0, x \mid Y_0) = p_0(x)$$

and

$$\int_\Omega p_N(0, t \mid Y_0)\Phi_N \, dx = \int_\Omega p_0(x)\Phi_N \, dx$$

implies that

$$\|\boldsymbol{b}_N(0) - \boldsymbol{c}_N(0)\| = 0$$

Therefore letting

$$K = \max_{0 \le \ell \le k} \left\{ K_\ell, K_\ell^- \right\}$$

proves the theorem. QED

*Remark 1.* The bound derived in the preceding proof is given qualitatively by Fig. 1.

*Remark 2.* The requirement in the hypothesis of Theorem 1 that $(\mu_N / \|\boldsymbol{A}_N\|)e^{\|\boldsymbol{A}\|} \to 0$ is somewhat unsatisfactory because it is not clear at this stage how to guarantee that this is true a priori. The requirement, however, can be tested for any given system by computing $\boldsymbol{A}_N$ and $\mu_N$ for various values of $N$ and testing to see if the expression is converging to zero. Further research could focus on deriving conditions under which this hypothesis holds.

*Remark 3.* The convergence proof given in this section does not give explicit bounds on the approximation error. In other words, given a desired $\epsilon$, we cannot say what $K$ must be to guarantee that an approximation error of $\epsilon$ is achieved. We can only say that such a $K$ does in fact exist.

*Remark 4.* The convergence proof guarantees a small approximation error for any finite time. The arguments given in this section do not say anything about the steady-state error.

## V. Cosine Basis

In this section we show how to implement Eqs. (10) and (11), using a cosine basis and the DCT fast transform algorithm. The reasons for choosing a cosine series as the basis for the Hilbert space are 1) fast algorithms for implementing the DCT exist[33] and may be used to approximate the integrals of Sec. III, and 2) each function in this set satisfies boundary conditions that adequately approximates the boundary conditions that should be met for this problem. The boundary condition that should be imposed is that of a totally absorbing boundary similar to matching the impedance at the terminal end of a wave guide. Because systematic methods for determining such boundary conditions for PDEs of the form of Eq. (3) are elusive at this point, we impose the condition that the spatial derivatives be zero at the boundaries. This is accomplished with a cosine basis because each function in the set satisfies this condition. Simulations have shown that this is an adequate approximation. We now proceed to show how the DCT may be used to implement the NPF.

Because the general $n$-dimensional case is notationally messy and the basic ideas are contained in one dimension, we will restrict ourselves to this case. Suppose we choose $\Omega = [a, b]$, then the cosine basis set is

$$\{\phi_\ell(x)\}_{\ell=0}^{N-1} = \begin{cases} \dfrac{1}{\sqrt{b-a}} & \ell = 0 \\ \sqrt{\dfrac{2}{b-a}} \cos\left( \dfrac{2\pi\ell}{b-a}(x-a) \right) & 1 \le \ell \le N-1 \end{cases} \tag{19}$$

This set of basis functions for $S_N$ are the first $N$ functions in a complete orthonormal sequence that spans the Hilbert space $L_2[a, b]$.

The DCT algorithm can be used to approximate the sequence of inner products of the real valued function $\eta(x)$ with each of the basis functions (19) simultaneously as follows:

$$\int_a^b \eta(x)\phi_k(x) \, dx \approx \sqrt{\frac{b-a}{N}} \mathrm{DCT}_k[\eta(\zeta)] \tag{20}$$

where the input to the DCT algorithm is the $N$ samples of $\eta(\cdot)$ over the interval $[a, b]$ given by

$$\zeta = [(2\varsigma + 1)/2N](b - a) + a, \qquad \varsigma = 0, \ldots, N-1 \tag{21}$$

Therefore, the DCT returns $N$-weighted integrals of $\eta$ that would normally require $\mathcal{O}(N^2)$ operations; because the DCT is a fast transform [$\mathcal{O}(N \log N)$ similar to the fast Fourier transform], this results in substantial reduction in computational burden for large $N$.

