Bayesian Quadrature in Nonlinear Filtering

Jakub Prüher and Miroslav Šimandl
European Centre of Excellence - New Technologies for the Information Society,
Faculty of Applied Sciences, University of West Bohemia, Univerzitní 18, Pilsen, Czech Republic

Keywords: Nonlinear Filtering, Bayesian Quadrature, Gaussian Process.

Abstract: The paper deals with the state estimation of nonlinear stochastic discrete-time systems by means of quadrature-based filtering algorithms. The algorithms use quadrature to approximate the moments given by integrals. The aim is at evaluation of the integral by Bayesian quadrature. The Bayesian quadrature perceives the integral itself as a random variable, on which inference is to be performed by conditioning on the function evaluations. Advantage of this approach is that in addition to the value of the integral, the variance of the integral is also obtained. In this paper, we improve estimation of covariances in quadrature-based filtering algorithms by taking into account the integral variance. The proposed modifications are applied to the Gauss-Hermite Kalman filter and the unscented Kalman filter algorithms. Finally, the performance of the modified filters is compared with the unmodified versions in numerical simulations. The modified versions of the filters exhibit significantly improved estimate credibility and a comparable root-mean-square error.

1 INTRODUCTION

Dynamic systems are widely used to model behaviour of real processes throughout the sciences. In many cases, it is useful to define a state of the system and consequently work with a state-space representation of the dynamics. When the dynamics exhibits stochasticity or can only be observed indirectly, the problem of state estimation becomes relevant. Estimating a state of the dynamic system from noisy measurements is a prevalent problem in many application areas such as aircraft guidance, GPS navigation (Grewal et al., 2007), weather forecast (Gillijns et al., 2006), telecommunications (Jiang et al., 2003) and time series analysis (Bhar, 2010). When the state estimator is required to produce an estimate using only the present and past measurements, this is known as the filtering problem.

For a discrete-time linear Gaussian systems, the best estimator in the mean-square-error sense is the much-celebrated Kalman filter (KF) (Kalman, 1960). First attempts to deal with the estimation of nonlinear dynamics can be traced to the work of (Smith et al., 1962), which resulted in the extended Kalman filter (EKF). The EKF algorithm uses the Taylor series expansion to approximate the nonlinearities in the system description. A disadvantage of the Taylor series is that it requires differentiability of the approximated functions. This prompted further development (Nørgaard et al., 2000; Šimandl and Duník, 2009) resulting in the derivative-free filters based on the Stirling's interpolation formula. Other approaches that approximate nonlinearities include the Fourier-Hermite KF (Sarmavuori and Särkkä, 2012), special case of which is the statistically linearized filter (Maybeck, 1982; Gelb, 1974).

Instead of explicitly dealing with nonlinearities in the system description, the unscented Kalman filter (UKF) (Julier et al., 2000) describes the densities by a finite set of deterministically chosen σ-points, which are then propagated through the nonlinearity. Other filters, such as the Gauss-Hermite Kalman filter (GHKF) (Ito and Xiong, 2000), the cubature Kalman filter (CKF) (Arasaratnam and Haykin, 2009) and the stochastic integration filter (Duník et al., 2013), utilize numerical quadrature rules to approximate moments of the relevant densities. These filters can be seen as representatives of a more general σ-point methodology.

A limitation of classical integral approximations, such as the Gauss-Hermite quadrature (GHQ), is that they are specifically designed to perform with zero error on a narrow class of functions (typically polynomials up to a given degree (Särkkä, 2013)). It is also possible to design rules, that have best average-case performance on a wider range of functions at the cost of permitting small non-zero error (Minka, 2000). In recent years, the Bayesian quadrature (BQ)

Prüher J. and Šimandl M..
Bayesian Quadrature in Nonlinear Filtering.
DOI: 10.5220/0005534003800387
In *Proceedings of the 12th International Conference on Informatics in Control, Automation and Robotics* (ICINCO-2015), pages 380-387
ISBN: 978-989-758-122-9

has become a focus of interest in probabilistic numerics community (Osborne et al., 2012). The BQ treats numerical integration as a problem of Bayesian inference and thus it is able to provide an additional information - namely, uncertainty in the computation of the integral itself. In (Särkkä et al., 2014), the authors work with the concept of BQ, but the algorithms derived therein do not make use of the uncertainty in the integral computations.

The goal of this paper is to augment the current σ-point algorithms so that the uncertainty associated with the integral approximations is also reflected in their estimates.

The rest of the paper is organized as follows. Formal definition of the Gaussian filtering problem is outlined in Section 2, followed by the exposition of the basic idea of Bayesian quadrature in Section 3. The main contribution, which is the design of the Bayes-Hermite Kalman filter (BHKF), is presented in Section 4. Finally, comparison of the BHKF with existing filters is made in Section 5.

