Robotics: Science and Systems 2015
Rome, Italy, July 13-17, 2015

A New Perspective and Extension of the Gaussian Filter

Manuel Wüthrich*, Sebastian Trimpe*, Daniel Kappler* and Stefan Schaal*†
*Autonomous Motion Department
Max Planck Institute for Intelligent Systems, Tübingen, Germany
Email: first.lastname@tuebingen.mpg.de
†Computational Learning and Motor Control lab
University of Southern California, Los Angeles, CA, USA

***Abstract*—The Gaussian Filter (GF) is one of the most widely used filtering algorithms; instances are the Extended Kalman Filter, the Unscented Kalman Filter and the Divided Difference Filter. GFs represent the belief of the current state by a Gaussian with the mean being an affine function of the measurement. We show that this representation can be too restrictive to accurately capture the dependences in systems with nonlinear observation models, and we investigate how the GF can be generalized to alleviate this problem. To this end, we view the GF from a variational-inference perspective. We analyse how restrictions on the form of the belief can be relaxed while maintaining simplicity and efficiency. This analysis provides a basis for generalizations of the GF. We propose one such generalization which coincides with a GF using a virtual measurement, obtained by applying a nonlinear function to the actual measurement. Numerical experiments show that the proposed Feature Gaussian Filter (FGF) can have a substantial performance advantage over the standard GF for systems with nonlinear observation models.**

I. Introduction

Decision making requires knowledge of some variables of interest. In the vast majority of real-world problems, these variables are latent, i.e. they cannot be observed directly and must be inferred from available measurements. To maintain an up-to-date belief over the latent variables, past measurements have to be fused continuously with incoming measurements. This process is called filtering and its applications range from robotics to estimating a communication signal using noisy measurements.

A. Dynamical Systems Modelling

Dynamical systems are typically modelled in a state-space representation, which means that the state is chosen such that the following two statements hold. First, the current observation depends only on the current state. Secondly, the next state of the system depends only on the current state. These assumptions can be visualized by the belief network shown in Figure 1.

We assume the system to be stationary, i.e. there is no explicit dependence on time. Therefore, the absolute time indices are irrelevant. Only the time difference within a figure or equation is of importance. To simplify notation, we will use the indices $1, 2, 3$ throughout the paper.

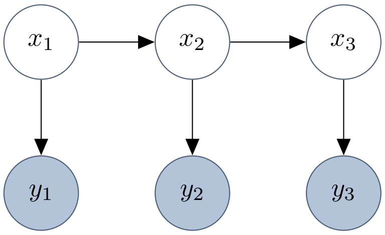

Figure 1. The belief network which characterizes the evolution of the state x and the observations y.

A stationary system can be characterized by two functions. The process model

$$x_2 = g(x_1, v_2) \tag{1}$$

describes the evolution of the state. Without loss of generality, we can assume the noise v_2 to be drawn from a Gaussian with zero mean and unit variance, since it can always be mapped onto any other distribution inside of the nonlinear function $g(\cdot)$. The observation model

$$y_2 = h(x_2, w_2) \tag{2}$$

describes how a measurement is produced from the current state. Following the same reasoning as above, we assume the noise w_2 to be Gaussian with zero mean and unit variance. The process and observation models can also be represented by distributions. The distributional form of both models are implied by their functional form

$$p(x_2|x_1) = \int_{v_2} \delta(x_2 - g(x_1, v_2))p(v_2) \tag{3}$$

$$p(y_2|x_2) = \int_{w_2} \delta(y_2 - h(x_2, w_2))p(w_2) \tag{4}$$

where δ is the Dirac delta function. While both representations contain the exact same information, sometimes one is more convenient than the other.