With Eq. (20), we can use the DCT to approximate Eqs. (7–9) and Eqs. (12–13) as

$$\boldsymbol{M} = \boldsymbol{I}$$

$$[\boldsymbol{A}_1(t)]_\ell = \sqrt{\frac{b-a}{N}} \mathrm{DCT}\left[ \left( \frac{\partial \phi_\ell f}{\partial x} \right)(\zeta) \right]$$

$$[\boldsymbol{A}_2(t)]_\ell = \sqrt{\frac{b-a}{N}} \mathrm{DCT}\left\{ \frac{Q}{2} \left[ \frac{\partial^2 \left( \phi_\ell G^2 \right)}{\partial x^2} \right](\zeta) \right\}$$

$$\boldsymbol{s} = \frac{\sqrt{b-a}}{N} \mathrm{DCT}[p(\zeta, t_0)]$$

$$[\mathbf{Y}_k]_\ell = \frac{\sqrt{b-a}}{N} \mathrm{DCT}[p(y_k \mid \zeta)\phi_\ell(\zeta)]$$

$$\boldsymbol{v}(y_k) = \frac{\sqrt{b-a}}{N} \mathrm{DCT}[p(y_k \mid \zeta)]$$

These equations show that all of the quantities needed to propagate $\boldsymbol{c}(t)$ between measurements and to update it at a measurement can be computed quickly using the DCT.

In particular, the procedure for implementing the filter is as follows. First note that $\boldsymbol{M}$ is the identity for this choice of basis functions and thus may be ignored. Second we compute $\boldsymbol{c}(t_0)$ by arraying the samples of $p(\zeta, t_0)$ [where $\zeta$ takes on the values given by Eq. (21)] in a vector, applying the DCT algorithm, and scaling the output by $\sqrt{[(b-a)/N]}$. Each column of $\boldsymbol{A}_1(t_0)$ and $\boldsymbol{A}_2(t_0)$ is computed in a similar manner and their sum becomes $\boldsymbol{A}_N(t_0)$ (actually the integrands associated with corresponding elements of $\boldsymbol{A}_1(t_0)$ and $\boldsymbol{A}_2(t_0)$ may be summed prior to taking the DCT). As previously stated, if $f$, $G$, and $Q$ are independent of time, then so is $\boldsymbol{A}_N(t_0)$, so that $\boldsymbol{A}_N$ and its associate state transition matrix $e^{\boldsymbol{A}_N(t_1-t_0)}$ may be computed off-line; otherwise, they must be computed at each time step. The parameter vector $\boldsymbol{c}(t_1^-)$ for the conditional density just prior to the first measurement is given by

$$\boldsymbol{c}\left(t_1^-\right) = e^{\boldsymbol{A}_N(t_1-t_0)}c(t_0) \tag{22}$$

At the observation time $t_1$, the measurement update is performed as follows. First, the vector $\boldsymbol{\upsilon}(y_1)$ and the $\ell$th column of $\mathbf{Y}(y_1)$ for each $\ell = 0, \ldots, N-1$ are calculated in a similar manner as $\boldsymbol{c}(t_0)$ and the columns of $\boldsymbol{A}_N(t_0)$. These quantities cannot be computed off-line because they depend on the measurement. The variables $\mathbf{Y}(y_1)$ and $\boldsymbol{\upsilon}(y_1)$ are then used to compute the updated parameter vector $\boldsymbol{c}(t_1)$ according to Eq. (11). This vector is then used as the initial condition to Eq. (22) for the next time interval, and the cycle repeats.

As the output of our filter we may wish to compute the conditional mean, the conditional covariance or the density function itself. These quantities can also be computed efficiently using the DCT as shown next.