2 PROBLEM FORMULATION

The discrete-time stochastic dynamic system is described by the following state-space model

$$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}) + \mathbf{q}_{k-1}, \qquad \mathbf{q}_{k-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \tag{1}$$

$$\mathbf{z}_k = \mathbf{h}(\mathbf{x}_k) + \mathbf{r}_k, \qquad \mathbf{r}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \tag{2}$$

with initial conditions $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{m}_0, \mathbf{P}_0)$, where $\mathbf{x}_k \in \mathbb{R}^n$ is the system state evolving according to the known nonlinear dynamics $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ perturbed by the white state noise $\mathbf{w}_{k-1} \in \mathbb{R}^n$. Measurement $\mathbf{z}_k \in \mathbb{R}^p$ is a result of applying known nonlinear transformation $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^p$ to the system state and white additive measurement noise $\mathbf{r}_k \in \mathbb{R}^p$. The mutual independence is assumed between the state noise $\mathbf{w}_k$, the measurement noise $\mathbf{r}_k$ and the system initial condition $\mathbf{x}_0$ for all $k \geq 1$.

The filtering problem is concerned with determination of the probability density function $p(\mathbf{x}_k \mid \mathbf{z}_{1:k})$. The shorthand $\mathbf{z}_{1:k}$ stands for the sequence of measurements $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k$. The general solution to the filtering problem is given by the Bayesian recursive relations in the form of density functions

$$p(\mathbf{x}_k | \mathbf{z}_{1:k}) = \frac{p(\mathbf{z}_k | \mathbf{x}_k)\, p(\mathbf{x}_k | \mathbf{z}_{1:k-1})}{p(\mathbf{z}_k | \mathbf{z}_{1:k-1})}, \tag{3}$$

with predictive density $p(\mathbf{x}_k \mid \mathbf{z}_{1:k-1})$ given by the Chapman-Kolmogorov equation

$$p(\mathbf{x}_k | \mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k | \mathbf{x}_{k-1}) p(\mathbf{x}_{k-1} | \mathbf{z}_{1:k-1})\, \mathrm{d}\mathbf{x}_{k-1}. \tag{4}$$

In this paper, the integral computation is assumed to take place over the support of $\mathbf{x}_{k-1}$. The likelihood term $p(\mathbf{z}_k \mid \mathbf{x}_k)$ in (3) is determined by the measurement model (2) and the transition probability $p(\mathbf{x}_k | \mathbf{x}_{k-1})$ in (4) by the dynamics model (1).

For tractability reasons, the Gaussian filters make simplifying assumption, that the joint density of state and measurement $p(\mathbf{x}_k, \mathbf{z}_k \mid \mathbf{z}_{1:k-1})$ is of the form

$$\mathcal{N}\left(\begin{bmatrix} \mathbf{x}_{k|k-1} \\ \mathbf{z}_{k|k-1} \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m}^x_{k|k-1} \\ \mathbf{m}^z_{k|k-1} \end{bmatrix}, \begin{bmatrix} \mathbf{P}^x_{k|k-1} & \mathbf{P}^{xz}_{k|k-1} \\ \mathbf{P}^{zx}_{k|k-1} & \mathbf{P}^z_{k|k-1} \end{bmatrix} \right). \tag{5}$$

Knowledge of the moments in (5) is fully sufficient (Deisenroth and Ohlsson, 2011) to express the first two moments, $\mathbf{m}^x_{k|k}$ and $\mathbf{P}^x_{k|k}$, of the conditional density $p(\mathbf{x}_k \mid \mathbf{z}_{1:k})$ using the conditioning formula for Gaussians as

$$\mathbf{m}^x_{k|k} = \mathbf{m}^x_{k|k-1} + \mathbf{K}_k \left(\mathbf{z}_k - \mathbf{m}^z_{k|k-1} \right), \tag{6}$$

$$\mathbf{P}^x_{k|k} = \mathbf{P}^x_{k|k-1} - \mathbf{K}_k \mathbf{P}^z_{k|k-1} \mathbf{K}_k^\top, \tag{7}$$

with the Kalman gain defined as $\mathbf{K}_k = \mathbf{P}^{xz}_{k|k-1} \left(\mathbf{P}^z_{k|k-1} \right)^{-1}$.

The problem of computing the moments in (5) can be seen, on a general level, as a computation of moments of a transformed random variable

$$\mathbf{y} = \mathbf{g}(\mathbf{x}), \tag{8}$$

where $\mathbf{g}$ is a nonlinear vector function. This invariably entails evaluation of the integrals of the following kind

$$\mathbb{E}[\mathbf{y}] = \int \mathbf{g}(\mathbf{x}) p(\mathbf{x})\, \mathrm{d}\mathbf{x} \tag{9}$$

with Gaussian $p(\mathbf{x})$. Since the integral is typically intractable, σ-point algorithms resort to the approximations based on weighted sum of function evaluations

$$\int \mathbf{g}(\mathbf{x}) p(\mathbf{x})\, \mathrm{d}\mathbf{x} \approx \sum_{i=1}^{N} w_i \mathbf{g}(\mathbf{x}^{(i)}). \tag{10}$$

The evaluation points $\mathbf{x}^{(i)}$ are also known as the σ-points, hence the name.