B. Exact Filtering

The desired posterior distribution over the current state $p(x_2|y_{:2})$ can be computed recursively from the distribution over the previous state $p(x_1|y_{:1})$; the subscript $(:t)$ denotes all time steps up to t. This recursion can be written in two

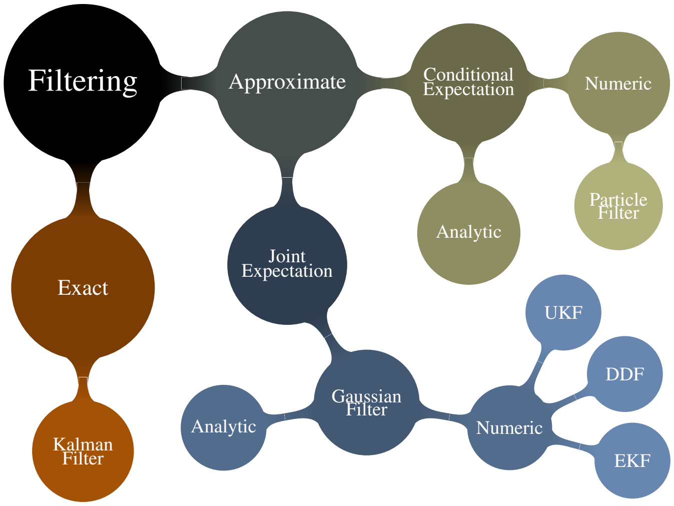

Figure 2. A taxonomy of filtering algorithms.

steps, a prediction step

$$p(x_2|y_{:1}) = \int_{x_1} p(x_2|x_1)p(x_1|y_{:1}) \tag{5}$$

and an update step

$$p(x_2|y_{:2}) = \frac{p(y_2|x_2)p(x_2|y_{:1})}{\int_{x_2} p(y_2|x_2)p(x_2|y_{:1})}. \tag{6}$$

Kalman [10] found the solution to these equations for linear process and observation models with additive Gaussian noise. However, filtering in nonlinear systems remains an important area of research. Exact solutions [2, 5] have been found for only a very restricted class of process and observation models. For more general dynamical systems, it is well known that the exact posterior distribution cannot be represented by a finite number of parameters [11]. Therefore, the need for approximations is evident.

C. Approximate Filtering

Approximate filtering methods are typically divided into deterministic, parametric methods, such as the Unscented Kalman Filter (UKF) [9] and the Extended Kalman Filter (EKF) [19], and stochastic, nonparametric methods such as the Particle Filter (PF) [7]. In this paper, we argue that there is a more fundamental division between filtering methods.

To the best of our knowledge, all existing filtering algorithms either compute expectations with respect to the conditional distribution $p(x_2|y_{:2})$ or with respect to the joint distribution $p(x_2, y_2|y_{:1})$. In Figure 2, we divide approximate filtering algorithms according this criterion. The computational power required to numerically compute expectations with respect to $p(x_2|y_{:2})$ increases exponentially with the state dimension, limiting the use of such methods to low dimensional problems. In contrast, expectations with respect to the joint distribution $p(x_2, y_2|y_{:1})$ can be approximated numerically with linear complexity in the state dimension. In Section III, we show how this fundamental difference arises.

Since conditional expectation methods suffer from the curse of dimensionality, we focus on joint expectation methods in this paper. To the best of our knowledge, all such methods approximate the true joint distribution $p(x_2, y_2|y_{:1})$ with a Gaussian distribution $q(x_2, y_2|y_{:1})$ and subsequently condition on y_2, which is easy due to the Gaussian form. This approach is called the Gaussian Filter, of which the well known EKF [19], the UKF [9] and the Divided Difference Filter (DDF) [15] are instances [21, 8].

Morelande and Garcia-Fernandez [14] show that for nonlinear dynamical systems, Gaussians can yield a poor fit to the true joint distribution $p(x_2, y_2|y_{:1})$, which in turn leads to bad filtering performance. To address this problem, we search for a more flexible representation of the belief that can accurately capture the dependences in the dynamical system, while maintaining the efficiency of the GF.