*Conditional Mean.* By definition the conditional mean of the state at time $t$ based on observations up to and including time $\tau \le t$ is

$$\bar{x}_t^\tau = \int \xi p(\xi, t \mid Y_\tau)\, \mathrm{d}\xi$$

We may compute the mean of our approximate conditional density as follows:

$$\hat{\bar{x}}_t^\tau = \int \xi p_N(\xi, t \mid Y_\tau)\, \mathrm{d}\xi = \int \xi \sum_{\ell=0}^{N-1} c_\ell(t)\phi_\ell(\xi)\, \mathrm{d}\xi$$

$$= \sum_{\ell=0}^{N-1} c_\ell(t) \int_\Omega x\phi_\ell\, \mathrm{d}x = \boldsymbol{\gamma}^T\boldsymbol{c}$$

where $\boldsymbol{\gamma} = \sqrt{[(b-a)/N]}\mathrm{DCT}[\zeta]$.

*Conditional Covariance.* The covariance of the approximate conditional density function is

$$\hat{P}_t^\tau = \int \xi^2 \hat{p}_N(\xi, t \mid Y_\tau)\, \mathrm{d}\xi$$

$$-\left(\hat{\bar{x}}_t^\tau\right)^2 = \sum_{\ell=0}^{N-1} c_\ell(t) \int_\Omega x^2\phi_\ell\, \mathrm{d}x - (\boldsymbol{\gamma}^T\boldsymbol{c})^2$$

$$= \boldsymbol{\Gamma}^T\boldsymbol{c} - (\boldsymbol{\gamma}^T\boldsymbol{c})^2$$

where $\boldsymbol{\Gamma} = \sqrt{[(b-a)/N]}\mathrm{DCT}[\zeta^2]$.

*Density Function.* To recover the approximate density function itself we have

$$p_N(x, t \mid Y_t) = \sum_{\ell=0}^{N-1} c_\ell(t)\phi_\ell(x) = \sqrt{\frac{N}{b-a}}\mathrm{IDCT}[\boldsymbol{c}(t)]$$

## VI. Examples

In this section we give simulation results comparing the NPF to the EKF for three continuous discrete filtering scenarios: 1) a standard nonlinear filtering problem, 2) an unobservable bimodal system, and 3) a two-dimensional system with an extremely poor sensor. The objective of the first example is to show that for a straightforward nonlinear system, the NPF and the EKF give similar results. The purpose of the second two examples is not to discredit the EKF, but rather to show that there are some simple examples whose nonlinearities are too difficult for the EKF, and yet can be handled by the NPF. In all three scenarios the EKF and the NPF have the same initial conditions and parameters. For comparison, state estimates are plotted with the true state trajectory.

### A. Nominal Example

This scenario compares the performance of the EKF and NPF for a standard filtering problem. The system dynamics and output equation are given by

$$\mathrm{d}x_t = \sin(x_t)\, \mathrm{d}t + \mathrm{d}\beta_t, \qquad y_k = x(t_k) + \upsilon_k$$

where $Q = 0.5$ and $R = 0.5$. The sampling period is $T = 0.1$ s. The initial state and the initial state estimate are both set to zero. The initial covariance estimate is one. The approximation interval is $\Omega = [-5\pi/2, 5\pi/2]$ and the number of Fourier basis elements is $N = 128$. Figure 2 shows the true state trajectory as well as the estimated state trajectories produced by the NPF and EKF. As a figure of merit, the rms errors between the true and estimated states is 0.275 for NPF and 0.235 for the EKF. Figure 2 shows that both the NPF and the EKF give very good state estimates.

