Optimal Parameterization of Posterior Densities Using Homotopy

Jonas Hagmar, Mats Jirstrand
Department of Systems Biology
and Bioimaging
Fraunhofer-Chalmers Centre
Gothenburg, Sweden
Email: {jonas,matsj}@fcc.chalmers.se

Lennart Svensson
Department of Signals and Systems
Chalmers University of Technology
Gothenburg, Sweden
Email: lennart.svensson@chalmers.se

Mark Morelande
Department of Electrical
and Electronic Engineering
The University of Melbourne
Melbourne, Australia
Email: mrmore@unimelb.edu.au

***Abstract*—In filtering algorithms, it is often desirable that the prior and posterior densities share a common density parameterization. This can rarely be done exactly. Instead it is necessary to seek a density from the same family as the prior which closely approximates the true posterior. We extend a method for computing the optimal parameter values for representing the posterior within a given parameterization. This is achieved by minimizing the deviation between the parameterized density and a homotopy that deforms the prior density into the posterior density. We derive novel results both for the general case, and for specific choices of measures of deviation. This includes approximate solution methods, that prove useful when we demonstrate how the method can be used with common density parameterizations. For an example with a non-linear measurement model, the method is shown to be more accurate than the Extended, Unscented and Cubature Kalman filters.**

Keywords: Nonlinear filtering, homotopy, measurement update, optimization, ordinary differential equations.

I. Introduction

The problem of computing the best estimate of the state of a dynamic system from a set of noisy measurements arises in many fields, such as navigation, radar tracking, computer vision and biology. The filtering problem can be stated as the task of finding the best estimate of the system state using all past measurements.

A common assumption is that the system state evolves according to a Markov process, and that noisy measurements of the system state are taken at discrete time points. Under these conditions, filtering operates in two phases; prediction and measurement update. Between measurements, the state of the system at future time points is predicted using the probability density of the system state at the last measurement, and the system equations, giving the so called *prior probability density* at the next measurement. The *posterior probability density*, which is the probability density of the state conditioned on all past measurements, including the new measurement, is computed from the prior probability density and the likelihood of the measurement, using Bayes' theorem. Using the posterior density as input for the next prediction phase, the two phases are repeated.

For linear systems and Gaussian random variables, the optimal solution to the filtering problem is given by the Kalman filter [6]. Exact solutions to the filtering problem exist for some other classes of systems [2], but in most cases, obtaining an explicit solution to the filtering problem is not possible. Performing the prediction step requires solving the Fokker-Planck equation (also known as the Kolmogorov forward equation), which is an intractable problem for all but a very limited class of systems. There are numerous techniques for approximating the solution of the filtering problem for non-linear systems, such as the Extended Kalman filter, the Unscented Kalman filter, and particle filters [4], [7]. The solution of the prediction step in the general case can be approximated by finite element methods applied to the Fokker-Planck equation, although the computational load becomes prohibitive even for systems of low dimension [10].

For all the methods mentioned, both exact and approximate, the probability density of the system state, obtained from the prediction step, is parameterized by a finite number of parameters. While the posterior density can be evaluated pointwise, an approximation of the posterior density within the same parameterization class as the prior density is often desirable, since it is used as an input to the next round of prediction. Finding the best parameter values directly is a problem of global optimization, which is very demanding.

Hanebeck et al. [5] introduced a general framework for performing the measurement update. In this method, a homotopy is formed, which gradually deforms an initial density into the posterior density as a scaling parameter λ increases from 0 to 1. A parameterized density is chosen, and λ-dependent parameter values are sought that minimize the deviation between this density and the homotopy, according to a given measure of deviation. The approximation of the posterior density is then retrieved as the parameterized density at $\lambda = 1$. In [5], the framework was used for examples of Gaussian mixtures. A serious limitation of the framework is that it assumes that integrals of functions of the homotopy and the parameterized density can be computed over the entire domain. This method shares some ideas with an approach that has been developed in a series of papers by Daum and Huang (see e.g. [3]). Even though Daum and Huang do not explicitly specify any density parameterization, they use a homotopy to propagate particles which characterize the posterior distribution.

978-0-9824438-2-8/11/$26.00 ©2011 IEEE

In many ways, this paper blends the work by Hanebeck et al. [5] with the ideas of Daum and Huang [3]. The similarities with Daum and Huang are the choice of homotopy, and that we initialize the homotopy with the prior density. Like Hanebeck et al. we use a cost based approach with a general but explicit density parameterization. The framework of Hanebeck et al. is extended by formalizing the minimization criterion for arbitrary measures of deviation and deriving the resulting system of ordinary differential equations (ODEs) to be solved for obtaining the parameter values as a function of λ. We further discuss how approximations can be introduced to facilitate the computations for non-Gaussian, non-linear and/or high-dimensional problems. In addition, we demonstrate how the method can be used with several different, commonly used, measures of deviation, density parameterizations, and measurement models. In these examples, we use the approximations that we have developed in order to make the computations tractable when the posterior density cannot be computed analytically. The method makes no assumption about the type of parameterization, other than appropriate smoothness of the measure of deviation with respect to the parameters and λ. We hope that the flexibility of the method will allow the use of new prediction algorithms, making it possible to address the large class of problems that are inaccessible with today's methods.

II. Background

In this section, we will present the theoretical background needed for the presentation of our method for computing the optimal parameterization of the posterior density. This includes presentations of the framework that we elaborate on and the specific choices of density parameterizations, measures of deviation, and measurement models used in our simulations.

A. The Filtering Problem and its Solution in the General Case

Given a time-dependent stochastic vector x_t, representing the system state, and noisy measurements $y_1, y_2, \ldots$, of the system, taken at corresponding time points $t_1 < t_2 < \ldots$, the filtering problem can be stated as finding the best estimate of x_t given the set of measurements $Y_t = \{y_l : t_l \leq t\}$.