Thus, for instance, to compute $\mathbf{m}^x_{k|k-1}$, $\mathbf{P}^x_{k|k-1}$ and $\mathbf{P}^{xz}_{k|k-1}$, the following expressions, given in the matrix notation, are used

$$\mathbf{m}^x_{k|k-1} \simeq \mathbf{F}^\top \mathbf{w}, \tag{11}$$

$$\mathbf{P}^x_{k|k-1} \simeq \tilde{\mathbf{F}}^\top \mathbf{W} \tilde{\mathbf{F}}, \tag{12}$$

$$\mathbf{P}^{xz}_{k|k-1} \simeq \tilde{\mathbf{X}}^\top \mathbf{W} \tilde{\mathbf{F}}, \tag{13}$$

where the weights are now $\mathbf{w} = [\, w_1, \ldots, w_N \,]^\top$, $\mathbf{W} = \mathrm{diag}\,([w_1, \ldots, w_N])$ and the i-th rows of $\mathbf{F}$, $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{X}}$ are defined as the transpose of $\mathbf{f}(\mathbf{x}^{(i)}_{k-1})$, $\mathbf{f}(\mathbf{x}^{(i)}_{k-1}) - \mathbf{m}^x_{k|k-1}$ and $\mathbf{x}^{(i)}_{k-1} - \mathbf{m}^x_{k|k-1}$ respectively.

All the information a quadrature rule has about the function behaviour is conveyed by the N function values $\mathbf{g}(\mathbf{x}^{(i)})$. Conversely, this means that any quadrature is uncertain about the true function values in between the σ-points. The importance of quantifying this uncertainty becomes particularly pronounced, when the function is not integrated exactly due to the inherent design limitations of the quadrature (such as the choice of weights and σ-points). All σ-point filters thus operate with the uncertainty, which is not accounted for in their estimates. The classical treatment of the quadrature does not lend itself nicely to the quantification of the uncertainty associated with a given rule. On the other hand, the Bayesian quadrature, which treats the integral approximation as a problem in Bayesian inference, is perfectly suited for this task.

The Idea of using Bayesian quadrature in the state estimation algorithms was already treated in (Särkkä et al., 2014). The derived filters and smoothers, however, do not fully utilize the potential of the Bayesian quadrature. Namely, variance of the integral is not reflected in their estimates, which still remains a problem to this day.

3 GAUSSIAN PROCESS PRIORS AND BAYESIAN QUADRATURE

In this section, we introduce the key concepts of Gaussian process priors and Bayesian quadrature, which are crucial to the derivation of the filtering algorithm in Section 4.

3.1 Gaussian Process Priors

Uncertainty over functions is naturally expressed by a stochastic process. In Bayesian quadrature, Gaussian processes (GP) are used for their favourable analytical properties. Gaussian process is a collection of random variables indexed by elements of an index set, any finite number of which has a joint Gaussian density (Rasmussen and Williams, 2006). That is, for any finite set of indices $\mathbf{X}' = \{\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_m\}$, it holds that

$$(\, g(\mathbf{x}'_1),\ g(\mathbf{x}'_2),\ \ldots,\ g(\mathbf{x}'_m)\,)^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \qquad (14)$$

where the kernel (covariance) matrix $\mathbf{K}$ is made up of pair-wise evaluations of the kernel function, thus $[\mathbf{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. Choosing a kernel, which in principle can be any symmetric positive definite function of two arguments, introduces assumptions about the underlying function we are trying to model. Bayesian inference allows to combine the GP prior $p(g)$ with the data, $\mathcal{D} = \{(\mathbf{x}_i, g(\mathbf{x}_i)), i = 1, \ldots, N\}$ comprising the evaluation points $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ and the function evaluations $\mathbf{y}_g = [\, g(\mathbf{x}_1), \ldots, g(\mathbf{x}_N)\,]^\top$, to produce a GP posterior $p(g \mid \mathcal{D})$ with moments given by (Rasmussen and Williams, 2006)

$$\mathbb{E}_g[g(\mathbf{x}')] = m_g(\mathbf{x}') = \mathbf{k}^\top(\mathbf{x}')\mathbf{K}^{-1}\mathbf{y}_g, \qquad (15)$$

$$\mathbb{V}_g[g(\mathbf{x}')] = \sigma_g^2(\mathbf{x}') = k(\mathbf{x}', \mathbf{x}') - \mathbf{k}^\top(\mathbf{x}')\mathbf{K}^{-1}\mathbf{k}(\mathbf{x}'), \qquad (16)$$

where $\mathbf{k}(\mathbf{x}') = [k(\mathbf{x}', \mathbf{x}_1), \ldots, k(\mathbf{x}', \mathbf{x}_N)]^\top$. Thus, for any test input $\mathbf{x}'$, we recover a Gaussian posterior predictive density over the function values $g(\mathbf{x}')$. Figure 1 depicts predictive moments of the GP posterior density. Notice, that in between the evaluations, where the true function value is not known, the GP model is uncertain.

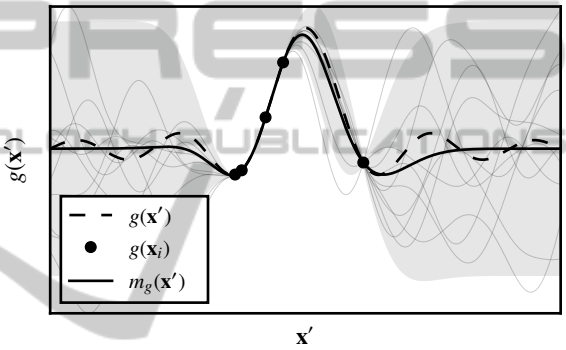

Figure 1: True function (dashed), GP posterior mean (solid), observed function values (dots) and GP posterior samples (grey). The shaded area represents GP posterior predictive uncertainty ($\pm 2\,\sigma_g(\mathbf{x}')$). Notice the collapse of uncertainty around the observations.