### B. Unobservable System

Consider the system

$$\mathrm{d}x_t = \sin(x_t)\, \mathrm{d}t + \mathrm{d}\beta_t \tag{23}$$

$$y_k = \left| x_{t_k} \right| + \upsilon_k \tag{24}$$

which is nonlinear unobservable at the origin. In the absence of noise, system (23) has stable equilibria at the points $\{\pi \pm 2\pi\ell, \ell = 0, 1, \ldots\}$ and unstable equilibria at the points $\{\pm 2\pi\ell, \ell = 0, 1, \ldots\}$. In the presence of noise the state $x_t$ will tend to the closest stable equilibrium. The interesting thing is that the output cannot be used to distinguish between positive and negative values of the state variable. This means that there is very little information available to distinguish between positive equilibria and the corresponding negative ones. Because we cannot distinguish between positive and negative values of $x$, we expect that the conditional density function for the state will be multimodal with major lobes centered at approximately $\pi \pm 2\pi\ell$, $\ell = 0, 1, \ldots$. It is interesting to note that the probability mass that is initially located to the right of the origin will migrate to an equilibrium point with a positive coordinate while the mass that starts out to the left of the origin will tend to migrate to a negative equilibrium. The conditional mean at each time instance will depend largely on the prior distribution that is "seldom precisely known and generally rather arbitrarily specified" (Ref. 3, p. 244). The EKF for this problem is given by

$$\frac{\mathrm{d}\hat{x}_t}{\mathrm{d}t} = \sin(\hat{x}_t), \qquad \dot{P}_t = 2P_t\cos(\hat{x}_t) + Q$$

$$K_k = \begin{cases} \dfrac{P_{t_k}^-}{P_{t_k}^- + R} & \hat{x}_{t_k}^- \ge 0 \\ \dfrac{-P_{t_k}^-}{P_{t_k}^- + R} & \hat{x}_{t_k}^- < 0 \end{cases}$$

$$P_{t_k} = \begin{cases} (1-K_k)P_{t_k}^- & \hat{x}_{t_k}^- \ge 0 \\ (1+K_k)P_{t_k}^- & \hat{x}_{t_k}^- < 0 \end{cases}$$

$$\hat{x}_{t_k} = \hat{x}_{t_k}^- + K_k\left(y_k - \left|\hat{x}_{t_k}^-\right|\right)$$