Assuming that the system state evolves according to a Markov process, the solution to the filtering problem consists of repeated cycles of two distinct phases. The first of these is the prediction phase, operating between measurements, in which the predictive density $p(x_t|Y_{t_{k-1}})$, $t_{k-1} < t < t_k$, $k = 1, 2, \ldots$, is computed from $p(x_{t_{k-1}}|Y_{t_{k-1}})$ and the system equations. We will denote $p(x_{t_k}|Y_{t_{k-1}})$ the *prior probability density*.

The purpose of the second phase, called the measurement update, is to compute the *posterior probability density* $p(x_{t_k}|Y_{t_k})$. The prior and posterior probability densities are related through Bayes' theorem, i.e.,

$$p(x_k|Y_{t_k}) = \frac{p(y_k|x_k)p(x_k|Y_{t_{k-1}})}{p(y_k|Y_{t_{k-1}})}, \tag{1}$$

where $p(y_k|x_k)$ is called the likelihood of the measurement. As the denominator of the right hand side of (1) acts as a normalization factor, the posterior probability density can be computed without explicit knowledge of $p(y_k|Y_{t_{k-1}})$, through normalization of the numerator.

B. Assumptions and Notation

For ease of notation, we consider the kth measurement update, and denote the prior density by $p(x)$ and the likelihood by $l(x)$. Since the filtering problem can be solved exactly only for a very limited class of systems, solutions to the problem have to be approximated in the vast majority of cases. We will by $p(x)$ mean the prior probability density obtained from the chosen solution method, whether it be exact or approximate. When necessary, we will assume that Leibniz rule for differentiation under the integral sign is applicable.

C. Progressive Bayes

The method used in this paper builds on the work of Hanebeck et al. [5], who establish the following framework for computing the optimal parameters for representing the posterior density within a given parameterization class:

1) Introduce a homotopy, $f(x;\lambda)$, where $f(x;0)$ is some density that is simple to approximate. As λ increases, $f(x;\lambda)$ continuously approaches $f(x;1)$, that equals the posterior density.
2) Define an approximation density, $f_a(x;\theta)$, where θ is a parameter vector of dimension r.
3) Define a measure of deviation, $G(\theta,\lambda)$, between $f(x;\lambda)$ and $f_a(x;\theta)$.
4) Derive a system of ODEs of the form

$$b(\theta,\lambda) = A(\theta)\theta'(\lambda), \tag{2}$$

where $\theta(0)$ is given. The optimal parameter values are then $\theta(1)$.

In the paper by Hanebeck et al., it is not specified more explicitly how to derive the system of ODEs for a general measure of deviation, but instead the authors show how the framework is implemented for a specific choice of the involved components.

D. Probability Density Parameterizations

A common family of parameterizations of probability densities follows from assuming that the probability density of x_t is a mixture of Gaussians, i.e.,

$$f_a(x;\theta) = \sum_{i=1}^{n} w_i N(x;\mu_i,\Sigma_i), \tag{3}$$

where $N(x;\mu,\Sigma)$ denotes the probability density function of a multivariate normal distribution with mean μ and covariance matrix Σ, $0 \leq w_i \leq 1$ are the weights, summing to 1, μ_i are the means, and Σ_i, $i = 1, \ldots, n$, are the covariance matrices of the Gaussians forming the sum. The corresponding parameter vector is

$$\theta = [w_1, \ldots, w_n, \mu_1, \ldots, \mu_n, \mathrm{vec}(\Sigma_1), \ldots, \mathrm{vec}(\Sigma_n)]^T. \tag{4}$$

A very flexible density parameterization consists of values of the density at specific points, and a method to interpolate between these points. In other words, we have the parameters

$$\theta = [x_1, \dots, x_n, v_1, \dots, v_n]^T, \tag{5}$$

where v_i is the value of the probability density at the point x_i, $i = 1, \dots, n$, where n is the number of points. The interpolation method determines the structure of $f_a(x;\theta)$.

E. Measures of Deviation

For our examples, we use two well-known measures of statistical distance. The first is the Kullback-Leibler divergence [9]

$$D_{\text{KL}}(Q\|R) = \int q(x) \log \frac{q(x)}{r(x)}\, dx, \tag{6}$$

where Q and R are continuous random variables with probability densities $q(x)$ and $r(x)$, respectively. Although the Kullback-Leibler divergence is not a true distance metric, since it is not symmetric, it will prove useful for our examples. The second choice of $G(\theta,\lambda)$ is the squared Hellinger distance [8],

$$H^2(Q,R) = \frac{1}{2} \int \left(\sqrt{q(x)} - \sqrt{r(x)}\right)^2 dx, \tag{7}$$

with Q, R, $q(x)$, and $r(x)$ defined similarly as in the Kullback-Leibler case. With slight abuse of notation, we will use $D_{\text{KL}}(q(x)\|r(x)) \equiv D_{\text{KL}}(Q\|R)$ and $H^2(q(x),r(x)) \equiv H^2(Q,R)$. We will see that these measures of deviation yield satisfactory results for our method, although the functions corresponding to $q(x)$ and $r(x)$ are not necessarily probability densities. In the example given by Hanebeck et al. [5], the measure of deviation is

$$\begin{aligned} \tilde{D}_{\text{SI}}(\theta,\lambda) &= \frac{1}{2}\int (f(x;\lambda) - f_a(x;\hat{\theta}) - \phi_a^T(\theta - \hat{\theta}))^2\, dx, \\ \phi_a &\equiv \left.\frac{\partial f_a(x;\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}}, \end{aligned} \tag{8}$$

which is a linearization of the squared integral deviation

$$D_{\text{SI}}(\theta,\lambda) = \frac{1}{2}\int (f(x;\lambda) - f_a(x;\theta))^2\, dx, \tag{9}$$

around the nominal parameter $\hat{\theta}$.