3.2 Bayesian Quadrature

The problem of numerical quadrature pertains to the approximate computation of the integral

$$\mathbb{E}_\mathbf{x}[g(\mathbf{x})] = \int g(\mathbf{x}) p(\mathbf{x})\, \mathrm{d}\mathbf{x}. \qquad (17)$$

The key distinguishing feature of the BQ is that it "treats the problem of numerical integration as the one of statistical inference." (O'Hagan, 1991) This is achieved by placing a prior density over the integrated functions themselves. Consequence of this is that the integral itself is then a random variable as well. Concretely, if GP prior density is used, then the value of the integral of the function will also be Gaussian distributed. This follows from the fact that integral is a linear operator acting on the GP distributed random function $g(\mathbf{x})$.

Following the line of thought of (Rasmussen and Ghahramani, 2003) we take expectation (with respect

to $p(g \mid \mathcal{D})$) of the integral (17) and obtain

$$\mathbb{E}_{g|\mathcal{D}}[\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})]] = \iint g(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}\,p(g \mid \mathcal{D})\,\mathrm{d}g$$
$$\iint g(\mathbf{x})p(g \mid \mathcal{D})\,\mathrm{d}g p(\mathbf{x})\,\mathrm{d}\mathbf{x} = \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{g|\mathcal{D}}[g(\mathbf{x})]]. \quad (18)$$

From (18) we see, that taking the expectation of integral is the same as integrating the GP posterior mean function, which effectively approximates the integrated function $g(x)$. Variance of the integral is

$$\mathbb{V}_{g|D}[\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})]] = \iint k(\mathbf{x}, \mathbf{x}')p(\mathbf{x})p(\mathbf{x}')\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}'. \quad (19)$$

A popular choice of kernel function, that enables the expressions (18) and (19) to be computed analytically is an Exponentiated Quadratic (EQ)

$$k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) = \alpha^2 \exp\left(-\frac{1}{2}\left(\mathbf{x}_i - \mathbf{x}_j\right)^\top \boldsymbol{\Lambda}^{-1}\left(\mathbf{x}_i - \mathbf{x}_j\right)\right), \quad (20)$$

where the vertical lengthscale α and the horizontal lengthscales on diagonal of $\boldsymbol{\Lambda} = \mathrm{diag}([\ell_1^2, \ldots, \ell_n^2])$ are kernel hyper-parameters, collectively denoted by the symbol $\boldsymbol{\theta}$. By using this particular kernel the assumption of smoothness (infinite differentiability) of the integrand is introduced (Rasmussen and Williams, 2006). Given the kernel function in the form (20) and $p(\mathbf{x}) = \mathcal{N}(\mathbf{m}, \mathbf{P})$, the expressions for the integral posterior mean and variance reduce to

$$\mathbb{E}_{g|D}[\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})]] = \mathbf{l}^\top \mathbf{K}^{-1}\mathbf{y}_g, \quad (21)$$

$$\mathbb{V}_{g|D}[\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})]] = \alpha^2 \left|2\boldsymbol{\Lambda}^{-1}\mathbf{P} + \mathbf{I}\right|^{-1/2} - \mathbf{l}^\top \mathbf{K}^{-1}\mathbf{l}, \quad (22)$$

with $\mathbf{l} = [l_1, \ldots, l_N]^\top$

$$l_i = \int k(\mathbf{x}, \mathbf{x}_i)\,\mathcal{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{P})\,\mathrm{d}\mathbf{x} = \alpha^2 \left|\boldsymbol{\Lambda}^{-1}\mathbf{P} + \mathbf{I}\right|^{-1/2} \times \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{m})^\top(\boldsymbol{\Lambda} + \mathbf{P})^{-1}(\mathbf{x}_i - \mathbf{m})\right). \quad (23)$$

Notice that we could define weights as $\mathbf{w} = \mathbf{l}^\top \mathbf{K}^{-1}$. Then the expression (21) is just a weighted sum of function evaluations, conforming to the general σ-point method as described by (11). As opposed to classical quadrature rules, that prescribe the precise locations of σ-points, BQ makes no such restrictions. In (Minka, 2000), the optimal placement is determined by minimizing the posterior variance of the integral (19).

In the next section, we show how the integral variance (19) can be reflected in the current nonlinear filtering quadrature-based algorithms.

4 BAYES-HERMITE KALMAN FILTER

In this section, we show how the integral variance can be incorporated into the moment estimates of the transformed random variable. Parallels are drawn with existing GP-based filters and the Bayes-Hermite Kalman filter algorithm is outlined.