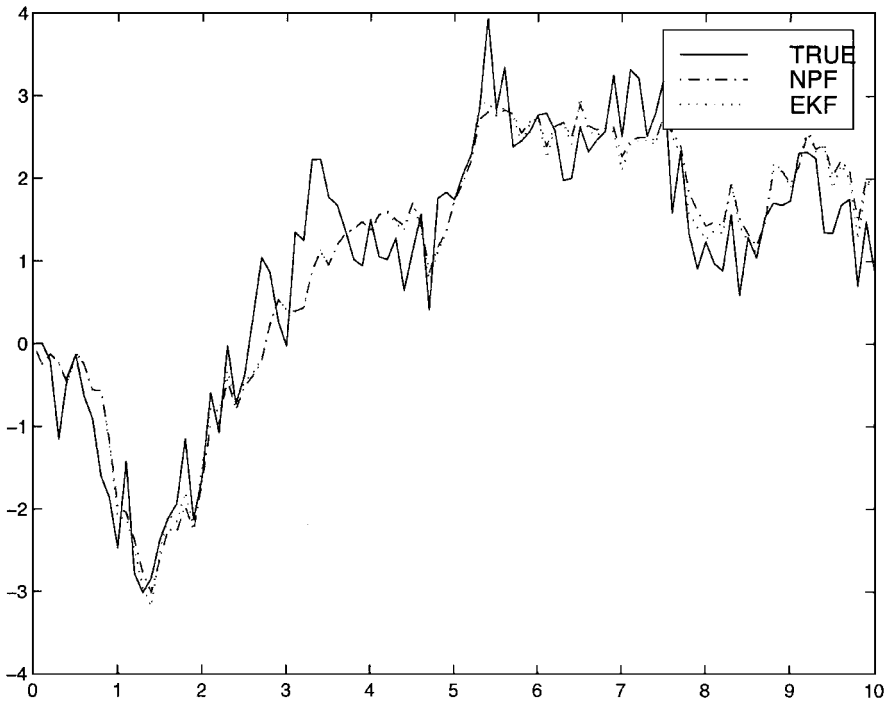


**Fig. 2 Comparison of state trajectory estimates of an EKF and the NPF. Points are plotted at observation times.**

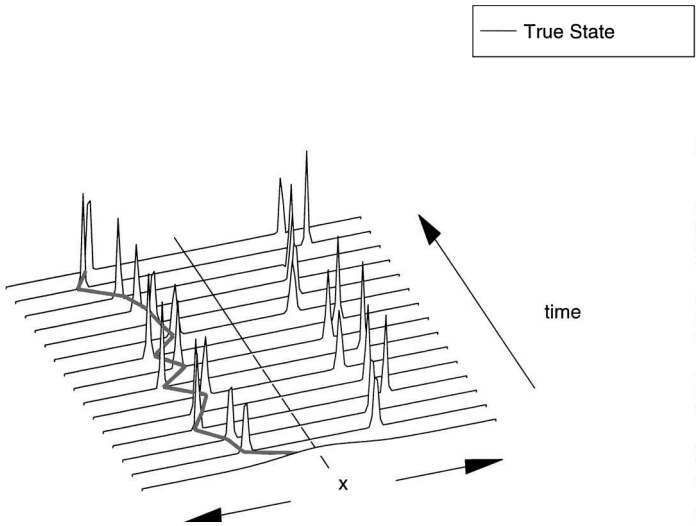


**Fig. 3 Approximate density function and the true state at the measurement updates.**

From the prediction equation we can see that the EKF has multiple equilibria as well. Which equilibria the filter predicts as the state will depend on the initial condition $\hat{x}(0)$, and it will not approximate the correct conditional mean unless the prior distribution is accurate and all of the prior probability density mass is located on one side of the origin. If $\hat{x}(0) < 0$, then $\hat{x}(t)$ goes to an equilibrium to the left and the covariance converges to an inaccurately low value unless the preceding condition is satisfied. An analogous situation holds for $\hat{x}(0) > 0$. If the initial state of the system is zero, then the noise during the initial transients of the system will determine whether the system goes to the left or to the right. It is therefore impossible to predict a priori the sign with which we should initialize the EKF.

The output of the NPF is shown in Fig. 3. The conditional density functions are plotted at discrete times. The thick line shows the true state of the system. After a few measurements the predicted density function is bimodal with modes approximately centered at plus and minus the absolute value of the actual state, which is what we expect.

### C. Second-Order System

In this section we consider the NPF for a system with a two-dimensional state space and a bounded region of attraction. The system dynamics are given by

$$d\boldsymbol{x} = \begin{bmatrix} -x_2 \\ 0.2\left(x_1^2 - 1\right)x_2 + x_1 \end{bmatrix} + d\beta \tag{25}$$

The covariance of $d\beta$ used for this example is $0.5\boldsymbol{I}$ for all time. This system has an unstable limit cycle that bounds the region of attraction to the equilibrium point at the origin. In the absence of noise, state trajectories with initial points inside of the region of convergence spiral inward to the origin, while those starting outside of this region diverge.

The measurement function for this system is the scalar function shown in Fig. 4a as a contour plot. This form of measurement function corresponds to a sensor that gives no directional information. The inner contour bounds a region where in the sensor is saturated and the signal is nearly one whenever the system state enters this region. Beyond the outer contour the signal is nearly zero. The region between the two contours corresponds to an approximately linear region. The covariance for the measurement noise used in this example is $R_k = 0.17$ for all $k$.