F. Measurement Models

For a measurement model

$$y_k = h(x_{t_k}) + v_k, \tag{10}$$

where v_k is normally distributed with mean 0 and covariance matrix R_k, the likelihood is

$$l(x) = N(y_k; h(x), R_k). \tag{11}$$

The measurement models considered in our simulations below are of two different types. The first type is the linear model $h(x) = M_k x$, since this gives a Gaussian likelihood, which is a common assumption. The second type of measurement model we consider is typical of radar systems, where measurements are taken in polar coordinates, while the system state is represented in Cartesian coordinates. This corresponds to

$$h(x) = [\sqrt{x_1^2 + x_2^2}, \arg(x_1 + i x_2)]^T, \tag{12}$$

where $x = [x_1, x_2]^T$.

III. Optimal Posterior Density Representation using Homotopy

Given the definitions in Section II, several methods for finding the optimal parameter values to represent the posterior density within a given parameterization are possible. A naive approach is to minimize $G(\theta, 1)$ w.r.t. θ directly. Performing optimization in this case is computationally a costly approach and/or does not guarantee that the global minimum is found. Another possible method is to iteratively compute the optimal parameters $\theta(\lambda)$ for small increments of λ by successive minimizations of $G(\theta,\lambda)$. This would hopefully make the optimizations involved simpler to perform, since only small adjustments of θ are necessary for every increment. The problems of choosing a suitable optimization method remains, however. The presented method circumvents this problem, by tracking the minimum of the measure of deviation for increasing values of λ. The necessary condition for optimality of the parameters is used to derive a system of ODEs governing $\theta(\lambda)$. We thus transform the optimization problem into solving a system of ODEs, for which efficient numerical methods exist.

We assume the conditions outlined above, i.e., a prior probability density $p(x)$ that belongs to a class of parameterized functions $f_a(x;\theta)$, where θ is a parameter vector. The parameter values θ_0 are such that $p(x) = f_a(x;\theta_0)$. The likelihood of the measurement is $l(x)$. We seek the parameter values θ_p, giving the optimal approximation of the unnormalized posterior probability density, within the class of functions defined by $f_a(x;\theta)$. To this end, we build on the general framework in Section II-C. The general idea is to calculate the optimal parameter values $\theta(\lambda)$ that minimize a measure of deviation, $G(\theta,\lambda)$, between $f_a(x;\theta)$ and a homotopy $f(x;\lambda)$, that continuously deforms $p(x)$ into the unnormalized posterior $p(x)l(x)$ as λ increases from 0 to 1. While different homotopies that fulfill these criteria are possible, we have chosen to use

$$f(x;\lambda) = p(x) l^{\lambda}(x), \tag{13}$$

when a specification of $f(x;\lambda)$ is needed.

A modification of the method presented in [5] is that we choose $f(x;0)$ equal to the prior density. As discussed above, in a typical filtering framework, we obtain an approximation of the prior within a given parameterization class from the preceding prediction step. Using the same parameterization for $f_a(x;\theta)$, we obtain a suitable representation of the posterior for the next round of prediction. These choices of $f(x;0)$, $f_a(x;\theta)$ and θ_0 also minimize $G(\theta,0)$, which is essential.

We extend the work in [5] by deriving the differential equation governing the evolution of $\theta(\lambda)$ for arbitrary measures of deviation $G(\theta,\lambda)$. This differential equation does not, strictly speaking, fall within the class of ODEs (2), considered by Hanebeck et al. We also develop approximate solution methods, that prove useful when analytic solutions of the right hand side of the ODE are not available.

More formally, we seek

$$\theta(\lambda) = \operatorname*{argmin}_{\theta} G(\theta, \lambda). \tag{14}$$

Using the notation

$$G_{\theta^i \lambda^j}(\hat{\theta}, \hat{\lambda}) \equiv \left. \frac{\partial^{i+j} G(\theta, \lambda)}{\partial \theta^i \partial \lambda^j} \right|_{\substack{\theta=\hat{\theta} \\ \lambda=\hat{\lambda}}}, \tag{15}$$

a necessary condition for optimality is

$$G_{\theta}(\theta(\lambda), \lambda) = 0. \tag{16}$$

Since this condition is true for all $\lambda \in [0, 1]$, we obtain, by taking the total derivative w.r.t. λ of the preceding equation,

$$G_{\theta\theta}(\theta(\lambda), \lambda)\theta'(\lambda) + G_{\theta\lambda}(\theta(\lambda), \lambda) = 0, \tag{17}$$

by applying the chain rule. Solving for $\theta'(\lambda)$, we arrive at, what we denote, the *homotopy differential equation* (HDE)

$$\theta'(\lambda) = -G_{\theta\theta}^{-1}(\theta(\lambda), \lambda) G_{\theta\lambda}(\theta(\lambda), \lambda), \tag{18}$$

where we, for the reasons mentioned above, have the initial condition $\theta(0) = \theta_0$, and solve it for $\lambda \in [0, 1]$.

A. Approximation of the HDE Using Restart

The solution of the HDE can be approximated by linearization on λ-intervals corresponding to a partition of $[0, 1]$, yielding an equation that we denote the *Homotopy Difference Equation* (HΔE). It is possible to simplify the HDE at $\lambda = 0$, by observing that $f(x; 0) = f_a(x; \theta(0)) = p(x)$. If the H$\Delta$E is restarted at λ-points corresponding to the left points of the λ-intervals of the partition, taking $f_a(x; \theta(\lambda))$ at each λ-point as the prior probability density for the restarted HΔE, we obtain what we denote the *Approximate Homotopy Difference Equation* (AHΔE), where the simplifications possible for the HDE at $\lambda = 0$ can be applied on each λ-interval. Using the same approximations in the continuous case, we obtain what we call the *Approximate Homotopy Differential Equation* (AHDE). It is worth noting that when $f(x; \lambda)$ can be perfectly represented by $f_a(x; \theta)$, the AHDE is equivalent to the HDE, i.e., it does not introduce any additional approximations.