In Sections II to IV, we first review existing filtering methods, in particular the GF. Then we find some desiderata for the form of the approximate belief in Section V to provide a basis for efficient generalizations of the GF. In Section VI, we propose one possible form of the approximate belief and show that this generalization coincides with the GF using a virtual measurement given by a nonlinear function of the actual measurement. Numerical examples in Section VII highlight the potential performance gains of the proposed filter over the standard GF.

II. Approximate Prediction

We start out with the distribution $p(x_1|y_{:1})$ computed in the previous time step. The representation of the beliefs might be parametric, such as a Gaussian, or it might be nonparametric, e.g. represented by a set of samples. In any case, the goal is to find the prediction $p(x_2|y_{:1})$ given the previous belief. When there is no closed form solution to (5), we have to settle for finding certain properties of the predicted belief $p(x_2|y_{:1})$ instead of the full distribution. For all filtering algorithms we are aware of, these desired properties can be written as expectations

$$\int_{x_2} f(x_2)p(x_2|y_{:1}). \tag{7}$$

For instance with $f(x_2) = x_2$, we obtain the mean μ, and with $f(x_2) = (x_2 - \mu)(x_2 - \mu)^T$, we obtain the covariance. These expectations can then be used to find the parameters of an approximate distribution. A widely used approach is moment matching, where the moments of the approximate distribution are set to the moments of the exact distribution. We will analyse such methods in more detail below. What is important here is that we are always concerned with finding expectations of the form of (7).

We substitute (5) in (7) in order to write this expectation in terms of the last belief and the process model:

$$\int_{x_2} f(x_2)p(x_2|y_{:1}) = \int_{x_2} f(x_2) \int_{x_1} p(x_2|x_1)p(x_1|y_{:1}). \tag{8}$$

Substituting the distributional process model (3) and solving the integral over x_2, which is easy due to the Dirac distribution δ, we obtain

$$\boxed{\int\limits_{x_2} f(x_2)p(x_2|y_{:1}) = \int\limits_{x_1,v_2} f(g(x_1,v_2))p(v_2)p(x_1|y_{:1}).} \quad (9)$$

For certain process models g and functions f, it is possible to find a closed form solution. In general, however, this integral has to be computed numerically. Since $p(v_2)$ is the Gaussian noise distribution and $p(x_1|y_{:1})$ is the previous belief in the representation of choice, it is generally possible to sample from these two distributions. This is crucial since it allows for efficient numerical integration.

One possibility is to use Monte Carlo sampling to approximate the expectation from (9). The standard deviation of the estimate is proportional to $\frac{1}{\sqrt{L}}$, with L being the number of samples. The dimension of the state does not affect the standard deviation of the estimate [16].

Another possibility is to use deterministic numerical integration algorithms, such as Gaussian quadrature methods. The complexity of such methods typically scales linearly with the state dimension [21].

Which particular numeric integration method is used to compute the approximate expectations is inconsequential for the results presented in this paper. What is important is that expectations of the type required in the prediction step can be approximated efficiently, even for a high dimensional state. This is unfortunately not the case for the update step, which is the issue we are addressing in this paper.

III. Approximate Update

The goal of the update step is to obtain an approximation of the posterior $p(x_2|y_{:2})$, based on the belief $p(x_2|y_{:1})$ which has been computed in the prediction step.

A. Computation of Conditional Expectations

As for the prediction, when there is no exact solution to (6), we compute expectations with respect to the posterior $\int_{x_2} r(x_2)p(x_2|y_{:2})$, where $r(\cdot)$ is an arbitrary function. We insert (6) to express this expectation in terms of the observation model and the predicted distribution:

$$\int\limits_{x_2} r(x_2)p(x_2|y_{:2}) = \frac{\int\limits_{x_2} r(x_2)p(y_2|x_2)p(x_2|y_{:1})}{\int\limits_{x_2} p(y_2|x_2)p(x_2|y_{:1})}. \quad (10)$$

Both the numerator and the denominator can be written as

$$\boxed{\int\limits_{x_2} f(x_2)p(y_2|x_2)p(x_2|y_{:1})} \quad (11)$$

with $f(x) = r(x)$ for the numerator and $f(x) = 1$ for the denominator. The update step thus amounts to computing expectations of the form of (11).