4.1 Incorporating Integral Uncertainty

Uncertainty over the function values is introduced by a GP posterior $p(g \mid \mathcal{D})$, whose mean function (15) acts effectively as an approximation to the deterministic function g. Note that the equations (15), (16) can only be used to model single output dimension of the vector function $\mathbf{g}$. For now, we will assume a scalar function g unless otherwise stated. To keep the notation uncluttered, conditioning on $\mathcal{D}$ will be omitted. Treating the function values $g(\mathbf{x})$ as random leads to the joint density $p(g, \mathbf{x})$ and thus, when computing the moments of $g(\mathbf{x})$, the expectations need to be taken with respect to both variables. This results in the following approximation of the true moments

$$\mu = \mathbb{E}_x[g(\mathbf{x})] \approx \mathbb{E}_{g,x}[g(\mathbf{x})], \quad (24)$$

$$\sigma^2 = \mathbb{V}_x[g(\mathbf{x})] \approx \mathbb{V}_{g,x}[g(\mathbf{x})]. \quad (25)$$

Using the law of iterated expectations, we get

$$\mathbb{E}_{g,x}[g(\mathbf{x})] = \mathbb{E}_g[\mathbb{E}_x[g(\mathbf{x})]] = \mathbb{E}_x[\mathbb{E}_g[g(\mathbf{x})]]. \quad (26)$$

This fact was used to derive weights for the filtering and smoothing algorithms in (Särkkä et al., 2014), where the same weights were used in computations of means and covariances. Our proposed approach, however, proceeds differently in derivation of weights used in the computation of covariance matrices.

Note that the term for variance can be written out using the decomposition formula either as

$$\mathbb{V}_{g,x}[g(\mathbf{x})] = \mathbb{E}_x[\mathbb{V}_g[g(\mathbf{x})]] + \mathbb{V}_x[\mathbb{E}_g[g(\mathbf{x})]] \quad (27)$$

or as

$$\mathbb{V}_{g,x}[g(\mathbf{x})] = \mathbb{E}_g[\mathbb{V}_x[g(\mathbf{x})]] + \mathbb{V}_g[\mathbb{E}_x[g(\mathbf{x})]] \quad (28)$$

depending on which factorization of the joint density $p(g, \mathbf{x})$ is used. The terms $\mathbb{V}_g[g(\mathbf{x})]$ and $\mathbb{V}_g[\mathbb{E}_x[g(\mathbf{x})]]$ can be identified as variance of the *integrand* and variance of the *integral* respectively. In case of deterministic g, both of these terms are zero.

With EQ covariance (20), the expression (26) for the first moment of a transformed random variable takes on the form (21). Since the variance decompositions in (27) and (28) are equivalent, both can be used to achieve the same goal.

The form (27) was utilized in derivation of the Gaussian process - assumed density filter (GP-ADF) (Deisenroth et al., 2012), which relies on the solution to the problem of prediction with GPs at uncertain inputs (Girard et al., 2003). So, even though these results were derived to solve a seemingly different problem, we point out, that by using the form (27), the uncertainty of the integral (as seen in the last term of (28)) is implicitly reflected in the resulting covariance. To conserve space, we only provide a summary of the results in (Deisenroth et al., 2009) and point reader to the said reference for detailed derivations. The expressions for the moments of transformed variable were rewritten into a form, which assumes that a single GP is used to model all the output dimensions of the vector function (8)

$$\boldsymbol{\mu} = \mathbf{G}^\top \mathbf{w}, \tag{29}$$

$$\boldsymbol{\Sigma} = \mathbf{G}^\top \mathbf{W} \mathbf{G} - \boldsymbol{\mu}\boldsymbol{\mu}^\top + \operatorname{diag}\left(\alpha^2 - \operatorname{tr}\left(\mathbf{K}^{-1}\mathbf{L}\right)\right), \tag{30}$$

with matrix $\mathbf{G}$ being defined analogously to $\mathbf{F}$ in (11)-(13). The weights are given as

$$\mathbf{w} = \mathbf{K}^{-1}\mathbf{l} \text{ and } \mathbf{W} = \mathbf{K}^{-1}\mathbf{L}\mathbf{K}^{-1}, \tag{31}$$

where

$$\mathbf{L} = \int k(\mathbf{X}, \mathbf{x}; \boldsymbol{\theta}_g)\, k(\mathbf{x}, \mathbf{X}; \boldsymbol{\theta}_g)\, \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{P})\, \mathrm{d}\mathbf{x}. \tag{32}$$

The equations (29) and (30) bear certain resemblance to the σ-point method in (11), (12); however, in this case matrix $\mathbf{W}$ is not diagonal. Note, that the weights depend on the current location of σ-points and need to be recomputed at every time step.

4.2 BHKF Algorithm

The filtering algorithm based on the BQ can now be constructed utilizing (29) and (30). The BHKF uses two GPs with the EQ covariance - one for each function in the state-space model (1)-(2), which means that the two sets of hyper-parameters are used; $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_h$. In the algorithm specification below, the lower index of $\mathbf{q}$ and $\mathbf{K}$ specifies the set of hyper-parameters used to construct these quantities.

Algorithm (Bayes-Hermite Kalman Filter)**.** In the following, let $\mathbf{x}_{0|0} \sim \mathcal{N}(\mathbf{m}_{0|0}, \mathbf{P}_{0|0})$, $i = 1, \ldots, N$ and $k = 1, 2, \ldots$.

Initialization:
Choose unit σ-points $\boldsymbol{\xi}^{(i)}$. Set hyper-parameters $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_h$. Proceed from the initial conditions $\mathbf{x}_{0|0}$, for all k, by alternating between the following prediction and filtering steps.