To be more explicit, we have, by Taylor's formula, for a partition $0 = \lambda_0 < \lambda_1 < \ldots < \lambda_n = 1$,

$$\theta(\lambda_{i+1}) = \theta(\lambda_i) + \theta'(\lambda_i)\Delta\lambda_i + o(\Delta\lambda_i), \tag{19}$$

$i = 0, \ldots, n-1$, where $\Delta\lambda_i = \lambda_{i+1} - \lambda_i$. Truncating the Taylor expansion after the linear term, we obtain the HΔE as

$$\bar{\theta}(\lambda_{i+1}) = \bar{\theta}(\lambda_i) + \bar{\theta}'(\lambda_i)\Delta\lambda_i, \tag{20}$$

$$\bar{\theta}(0) = \theta(0) \tag{21}$$

where $\bar{\theta}(\lambda_{i+1})$ is the approximation of $\theta(\lambda_{i+1})$ obtained from the HΔE, and $\bar{\theta}'(\lambda) = -G_{\theta\theta}^{-1}(\bar{\theta}(\lambda), \lambda) G_{\theta\lambda}(\bar{\theta}(\lambda), \lambda)$.

For the computation of the AHΔE, the HΔE is restarted at every λ_i, in the sense that $f_a(x; \bar{\theta}(\lambda_i))$ is used as the prior. This is reasonable, since $f_a(x; \theta(\lambda))$ is the best possible approximation of $f(x; \lambda)$, by design. As mentioned above, it is then possible to make simplifications to the AHΔE

$$\tilde{\bar{\theta}}(\lambda_{i+1}) = \tilde{\bar{\theta}}(\lambda_i) + \tilde{\bar{\theta}}'(\lambda)\Delta\lambda_i, \tag{22}$$

$$\tilde{\bar{\theta}}(0) = \theta(0), \tag{23}$$

where $\tilde{\bar{\theta}}(\lambda_{i+1})$ is the AHΔE approximation of $\theta(\lambda_{i+1})$, and $\tilde{\bar{\theta}}'(\lambda) = -G_{\theta\theta}^{-1}(\tilde{\bar{\theta}}(\lambda), \lambda) G_{\theta\lambda}(\tilde{\bar{\theta}}(\lambda), \lambda)$, with the substitutions $\lambda = 0$ and $f(x; \lambda) := p(x) := f_a(x; \theta)$.

To derive the AHDE, we let $\Delta\lambda_i \to 0$, and obtain, in the limit,

$$\tilde{\theta}'(\lambda) = -G_{\theta\theta}^{-1}(\tilde{\theta}(\lambda), \lambda) G_{\theta\lambda}(\tilde{\theta}(\lambda), \lambda), \tag{24}$$

$$\tilde{\theta}(0) = \theta(0), \tag{25}$$

where $\tilde{\theta}(\lambda)$ is the AHDE approximation of $\theta(\lambda)$, and the right hand side of the differential equation is computed using the substitutions $\lambda = 0$ and $f(x; \lambda) := p(x) := f_a(x; \theta)$.

1) AHDE for Specific Measures of Deviation: We will study how the AHDE can be simplified for specific choices of the measure of deviation, namely the Kullback-Leibler divergence and the squared Hellinger distance. We will use $f \equiv f(x; \lambda)$ and $f_a \equiv f_a(x; \theta)$ for ease of notation. Using the Kullback-Leibler divergence as our choice of measure of deviation, we have

$$\left. \frac{\partial^2 D_{\mathrm{KL}}(f \| f_a)}{\partial \theta^2} \right|_{f=f_a} = \tilde{\mathcal{I}}(\theta) - \int \frac{\partial^2 f_a}{\partial \theta^2} \, dx \tag{26}$$

$$\left. \frac{\partial^2 D_{\mathrm{KL}}(f_a \| f)}{\partial \theta^2} \right|_{f=f_a} = \tilde{\mathcal{I}}(\theta) + \int \frac{\partial^2 f_a}{\partial \theta^2} \, dx, \tag{27}$$

$$\left. \frac{\partial^2 D_{KL}(f \| f_a)}{\partial \lambda \partial \theta} \right|_{\substack{\lambda=0 \\ f=f_a}} = \left. \frac{\partial^2 D_{KL}(f_a \| f)}{\partial \lambda \partial \theta} \right|_{\substack{\lambda=0 \\ f=f_a}} \tag{28}$$

$$= -\int \log(l) \frac{\partial f_a}{\partial \theta} \, dx, \tag{29}$$

where

$$\tilde{\mathcal{I}}(\theta) = \int \frac{\partial \log f_a}{\partial \theta} \frac{\partial \log f_a}{\partial \theta}^T f_a \, dx \tag{30}$$

is the Fisher information matrix, $\mathcal{I}(\theta)$, if $f_a(x; \theta)$ is a probability density. If this holds, the second terms of the right hand sides of (26) and (27) disappear, making them equal to the Fisher information matrix.

If the measure of deviation is instead taken to be the squared Hellinger distance, we obtain the corresponding expressions

$$\left. \frac{\partial^2 H^2(f, f_a)}{\partial \theta^2} \right|_{f=f_a} = \frac{1}{4} \tilde{\mathcal{I}}(\theta), \tag{31}$$

$$\left. \frac{\partial^2 H^2(f, f_a)}{\partial \theta \partial \lambda} \right|_{\substack{\lambda=0 \\ f=f_a}} = -\frac{1}{4} \int \log(l) \frac{\partial f_a}{\partial \theta} \, dx. \tag{32}$$

We see that for both the Kullback-Leibler divergence and the squared Hellinger distance, the AHDE is

$$\frac{d\bar{\theta}(\lambda)}{d\lambda} = \mathcal{I}^{-1}(\theta) \int \log(l) \frac{\partial f_a}{\partial \theta}\, dx, \tag{33}$$

$$\bar{\theta}(0) = \theta(0), \tag{34}$$

when $f_a(x;\theta)$ is a probability density.