The simulated results for this example are shown in Fig. 4b. The true state trajectory and the trajectories of the conditional means produced by the NPF and the EKF are plotted. The prior density for both filters is Gaussian with covariance $0.1\boldsymbol{I}$. After a few measurement updates we see that the EKF has made a fatal error. The EKF measurement function at each measurement is linearized about the current mean estimate, but the actual state is outside of the region where the linearization is valid. The result is that the EKF estimate jumps outside of the region of attraction. The linearization point is now in a region where the measurement function has nearly zero gradient, so that the linearized system is almost completely unobservable. The EKF state estimate now evolves open loop, and because it is outside of the region of attraction, it ultimately diverges even though the true state remains inside the region of attraction. As seen in Fig. 4b the NPF performs reasonably well given the limited information available from the measurements and does not fail because of the nonlinearities in the system and measurement characteristics.

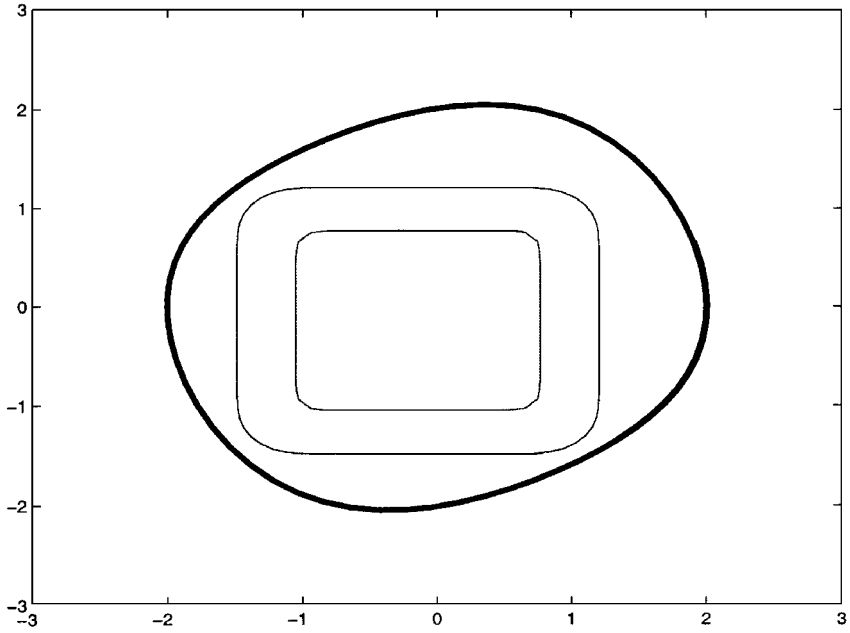


**Fig. 4a Bound for the region of attraction for the dynamic system (25) and the contour plot of the measurement function.**