As in the prediction step, we can approximate this expectation either by sampling, which is used in Sequential Monte Carlo (SMC) [7, 4], or by applying deterministic methods such as Gaussian quadrature [12].

There is, however, a very important difference to the prediction step. We now need to compute the expectation of a function f weighted with the observation model $p(y_2|x_2)$. If these weights are very small at most evaluation points, the numeric integration becomes inaccurate, an effect known as particle deprivation in particle filters [4].

Unfortunately, this effect becomes worse with increasing dimensionality. To see this, consider a simple example with predictive distribution $p(x_2|y_{:1}) = \mathcal{N}(x_2|0,I)$ and observation model $p(y_2|x_2) = \mathcal{N}(y_2|x_2,I)$. Both the state and measurement dimensions are equal to D. Computing the expected weight, i.e. the expected value of the likelihood, yields

$$E[p(y_2|x_2)] = \int\limits_{x_2,y_2} p(y_2|x_2)p(y_2|x_2)p(x_2|y_{:1}) = (2\sqrt{\pi})^{-D}. \quad (12)$$

That is, the expected weight decreases exponentially with the dimension D. In fact, it is well known that the computational demands of such methods increase exponentially with the state dimensionality [13, 3, 16]. Thus, methods that rely on the computation of conditional expectations are restricted to dynamical systems which either have a simple structure such that expectations can be computed analytically, or are low dimensional such that numeric methods can be used.

B. Computation of Joint Expectations

There are a number of approaches which avoid computing such expectations with respect to the conditional distribution $p(x_2|y_{:2})$. Instead, these methods express the parameters of the approximate posterior $q(x_2|y_{:2})$ as a function of expectations with respect to the joint distribution:

$$\int\limits_{x_2,y_2} f(x_2,y_2)p(y_2,x_2|y_{:1}) = \int\limits_{x_2,y_2} f(x_2,y_2)p(y_2|x_2)p(x_2|y_{:1}) \quad (13)$$

Inserting the observation model from (4) into the joint expectation above and solving the integral over y_2 yields

$$\boxed{\begin{aligned}&\int\limits_{x_2,y_2} f(x_2,y_2)p(y_2,x_2|y_{:1}) = \\ &\qquad \int\limits_{x_2,w_2} f(x_2,h(x_2,w_2))p(w_2)p(x_2|y_{:1}).\end{aligned}} \quad (14)$$

This term has the same form as the expectation in the prediction step (9). It is an integral of an arbitrary function with respect to probability densities that can be sampled. This allows us to approximate this expectation efficiently, even for high dimensional states.

C. Conclusion

The insight of this section is that computing expectations numerically with respect to the conditional distribution $p(x_2|y_{:2})$ requires exponential computational power in the state dimension, whereas the complexity of computing expectations with respect to the joint distribution $p(x_2,y_2|y_{:1})$ scales linearly with the state dimension. Note that expectations with respect to the marginals $p(x_2|y_{:1})$ and $p(y_2|y_{:1})$ are a special case of an expectation with respect to the joint distribution and can be computed efficiently as well.

In the remainder of the paper, we only consider the update step. Thus, the only variables we require are x_2 and y_2; x_1 will not be considered. Therefore, we drop the indices for ease of notation. Furthermore, we make the dependence on $y_{:1}$ implicit. That is, $p(x_2, y_2|y_{:1})$ becomes $p(x, y)$ and $p(x_2|y_{:2})$ becomes $p(x|y)$, etc.