2) Connection Between Linearized Squared Integral Deviation HDE and AHDE: Hanebeck et al. use (8) as the measure of deviation, and derive the ODE

$$\int \frac{\partial f}{\partial \lambda} \frac{\partial f_a}{\partial \theta}\, dx = \left(\int \frac{\partial f_a}{\partial \theta} \frac{\partial f_a}{\partial \theta}^T dx \right) \theta'(\lambda), \tag{35}$$

corresponding to the HDE.

For the squared integral deviation, (9), we have

$$\frac{\partial^2 D_{\mathrm{SI}}}{\partial \theta^2} = \int \left(\frac{\partial f_a}{\partial \theta} \frac{\partial f_a}{\partial \theta}^T + (f_a - f) \frac{\partial^2 f_a}{\partial \theta^2} \right) dx, \tag{36}$$

$$\frac{\partial^2 D_{\mathrm{SI}}}{\partial \lambda \partial \theta} = - \int \frac{\partial f}{\partial \lambda} \frac{\partial f_a}{\partial \theta}\, dx. \tag{37}$$

By noting that

$$\left. \frac{\partial^2 D_{\mathrm{SI}}}{\partial \theta^2} \right|_{f = f_a} = \int \frac{\partial f_a}{\partial \theta} \frac{\partial f_a}{\partial \theta}^T dx, \tag{38}$$

we see that the ODE derived in the example in [5] is in fact the AHDE with D_{SI} as the choice of measure of deviation.

B. Approximation of the HDE Using Numerical Integration

The integrals of the HDE can be approximated by numerical methods. Write the chosen measure of deviation as

$$G(\theta, \lambda) = \int_{\Omega} g(x; \theta, \lambda)\, dx, \tag{39}$$

where $g(x;\theta,\lambda)$ is an expression formed by $f(x;\lambda)$ and $f_a(x;\theta)$, and Ω is the union of the supports of $f(x;\lambda)$ and $f_a(x;\theta)$. This is applicable for both the Kullback-Leibler divergence and the squared Hellinger distance, used as measures of deviation in the examples below. We then have

$$\frac{\partial^2 G(\theta, \lambda)}{\partial \theta_i \partial \theta_j} = \int_{\Omega} \frac{\partial^2 g(x; \theta, \lambda)}{\partial \theta_i \partial \theta_j}\, dx. \tag{40}$$

$$\approx \sum_{k=1}^{n} \frac{\partial^2 g(\xi_k; \theta, \lambda)}{\partial \theta_i \partial \theta_j} \Delta x_k, \tag{41}$$

where we have divided Ω into disjoint elements ω_k, $k = 1, \ldots, n$, with respective volumes Δx_k, $\xi_k \in \omega_k$, and $i, j = 1, \ldots, r$. Even for $f(x;\lambda)$ or $f_a(x;\theta)$ with infinite support, it is often possible to introduce such a division of Ω with finite n, by neglecting any element $\tilde{\omega}$, with volume $\Delta\tilde{x}$, such that

$$\frac{\partial^2 g(\tilde{\xi}; \theta, \lambda)}{\partial \theta_i \partial \theta_j} \Delta \tilde{x} \lll \underbrace{\sum_{k=1}^{n} \frac{\partial^2 g(\xi_k; \theta, \lambda)}{\partial \theta_i \partial \theta_j}}_{\equiv A_{ij}(\theta, \lambda)} \Delta x_k, \tag{42}$$

where $\tilde{\xi} \in \tilde{\omega}$. If $\Delta x = \Delta x_k$, we thus have

$$\frac{\partial^2 G(\theta, \lambda)}{\partial \theta^2} \approx \Delta x A(\theta, \lambda). \tag{43}$$

By a similar argument,

$$\frac{\partial^2 G(\theta, \lambda)}{\partial \theta \partial \lambda} \approx \Delta x \underbrace{\sum_{k=1}^{n} \frac{\partial^2 g(\xi_k; \theta, \lambda)}{\partial \theta_i \partial \lambda}}_{b_i(\theta, \lambda)}. \tag{44}$$

Using this approximation with the HDE, we obtain

$$\theta'(\lambda) = -A(\theta, \lambda)^{-1} b(\theta, \lambda) \big|_{\theta = \theta(\lambda)}. \tag{45}$$

This gives some justification to computing the measure of deviation over a grid of discrete points instead of over Ω. In the examples below, we will see that this method gives satisfactory results, even for grids that only partly cover Ω.

IV. Examples

In this section, we demonstrate how the posterior probability density can be computed using our framework, for the examples mentioned in Section II, beginning with an example for which the integrals of the HDE can be solved exactly. In the later examples, we show how even coarse approximations of the HDE yield satisfactory results. Throughout the examples, we use the homotopy (13).

A. Univariate Gaussian Prior and Likelihood

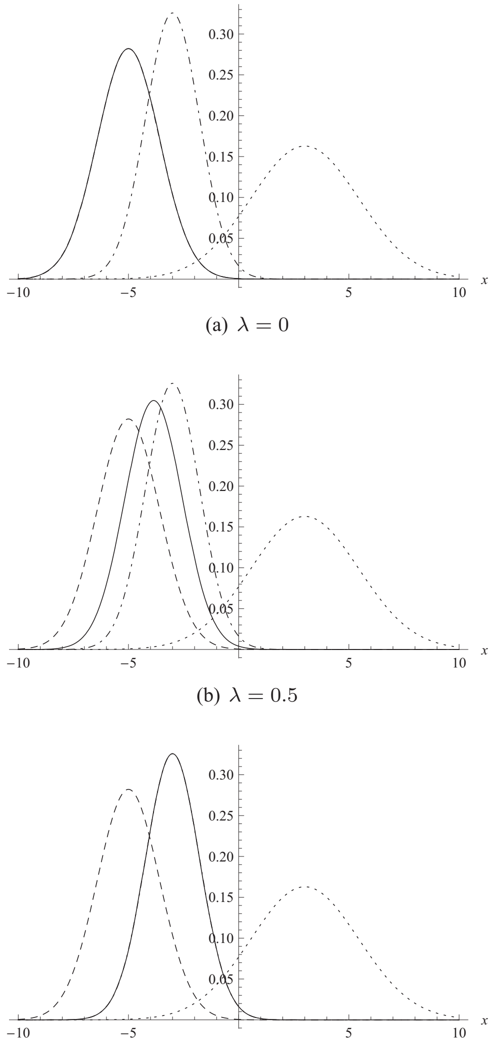
As a first example, we assume a prior given by a univariate normal density, with mean μ_0 and variance σ_0^2, i.e., $p(x) = N(x; \mu_0, \sigma_0^2)$. The likelihood is given by $l(x) = N(\mu_l; x, \sigma_l^2) = N(x; \mu_l, \sigma_l^2)$. We use the Kullback-Leibler divergence, $D_{\mathrm{KL}}(f(x;\lambda) \| f_a(x;\theta))$, as the measure of deviation between $f(x;\lambda)$ and $f_a(x;\theta)$. Furthermore, we let $f_a(x;\theta) = N(x; \mu, \sigma^2)$, and thus $\theta = [\mu, \sigma^2]^T$. We obtain the HDE