Prediction:

1. Form the σ-points $\mathbf{x}_{k-1}^{(i)} = \mathbf{m}_{k-1|k-1}^x + \sqrt{\mathbf{P}_{k-1|k-1}^x}\,\boldsymbol{\xi}^{(i)}$.
2. Propagate σ-points through the dynamics model $\mathbf{x}_k^{(i)} = \mathbf{f}\left(\mathbf{x}_{k-1}^{(i)}\right)$ and form $\mathbf{F}$ as in (11)-(13).
3. Using $\mathbf{x}_k^{(i)}$ and hyper-parameters $\boldsymbol{\theta}_f$, compute weights $\mathbf{w}^x$ and $\mathbf{W}^x$ according to (31) and (32).
4. Compute predictive mean $\mathbf{m}_{k|k-1}^x$ and predictive covariance $\mathbf{P}_{k|k-1}^x$

$$\begin{aligned}\mathbf{m}_{k|k-1}^x &= \mathbf{F}^\top \mathbf{w}^x \\ \mathbf{P}_{k|k-1}^x &= \mathbf{F}^\top \mathbf{W}^x \mathbf{F} - \mathbf{m}_{k|k-1}^x (\mathbf{m}_{k|k-1}^x)^\top \\ &\quad + \operatorname{diag}\left(\alpha^2 - \operatorname{tr}\left(\mathbf{K}^{-1}\mathbf{L}\right)\right) + \mathbf{Q}\end{aligned}$$

Filtering:

1. Form the σ-points $\mathbf{x}_k^{(i)} = \mathbf{m}_{k|k-1}^x + \sqrt{\mathbf{P}_{k|k-1}^x}\,\boldsymbol{\xi}^{(i)}$.
2. Propagate the σ-points through the measurement model $\mathbf{z}_k^{(i)} = \mathbf{h}\left(\mathbf{x}_k^{(i)}\right)$, and form $\mathbf{H}$ as in (11)-(13)
3. Using $\mathbf{x}_k^{(i)}$ and hyper-parameters $\boldsymbol{\theta}_h$, compute weights $\mathbf{w}^z$ and $\mathbf{W}^z$ according to (31) and (32). Construct $\mathbf{W}^{xz} = \operatorname{diag}(\mathbf{l}_h)\,\mathbf{K}_h^{-1}$.
4. Compute measurement mean, covariance and state-measurement cross-covariance

$$\begin{aligned}\mathbf{m}_{k|k-1}^z &= \mathbf{H}^\top \mathbf{w}^z \\ \mathbf{P}_{k|k-1}^z &= \mathbf{H}^\top \mathbf{W}^z \mathbf{H} - \mathbf{m}_{k|k}^z (\mathbf{m}_{k|k}^z)^\top, \\ &\quad + \operatorname{diag}\left(\alpha^2 - \operatorname{tr}\left(\mathbf{K}^{-1}\mathbf{L}\right)\right) + \mathbf{R} \\ \mathbf{P}_{k|k-1}^{xz} &= \mathbf{P}_{k|k-1}^x (\mathbf{P}_{k|k-1}^x + \boldsymbol{\Lambda})^{-1} \tilde{\mathbf{X}} \mathbf{W}^{xz} \mathbf{H},\end{aligned}$$

   where the i-th row of $\tilde{\mathbf{X}}$ is $\mathbf{x}_k^{(i)} - \mathbf{m}_{k|k-1}^x$
5. Compute the filtered mean $\mathbf{m}_{k|k}^x$ and filtered covariance $\mathbf{P}_{k|k}^x$

$$\begin{aligned}\mathbf{m}_{k|k}^x &= \mathbf{m}_{k|k-1}^x + \mathbf{K}_k \left(\mathbf{z}_k - \mathbf{m}_{k|k-1}^z\right), \\ \mathbf{P}_{k|k}^x &= \mathbf{P}_{k|k-1}^x - \mathbf{K}_k \mathbf{P}_{k|k-1}^z \mathbf{K}_k^\top.\end{aligned}$$

   with Kalman gain $\mathbf{K}_k = \mathbf{P}_{k|k-1}^{xz} \left(\mathbf{P}_{k|k-1}^z\right)^{-1}$.

5 NUMERICAL ILLUSTRATION

In the numerical simulations the performance of the filters was tested on a univariate non-stationary growth model (UNGM) (Gordon et al., 1993)

$$x_k = \frac{1}{2}x_{k-1} + \frac{25 x_{k-1}}{1 + x_{k-1}^2} + 8\cos(1.2\,k) + q_{k-1}, \tag{33}$$

$$z_k = \frac{1}{20}x_{k-1}^2 + r_k, \tag{34}$$

with the state noise $q_{k-1} \sim \mathcal{N}(0, 10)$, measurement noise $r_k \sim \mathcal{N}(0, 1)$ and initial conditions $x_{0|0} \sim \mathcal{N}(0, 5)$.