IV. The Gaussian Filter

The advantage in terms of computational complexity of joint expectation filters over conditional expectation filters comes at a price: The approximate posterior $q(x|y)$ must have a functional form such that its parameters can be computed efficiently from these joint expectations. To the best of our knowledge, all existing joint expectation filters solve this issue by approximating the true joint distribution $p(x, y)$ with a Gaussian distribution:

$$q(x, y) = \mathcal{N}\left(\begin{pmatrix} x \\ y \end{pmatrix} \middle| \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}\right). \tag{15}$$

The parameters of this approximation are readily obtained by moment matching, i.e. the moments of the Gaussian are set to the moments of the exact distribution:

$$\boxed{\begin{aligned} \mu_x &= \int_x x p(x) \\ \mu_y &= \int_y y p(y) \\ \Sigma_{xx} &= \int_x (x-\mu_x)(x-\mu_x)^T p(x) \\ \Sigma_{yy} &= \int_y (y-\mu_y)(y-\mu_y)^T p(y) \\ \Sigma_{xy} &= \int_{x,y} (x-\mu_x)(y-\mu_y)^T p(x,y). \end{aligned}} \tag{16}$$

All of these expectations can be computed efficiently for reasons explained in the previous section.

After the moment matching step, we condition on y to obtain the desired posterior, which is a simple operation since the approximation is Gaussian:

$$q(x|y) = \mathcal{N}(x|\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}^T). \tag{17}$$

This approach is called the Gaussian Filter (GF) [8, 18, 21]. Widely used filters such as the EKF [19], the UKF [9] and the DDF [15] are instances of the Gaussian Filter, differing only in the numeric integration method used for computing the expectations in (16).

While much effort has been devoted to finding accurate numeric integration schemes for computing these expectations, there seems to be no joint expectation method using a non-Gaussian joint approximation $q(x, y)$. The posterior (17), which we ultimately care about, is therefore Gaussian in the state x with the mean being an affine function of y. As we show in the experimental section, this form can be too restrictive to accurately capture the relationship between the measurement and the state in nonlinear settings. This leads to information about the state being discarded and ultimately to poor filtering performance.

V. Generalization of the Gaussian Filter

In this section, we investigate whether it is possible to find a more general form of the approximate posterior $q(x|y)$ that still allows for efficient computation of the parameters. To this end, we write the problem of finding the parameters of the approximation as an optimization problem.

In the GF, the parameters Θ of the Gaussian belief $q(x, y|\Theta)$ are found by moment matching. For a Gaussian approximation, moment matching is equivalent to minimizing the KL-divergence [1]

$$\mathrm{KL}[p(x,y)|q(x,y|\Theta)] = \int_{x,y} \log\left(\frac{p(x,y)}{q(x,y|\Theta)}\right) p(x,y). \tag{18}$$

By minimizing (18) with respect to Θ, we can thus retrieve the GF. Furthermore, the KL-divergence has convenient analytic properties. It is a widely used objective for matching distributions and can be justified from an information theoretic point of view [1].

Having found an appropriate objective for the approximation, it is natural to ask if it is possible to find more general, non-Gaussian approximations. The form of $q(x, y|\Theta)$ is restricted by the requirement of being able to condition on y in closed form in order to find the approximate conditional $q(x|y, \Theta)$. This requirement is met automatically if we choose the form of the conditional distribution and the marginal distribution separately, instead of picking a form for the joint distribution. The joint distribution is then given by $q(x, y|\Theta) = q(x|y, \theta)q(y|\vartheta)$, where we have split the parameter set Θ into θ and ϑ. Any conditional and marginal distributions can be combined to form a valid joint distribution. Hence, the respective parameter sets θ and ϑ can be chosen independently. Imposing any constraints tying the two parameter sets together would restrict the flexibility of the joint distribution unnecessarily.