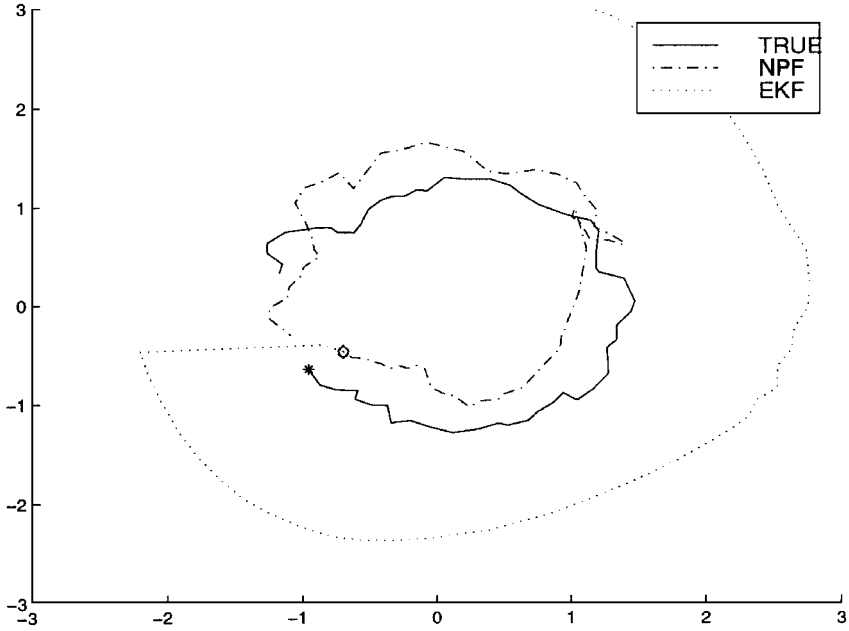


**Fig. 4b Simulated state and estimated state trajectories.**

## VII. Conclusions

In this paper we have used the Galerkin spectral method to derive a nonlinear filter that approximates the exact evolution of the conditional density function of a nonlinear system. We have shown that our filter converges to the true density function as the order of the approximation is increased. We have also shown how to use the DCT algorithm efficiently to implement the algorithm and demonstrated several simple examples where our filter outperforms the EKF. The results contribute to the state of the art in nonlinear filtering by deriving an approximate filter that returns the entire density function and is not based on the linearization of the system. Various filters other than the minimum variance filter could be derived from our approximation because we approximate the conditional density on the state and not its moments. The major limitation of the NPF filter is the curse of dimensionality in that the projection equations must be implemented over an $n$-dimensional state space. The computational complexity and the memory requirements grow exponentially with $n$. This currently limits the practical applicability of this filter to low-order systems. It is envisioned that the filter could be used to process data from sensors with severely nonlinear output characteristics.

## Acknowledgment

This paper was presented at the 1997 American Control Conference.

## References

[1] Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," *Transactions of the ASME, Journal of Basic Engineering*, Vol. 82, No. 1, 1960, pp. 34, 35.

[2] Kalman R. E., and Bucy, R. S., "New Results in Linear Filtering and Prediction Theory," *Transaction of the ASME, Journal of Basic Engineering*, Vol. 83, 1961, 95–108.

[3] Jazwinski, A. H., *Stochastic Processes and Filtering Theory*, Vol. 64, *Mathematics in Science and Engineering*, Academic Press, New York, 1970.

[4] Kushner, H. J., "Dynamical Equations for Optimal Nonlinear Filter," *Journal of Differential Equations*, Vol. 3, No. 2, 1967, pp. 179–190.

[5]Kushner, H. J., "Nonlinear Filtering: The Exact Dynamical Equations Satisfied by the Conditional Mode," *IEEE Transactions on Automatic Control*, Vol. 12, 1967, pp. 262–267.

[6]Stratonovich, R. L., "Conditional Markov Processes," *Theory of Probability and Its Applications*, Vol. 5, 1960, pp. 156–178.

[7]Bryson, A. E., and Ho, Y. C., *Applied Optimal Control*, Hemisphere, New York, 1975.

[8]Lewis, F. L., *Optimal Estimation: With an Introduction to Stochastic Control Theory*, Wiley, New York, 1986.

[9]Ohap, R. F., and Stubberud, A. R., "A Technique for Estimating the State of a Nonlinear System," *IEEE Transactions on Automatic Control*, Vol. 10, No. 4, 1965, pp. 150–155.

[10]Gelb, A. (ed.), *Applied Optimal Estimation*, MIT Press, Cambridge, MA, 1974.

[11]Ljung, L., "Asymptotic Behavior of the Extended Kalman Filter Using Filtered State Estimates," *IEEE Transactions on Automatic Control*, Vol. 24, No. 1, 1979, pp. 36–50.

[12]Ursin, B., "Asymptotic Convergence Properties of the Extended Kalman Filter Using Filtered State Estimates," *IEEE Transactions on Automatic Control*, Vol. 25, No. 12, 1980, pp. 1207–1211.

[13]Galkowski, P. J., and Islam, M. A., "An Alternative Derivation of the Modified Gain Function of Song and Speyer," *IEEE Transactions on Automatic Control*, Vol. 36, No. 11, 1991, pp. 1323–1326.

[14]Bell, B. M., and Cathey, F. W., "The Iterated Kalman Filter Update as a Gauss-Newton Method," *IEEE Transactions on Automatic Control*, Vol. 38, No. 2, 1993, pp. 294–297.