Since the BHKF does not prescribe the σ-point locations, they can be chosen at will. The GHKF based on the r-th order Gauss-Hermite (GH) quadrature rule uses σ-points, which are determined as the roots of the r-th degree univariate Hermite polynomial $H_r(x)$. When it is required to integrate function of a vector argument ($n > 1$), a multidimensional grid of points is formed by the Cartesian product, leading to their exponential growth ($N = r^n$). The GH weights are computed according to (Särkkä, 2013) as

$$w_i = \frac{r!}{[rH_{r-1}(x^{(i)})]^2}. \tag{35}$$

The Unscented Transform (UT) is also a simple quadrature rule (Ito and Xiong, 2000), that uses $N = 2n + 1$ deterministically chosen σ-points,

$$\mathbf{x}^{(i)} = \mathbf{m} + \sqrt{\mathbf{P}}\boldsymbol{\xi}^{(i)} \tag{36}$$

with unit σ-points defined as columns of the matrix

$$\left[\, \boldsymbol{\xi}^{(0)},\ \boldsymbol{\xi}^{(1)},\ \ldots,\ \boldsymbol{\xi}^{(2n+1)} \,\right] = \left[\, \mathbf{0},\ c\mathbf{I}_n,\ -c\mathbf{I}_n \,\right] \tag{37}$$

where $\mathbf{I}_n$ denotes $n \times n$ identity matrix. The corresponding weights are defined by

$$w_0 = \frac{\kappa}{n+\kappa}, \quad w_i = \frac{1}{2(n+\kappa)}, \quad i = 1, \ldots, 2n \tag{38}$$

with scaling factor $c = \sqrt{n+\kappa}$. All of the BHKFs used the same set of hyper-parameters $\boldsymbol{\theta}_f = \boldsymbol{\theta}_h = [\ell, \alpha]^\top = [3, 1]^\top$. UKF operated with $\kappa = 2$. BHKFs that used UT and GH σ-points of order 5, 7, and 10 were compared with their classical quadrature-based counterparts, namely, UKF and GHKF of order 5, 7 and 10.

We performed 100 simulations, each for $K = 500$ time steps. Root-mean-square error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{x}_k - \mathbf{m}_{k|k}^x\right)^2} \tag{39}$$

was used to measure the overall error in the state estimate $\mathbf{m}_{k|k}^x$ across all time steps. As a metric that takes into account the estimated state covariance, the Non-credibility Index (NCI) (Li and Zhao, 2006) given by

$$\text{NCI} = \frac{10}{K}\sum_{k=1}^{K}\log_{10}\frac{\left(\mathbf{x}_k - \mathbf{m}_{k|k}^x\right)^\top \mathbf{P}_{k|k}^{-1}\left(\mathbf{x}_k - \mathbf{m}_{k|k}^x\right)}{\left(\mathbf{x}_k - \mathbf{m}_{k|k}^x\right)^\top \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_k - \mathbf{m}_{k|k}^x\right)} \tag{40}$$

was used, where $\boldsymbol{\Sigma}_k$ is the mean-square-error matrix. The filter is said to be optimistic if it underestimates the actual error, which is indicated by NCI > 0. Perfectly credible filter would provide NCI = 0, that is, it would neither overestimate nor underestimate the actual error.

Table 1: The average root-mean-square error.

	BQ	Classical
UT	10.544 ± 0.048	11.081 ± 0.159
GH5	10.740 ± 0.070	10.257 ± 0.133
GH7	10.306 ± 0.053	9.855 ± 0.133
GH10	10.431 ± 0.058	9.705 ± 0.120

Tables show average values of the performance metrics across simulations with estimates of ±2 standard deviations (obtained by bootstrapping (Wasserman, 2006)). As evidenced by the results in Table 1, the BQ provides superior RMSE performance only for the case of UT σ-points. In the classical quadrature case the performance improves with increasing number of σ-points used. This trend is not observed in the BQ case. We suspect that this is due to the hyper-parameters $\boldsymbol{\theta}_f, \boldsymbol{\theta}_h$ being fixed to the same value regardless of the number of σ-points. These act effectively as a training set of the GP model and thus, it would make sense to use different values of hyper-parameters with training sets of different sizes. Figure 2 illustrates the effect of changing lengthscale on the overall performance of the BHKF with UT σ-points. The self-assessment of the filter performance

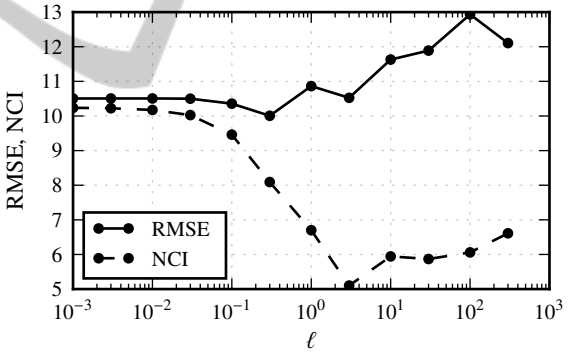

Figure 2: Sensitivity of BHKF performance to changes in the lengthscale hyperparameter ℓ. The choice $\ell = 3$ minimizes NCI at the cost of slightly higher RMSE.

Table 2: The average non-credibility index.

	BQ	Classical
UT	5.106 ± 0.010	12.071 ± 0.045
GH5	4.977 ± 0.013	10.228 ± 0.065
GH7	4.321 ± 0.009	9.256 ± 0.070
GH10	3.647 ± 0.010	8.042 ± 0.077

is less optimistic in the case of BQ, as indicated by the lower NCI in the Table 2. This indicates that the BQ based filters are more conservative in their covariance estimates. This is a consequence of including additional uncertainty (integral variance), which the classical quadrature-based filters do not utilize.