$$\frac{d\theta(\lambda)}{d\lambda} = I^{-1} \int \frac{\partial f}{\partial \lambda} \frac{\partial \log f_a}{\partial \theta}\, dx, \tag{46}$$

$$I = \int f \left(\frac{\partial \log f_a}{\partial \theta} \frac{\partial \log f_a}{\partial \theta}^T - \frac{1}{f_a} \frac{\partial^2 f_a}{\partial \theta^2} \right) dx, \tag{47}$$

$$\theta(0) = [\mu_0, \sigma_0^2]^T. \tag{48}$$

For numerical values of the parameters of the prior probability density and the likelihood, the integrals of this HDE can be solved analytically. In this example, we arbitrarily choose $\mu_0 = -5$, $\sigma_0^2 = 2$, $\mu_l = 3$ and $\sigma_l^2 = 6$. The true posterior probability density is normal with mean -3 and variance $3/2$. The HDE can now be solved numerically, yielding $\theta(1) = [-3.0, 1.5]^T$, which is the expected result, since no approximations are introduced, other than those resulting from the error tolerances used by the numerical ODE solver. The solution of the HDE is visualized in Figure 1. The homotopy $f(x;\lambda)$ coincides with $f(x;\theta(\lambda))$ for $0 \le \lambda \le 1$ (not shown).



(a) $\lambda = 0$

(b) $\lambda = 0.5$

(c) $\lambda = 1$

Figure 1. The prior (dashed) and posterior (dash-dotted) probability densities, likelihood (dotted), and $f_a(x;\theta(\lambda))$ (solid) for $\lambda = 0$ (a), $\lambda = 0.5$ (b) and $\lambda = 1$ (c), for the example in Section IV-A. For $\lambda = 0$ and $\lambda = 1$, $f_a(x;\theta(\lambda))$ coincides with the prior and posterior probability densities, respectively.

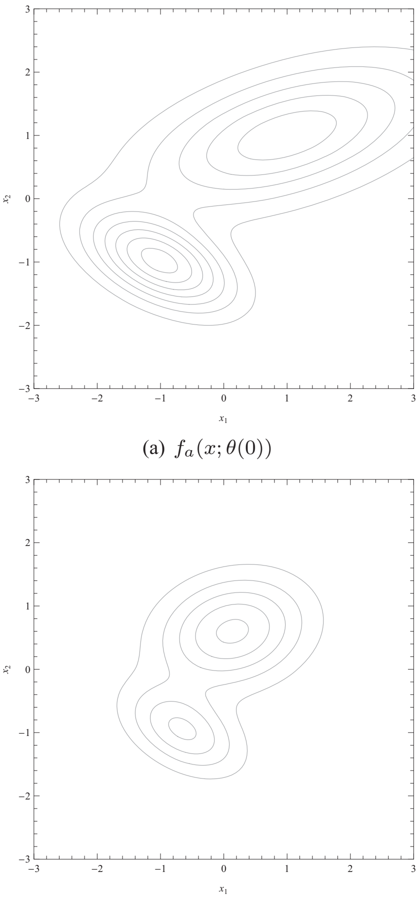
B. Gaussian Mixture Prior

As discussed in Section II, a sum of Gaussian kernels is a possible choice of prior. In this example, using bivariate densities, we assume

$$f_a(x;\theta) = \sum_{i=1}^{2} w_i N(x;\mu_i,\Sigma_i), \tag{49}$$

$$\theta = [w_1, \mu_{1,x_1}, \mu_{1,x_2}, \sigma_{1,x_1}, \sigma_{1,x_2}, \rho_1, w_2, \ldots, \rho_2]^T, \tag{50}$$

$$\Sigma_i = \left[\begin{array}{cc} \sigma_{i,x_1}^2 & \rho_i \sigma_{i,x_1} \sigma_{i,x_2} \\ \rho_i \sigma_{i,x_1} \sigma_{i,x_2} & \sigma_{i,x_2}^2 \end{array} \right]. \tag{51}$$



(a) $f_a(x;\theta(0))$

(b) $f_a(x;\theta(1))$

Figure 2. Contour plots of $f_a(x;\theta(\lambda))$, $\lambda = 0$ (a) and $\lambda = 1$ (b), for the example in Section IV-B
.