[15]Bell, B. M., "The Iterated Kalman Smoother as a Gauss-Newton Method," *SIAM Journal of Optimization*, Vol. 4, No. 8, 1994, pp. 626–636.

[16]Mikhlin, S. G., *Variational Methods in Mathematical Physics*, MacMillan, New York, 1964.

[17]Kantorovich, L. V., and Krylov, V. I., *Approximate Methods of Higher Analysis*, Interscience Publishers, New York, 1958.

[18]Risken, H., *The Folker-Planck Equation: Methods of Solution and Application*, Vol. 18, *Springer Series in Synergetics*, Springer–Verlag, Berlin, 1989.

[19]Ling, F. H., and Wu, X. X., "Fast Galerkin Method and Its Application to Determine Periodic Solutions of Non-Linear Oscillators," *International Journal of Non-Linear Mechanics*, Vol. 22, No. 2, 1987, pp. 89–98.

[20]Bouc, R., and Defilippi, M., "A Galerkin Multiharmonic Procedure for Nonlinear Multidimentional Random Vibration," *International Journal of Engineering Science*, Vol. 25, No. 6, 1987, pp. 723–733.

[21]Weber, H., "Nonlinear Solution of a Class of Nonlinear Boundary Value Problems for Analytic Functions," *Journal of Applied Mathematics and Physics*, Vol. 33, May 1982, pp. 301–314.

[22]Beard, R., "Improving the Closed-Loop Performance of Nonlinear Systems," Ph.D. Thesis, Dept. of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Inst., Troy, NY, 1995.

[23]Beard, R., Saridis, G., and Wen, J., "Improving the Performance of Stabilizing Control for Nonlinear Systems," *Control Systems Magazine*, Vol. 16, No. 10, 1996, pp. 27–35.

[24]Beard, R., Saridis, G., and Wen, J., "Galerkin Approximation of the Generalized Hamilton-Jacobi-Bellman Equation," *Automatica*, Vol. 33, No. 12, 1997, pp. 2159–2177.

[25]Beard, R., Saridis, G., and Wen, J., "Approximate Solution to the Time-Invariant Hamilton-Jacobi-Bellman Equation," *Journal of Optimization Theory and Application*, Vol. 96, March 1998, pp. 589–626.

[26]Baras, J. S., "Real Time Architectures for the Zakai Equation and Application," *Stochastic Analysis: Liber Amicorum for Moshe Zakai*, edited by E. Mayer-Wolf, E. Merzbach, and A. Shwartz, Academic Press, Boston, MA, 1991, pp. 15–38.

[27]Zakai, M., "On the Optimal Filtering of Diffusion Processes," *Zeitschrift für Wahrscheinlichkeitstheorie und verw. Geb.*, Vol. 2, 1969, pp. 230–243.

[28]Daum, F. E., "Solution of the Zakai, Equation by Separation of Variables," *IEEE Transactions on Automatic Control*, Vol. 32, No. 10, 1987, pp. 941–943.

[29]Daum, F. E., "Exact Finite-Dimensional Nonlinear Filters," *IEEE Transactions on Automatic Control*, Vol. 31, No. 7, 1986, pp. 616–622.

[30]Brigo, D., Hanzon, B., and LeGlanc, F., "A Differential Geometric Approach to Nonlinear Filtering: The Projection Filter," *IEEE Transactions on Automatic Control*, Vol. 43, No. 2, 1998, pp. 247–252.

[31]Zeidler, E., *Nonlinear Functional Analysis and Its Applications, II/A: Linear Monotone Operators*, Springer–Verlag, Berlin, 1990.

[32]H. K., Khalil, *Nonlinear Systems*, Macmillan, New York, 1992.

[33]Jain, A. K., *Fundamentals of Digital Image Processing*, Prentice–Hall, Englewood Cliffs, NJ, 1989.