6 CONCLUSIONS

In this paper, we proposed a way of utilizing uncertainty associated with integral approximations in the nonlinear quadrature-based filtering algorithms. This was enabled by the Bayesian treatment of quadrature.

The proposed filtering algorithms were tested on a univariate benchmarking example. The results show that the filters utilizing additional uncertainty provided by the BQ show significant improvement in terms of credibility of their estimates.

Proper setting of the hyper-parameters is crucially important for achieving competitive results. The need for a principled approach for dealing with the hyper-parameters could prompt further research. Another possible research direction could be concerned with the adaptive placement of σ-points based on the posterior integral variance.

ACKNOWLEDGEMENTS

This work was supported by the Czech Science Foundation, project no. GACR P103-13-07058J.

REFERENCES

Arasaratnam, I. and Haykin, S. (2009). Cubature Kalman Filters. *IEEE Transactions on Automatic Control*, 54(6):1254–1269.

Bhar, R. (2010). *Stochastic filtering with applications in finance*. World Scientific.

Deisenroth, M. P., Huber, M. F., and Hanebeck, U. D. (2009). Analytic moment-based Gaussian process filtering. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8. ACM Press.

Deisenroth, M. P. and Ohlsson, H. (2011). A General Perspective on Gaussian Filtering and Smoothing: Explaining Current and Deriving New Algorithms. In *American Control Conference (ACC), 2011*, pages 1807–1812. IEEE.

Deisenroth, M. P., Turner, R. D., Huber, M. F., Hanebeck, U. D., and Rasmussen, C. E. (2012). Robust Filtering and Smoothing with Gaussian Processes. *IEEE Transactions on Automatic Control*, 57(7):1865–1871.

Duník, J., Straka, O., and Šimandl, M. (2013). Stochastic Integration Filter. *IEEE Transactions on Automatic Control*, 58(6):1561–1566.

Gelb, A. (1974). *Applied Optimal Estimation*. The MIT Press.

Gillijns, S., Mendoza, O., Chandrasekar, J., De Moor, B., Bernstein, D., and Ridley, A. (2006). What is the ensemble kalman filter and how well does it work? In *American Control Conference, 2006*, page 6.

Girard, A., Rasmussen, C. E., Quiñonero Candela, J., and Murray-Smith, R. (2003). Gaussian Process Priors With Uncertain Inputs Application to Multiple-Step Ahead Time Series Forecasting. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 545–552. MIT Press.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140(2):107–113.

Grewal, M. S., Weill, L. R., and Andrews, A. P. (2007). *Global Positioning Systems, Inertial Navigation, and Integration*. Wiley.

Ito, K. and Xiong, K. (2000). Gaussian Filters for Nonlinear Filtering Problems. *IEEE Transactions on Automatic Control*, 45(5):910–927.

Jiang, T., Sidiropoulos, N., and Giannakis, G. (2003). Kalman filtering for power estimation in mobile communications. *Wireless Communications, IEEE Transactions on*, 2(1):151–161.

Julier, S. J., Uhlmann, J. K., and Durrant-Whyte, H. F. (2000). A New Method for the Nonlinear Transformation of Means and Covariances in Filters and Estimators. *IEEE Transactions on Automatic Control*, 45(3):477–482.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45.

Li, X. R. and Zhao, Z. (2006). Measuring Estimator's Credibility: Noncredibility Index. In *Information Fusion, 2006 9th International Conference on*, pages 1–8.

Maybeck, P. S. (1982). *Stochastic Models, Estimation and Control: Volume 2*. Academic Press.

Minka, T. P. (2000). Deriving Quadrature Rules from Gaussian Processes. Technical report, Statistics Department, Carnegie Mellon University, Tech. Rep.

Nørgaard, M., Poulsen, N. K., and Ravn, O. (2000). New developments in state estimation for nonlinear systems. *Automatica*, 36:1627–1638.

O'Hagan, A. (1991). Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260.

Osborne, M. A., Rasmussen, C. E., Duvenaud, D. K., Garnett, R., and Roberts, S. J. (2012). Active Learning of Model Evidence Using Bayesian Quadrature. In *Advances in Neural Information Processing Systems (NIPS)*, pages 46–54.

Rasmussen, C. E. and Ghahramani, Z. (2003). Bayesian monte carlo. In S. Becker, S. T. and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 489–496. MIT Press, Cambridge, MA.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press, New York.

Särkkä, S., Hartikainen, J., Svensson, L., and Sandblom, F. (2014). Gaussian Process Quadratures in Nonlinear Sigma-Point Filtering and Smoothing. In *Informa-*

tion Fusion (FUSION), 2014 17th International Conference on, pages 1–8.

Sarmavuori, J. and Särkkä, S. (2012). Fourier-Hermite Kalman Filter. *IEEE Transactions on Automatic Control*, 57(6):1511–1515.

Šimandl, M. and Duník, J. (2009). Derivative-free estimation methods: New results and performance analysis. *Automatica*, 45(7):1749–1757.

Smith, G. L., Schmidt, S. F., and McGee, L. A. (1962). Application of statistical filter theory to the optimal estimation of position and velocity on board a circumlunar vehicle. Technical report, NASA Tech. Rep.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer-Verlag New York.