The prior probability density is given by $p(x) = f_a(x;\theta(0))$, where $\theta(0)$ is chosen arbitrarily to

$$[.3, -1., -1., .75, .5, -.5, .7, 1., 1., 1.5, .75, .5]^T. \tag{52}$$

Furthermore, we assume the likelihood $l(x) = N(x;\mu_l,\Sigma_l)$ with arbitrarily chosen values $\mu_l = [0,0]^T$ and Σ_l defined similarly as Σ_i above, with $\sigma_{l,x_1} = 1.$, $\sigma_{l,x_2} = 1.25$ and $\rho_l = -0.25$. To approximate the integrals of the resulting HDE, where the measure of deviation is given by the squared Hellinger distance, the integrals are replaced by sums of the integrands at the points of a uniform 7 by 7 grid over $[-3,3] \times [-3,3]$ (cf. Section III-B). In Figure 2, $f_a(x;\theta(\lambda))$, with $\theta(\lambda)$ obtained from the numerical solution of the HDE, is plotted for $\lambda = 0$ and $\lambda = 1$. In similar contour plots of $f(x;\lambda)$ superimposed on $f_a(x;\theta(\lambda))$, the contours coincide for $\lambda \in [0,1]$ (not shown). The solution of the AHDE yields equally similar $f(x;\lambda)$ and $f_a(x;\theta(\lambda))$ (not shown). The Hellinger distance between $f(x;1)$ and $f_a(x;\theta(1))$ computed numerically over $[-3,3] \times [-3,3]$ is very close to zero, the difference most probably due to limited machine precision.

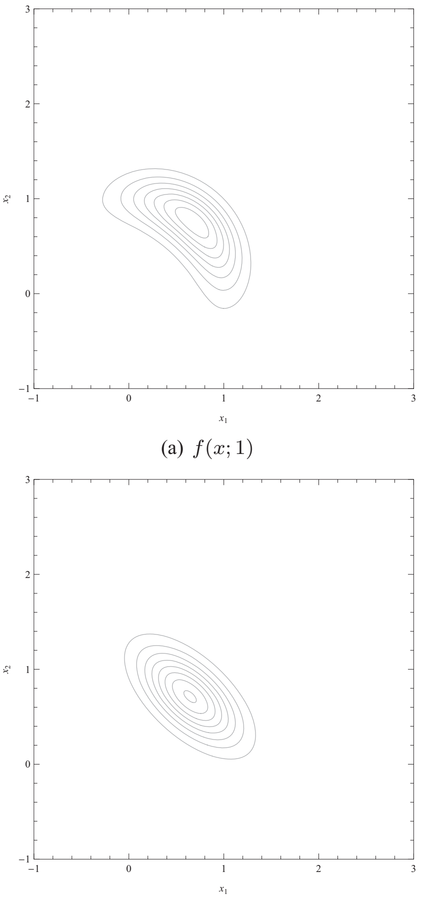
C. Non-linear Measurement Model

The next example is inspired by the filtering problem arising in radar observations. We assume a two-dimensional Gaussian prior, with $\mu_{x_1} = \mu_{x_2} = \sigma_{x_1} = \sigma_{x_2} = 1$ and $\rho = 0$, using notation similar to the that of the example in Section IV-B. The measurement model is given by (11) and (12), with the measurement $y = [y_1, y_2]^T$, where $y_1 = 1$ and $y_2 = \pi 5/18$ are the measured distance and angle, respectively. The covariance matrix of the measurement noise is

$$R = \begin{bmatrix} 0.2^2 & 0 \\ 0 & 0.6^2 \end{bmatrix}. \qquad (53)$$

The measure of deviation is given by the squared Hellinger distance. In this example, the integrals of the HDE and the AHDE are replaced by sums of the integrands at the points of a uniform 9 by 9 grid over the region $[-0.01, 1.99] \times [-0.01, 1.99]$ (cf. Section III-B). The resulting $f_a(x; \theta(1))$ from the numerical solution of the HDE is compared to the true posterior density $f(x; 1)$ in Figure 3. The posterior density inherits its non-elliptical shape, when visualized in a contour plot, from the likelihood function. This shape is typical of a radar measurement, where the uncertainty of the distance measurement is less than the uncertainty of the measurement of the angle. Since we have limited the approximation of the homotopy to a Gaussian, it is clear that $f_a(x; \theta(1))$ cannot fully capture the shape of the unnormalized posterior probability density. However, as seen in the figure, the two functions overlap to a great extent in the areas of high density.

To compare the accuracy of the proposed method to the Extended Kalman Filter (EKF), Unscented Kalman Filter (UKF) [7], the Cubature Kalman Filter (CKF) [1], and global optimization of the parameters, we compute the Hellinger distance between the posterior and the approximations of the posterior obtained from the different methods. The UKF is implemented from [11], with parameters $\alpha = 10^{-3}$, $\beta = 2$ and $\kappa = 0$, and resulting sigma point weights, as suggested in [12]. The Hellinger distance is approximated by taking the square root of (7), evaluated numerically over $[-1, 2] \times [-1, 2]$, which is a reasonable approximation, since the probability densities are close to zero outside this region. The global optimization is performed with the built-in function `NMinimize` (using the Nelder-Mead method) in MATHEMATICA, using the Hellinger distance as the objective function, and serves as a benchmark. The results are compared in Table I, where the abbreviations introduced in the text are used to denote the respective methods, and ”Global” denotes global optimization. The distances corresponding to our method, using the HDE and AHDE, are close to the distance of the benchmark, with the AHDE being less accurate than the HDE, as expected. The EKF, UKF and CKF all give greater values for the Hellinger distance than our method, for this example. We note the surprising result that in this simulation, the UKF and CKF perform less well than the EKF. The computational complexities of the HDE and AHDE, requiring the solution of a system of ODEs, are significantly greater than those of the EKF, UKF and CKF. The HDE and AHDE were solved in $1.078\,\mathrm{s}$ and $0.890\,\mathrm{s}$ on a regular desktop computer, respectively. This should be compared to the $174\,\mathrm{s}$ spent by the global optimization algorithm, with only a modest improvement in the Hellinger distance between the resulting approximation and the true posterior.



(a) $f(x; 1)$

(b) $f_a(x; \theta(1))$

Figure 3. Contour plots of the true posterior density $f(x; 1)$ (a) and $f_a(x; \theta(1))$ (b) for the example in Section IV-C.

Table I
APPROXIMATE HELLINGER DISTANCE BETWEEN THE TRUE POSTERIOR AND ITS APPROXIMATION COMPUTED USING DIFFERENT METHODS, FOR THE EXAMPLE IN SECTION IV-C.

Method	Hellinger distance
HDE	0.172
AHDE	0.199
EKF	0.300
UKF	0.415
CKF	0.703
Global	0.167

D. Point-Value Parameterization

This last example demonstrates how a very flexible parameterization, approximating the densities through interpolation

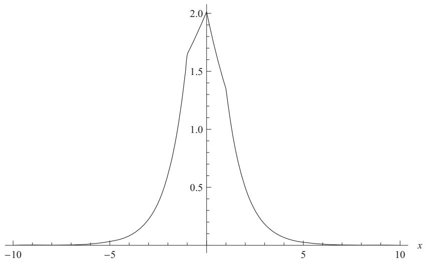


Figure 4. The prior density, obtained from interpolation between point-value pairs, for the example in section IV-D.

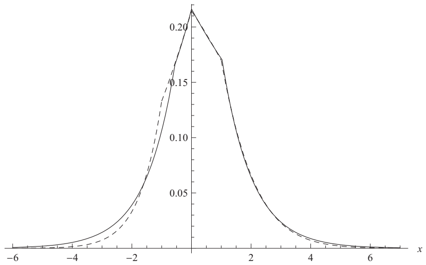


Figure 5. The unnormalized posterior (dashed) and its approximation (solid), obtained from the HDE as $f_a(x;\theta(1))$, for the example in section IV-D.

between point-value pairs, as outlined in Section II, can be used with the proposed framework. We use the parameterized density function

$$f_a(x;\theta) = \exp(g(x;\theta)), \tag{54}$$

where θ is given by (5), and $g(x;\theta)$ gives the linear interpolation of the point-value pairs $\{x_i, v_i\}$, $i = 1, \ldots, n$, with the boundary conditions

$$\frac{dg(x;\theta)}{dx} = \begin{cases} 1, & x < \min_{1\leq i\leq n} x_i, \\ -1, & x > \max_{1\leq i\leq n} x_i, \end{cases} \tag{55}$$

which were chosen arbitrarily for the sake of our example. We let $n = 3$, $\theta(0) = [-1, 0, 1, 0.5, 0.7, 0.3]^T$, giving a prior density as shown in Figure 4. While this is not a probability density, the given parameter values are used for ease of presentation. The parameter values could be scaled to yield a prior density that integrates to 1, to make the example more true to the intended application. The likelihood is taken to be $l(x) = N(x; 2, 9)$. Using the method described in Section III-B to approximate the integrals of the HDE, in which the measure of deviation is given by the squared Hellinger distance, a solution of the HDE at $\lambda = 1$ is obtained that is close to the posterior computed from pointwise multiplication of the prior density and the likelihood, as shown in Figure 5.

V. Conclusion

We have elaborated on a method for computing the optimal parameter values to represent the posterior probability density of a Bayesian measurement updated. The method is based on tracking the minimum of a measure of deviation between a parametrized density function and a homotopy, that deforms the prior density into the posterior density. Our contributions are a formalization of the minimization criterion and the resulting system of ODEs for both arbitrary and specific measures of deviation, and methods for approximate solutions of the problem. We have shown that the method and the approximations can be used successfully with common parameterizations of density functions, including cases where no analytical solution is available, where the method yielded more accurate results than the Extended, Unscented and Cubature Kalman Filters. We hope that the framework will aid the development of filtering algorithms for non-linear and non-Gaussian problems, where flexible parameterizations of the probability densities are needed.

VI. Acknowledgement

This work has been supported by the European Commission 7th Framework Programme (UNICELLSYS, grant No 201142) and the Swedish Foundation for Strategic Research through the Gothenburg Mathematical Modelling Centre.

References

[1] I. Arasaratnam and S. Haykin, "Cubature Kalman Filters," *Automatic Control, IEEE Transactions on*, vol. 54, no. 6, pp. 1254–1269, 2009.

[2] F. Daum, "Nonlinear Filters: Beyond the Kalman Filter," *IEEE Aerospace And Electronic Systems Magazine*, vol. 20, no. 8, pp. 57–69, 2005.

[3] F. Daum and J. Huang, "Nonlinear Filters with Particle Flow Induced by Log-homotopy," in *Proc. of SPIE, Signal Processing, Sensor Fusion, and Target Recognition XVIII*, I. Kadar, Ed., vol. 7336 733603-1, Orlando, FL, USA, April 13, 2009.

[4] N.J. Gordon, D.J. Salmond and A.F.M. Smith, "Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation," *Radar and Signal Processing, IEE Proceedings F*, vol. 140, no. 2, pp. 107–113, 1993.

[5] U.D. Hanebeck, K. Briechle and A. Rauh, "Progressive Bayes: a New Framework for Nonlinear State Estimation," in *Proc. SPIE 2003, Multisource Information Fusion: Architectures, Algorithms, and Applications*, B.V. Dasarathy, Ed., vol. 5099, Orlando, FL, April 23, 2003, pp. 256–267.

[6] A.H. Jazwinski, *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.

[7] S.J. Julier and J.K. Uhlmann, "Unscented Filtering and Nonlinear Estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[8] S. Kakutani, "On Equivalence of Infinite Product Measures," *The Annals of Mathematics*, vol. 49, no. 1, pp. 214–224, 1948.

[9] S. Kullback and R.A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[10] A. Masud and L.A. Bergman, "Solution of the Four Dimensional Fokker-Planck Equation: Still a Challenge," in *Proceedings of the ICOSSAR 2005 Conference*, G. Augusti, G.I. Schuëller, M. Ciampoli, Eds., Rome, Italy, June 22-26, 2005. pp. 1911–1916.

[11] S. Särkkä, "On Unscented Kalman Filtering for State Estimation of Continuous-Time Nonlinear Systems," *Automatic Control, IEEE Transactions on*, vol. 52, no. 9, pp. 1631–1641, 2007.

[12] E.A. Wan and R. van der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, Lake Louise, Alta., Canada, October 1-4, 2000, pp. 153–158.