Inserting this factorization into (18), we obtain

$$\mathrm{KL}[p(x,y)|q(x,y|\Theta)] = c(\vartheta) + \mathrm{KL}[p(x,y)|q(x|y,\theta)] \tag{19}$$

where we have collected all terms independent of θ in $c(\vartheta)$. Since only the conditional distribution is of interest, we will maximize with respect to θ. Hence, we can drop the terms which do not depend on θ, which leads to the objective function

$$\boxed{\mathrm{KL}[p(x,y)|q(x|y,\theta)] = \int_{y,x} \log\left(\frac{p(x,y)}{q(x|y,\theta)}\right) p(x,y).} \tag{20}$$

Note that this is a somewhat unusual KL-divergence, since it compares a joint distribution with a conditional distribution. However, this configuration is very desirable in this context. We can directly obtain the approximate posterior distribution $q(x|y, \theta)$ from the exact joint distribution $p(x, y)$ by minimizing (20) with respect to θ. Only expectations with respect to the joint distribution $p(x, y)$ are required, and we have seen that these can be approximated efficiently.

A. *Desiderata for the Form of the Approximation*

In the following, we seek conditions on the form of $q(x|y, \theta)$ that allow for an efficient minimization of (20) with respect to θ.

First, $q(x|y, \theta)$ has to integrate to one in x since it is a probability distribution. We can enforce this condition by writing

$$q(x|y, \theta) = \frac{r(x, y, \theta)}{\int_x r(x, y, \theta)} \tag{21}$$

with $r(x, y, \theta)$ being any positive function whose integral in x over the real domain is finite and non-zero.

Furthermore, for the objective in (20) to be well defined, the support of $q(x|y, \theta)$ has to contain the support of $p(x, y)$. Since $p(x, y)$ could be any distribution, we will choose the form $q(x|y, \theta)$ such that it has infinite support; that is, $q(x|y, \theta) > 0$ everywhere, which implies $r(x, y, \theta) > 0$. This condition is enforced by writing the approximate distribution as

$$\boxed{q(x|y, \theta) = \frac{e^{f(x,y,\theta)}}{\int_x e^{f(x,y,\theta)}}} \tag{22}$$

with $f(x, y, \theta) = \log(r(x, y, \theta))$. The question we will address in the following is what f has to look like in order to obtain an efficient filtering algorithm.

Substituting $q(x|y, \theta)$ in (20), we obtain

$$\boxed{\begin{aligned} &\mathrm{KL}[p(x, y)|q(x|y, \theta)] = C + \\ &\int\limits_y \log\left(\int\limits_x e^{f(x,y,\theta)}\right) p(y) - \int\limits_{y,x} f(x, y, \theta)p(x, y) \end{aligned}} \tag{23}$$

where we have collected the terms which do not depend on θ in C. By setting the derivative with respect to θ to zero, we obtain a criterion for stationarity

$$\boxed{\int\limits_y\left(\int\limits_x \frac{\partial f(x, y, \theta)}{\partial \theta} q(x|y, \theta)\right) p(y) = \int\limits_{y,x} \frac{\partial f(x, y, \theta)}{\partial \theta} p(x, y).} \tag{24}$$

If we choose $f(\cdot)$ such that the objective (23) is convex in θ, then (24) is a sufficient condition for optimality.

Before this system of equations can be solved, all the integrals have to be computed. The integral over x on the left-hand side of (24) is an expectation with respect to the parametric approximation. Since the integrand depends on unknown parameters, this inner integral cannot be approximated numerically. Therefore, f has to be chosen such that there is a closed form solution.

In general, the outer integral over y cannot be solved in closed form since $p(y)$ can have a very complex form, depending on the dynamical system. However, expectations with respect to $p(y)$ can be efficiently approximated numerically, as discussed above. Numeric integration is possible only if the integrand depends on no other variable than the ones we integrate out. Therefore, we require f to be such that, after analytically solving the inner integral over x, all the dependences on θ can be moved outside of the integral over y.

On the right-hand side of (24), we evaluate an expectation with respect to $p(x, y)$. Again, it is not possible for general dynamical systems to find a closed form solution, but numerical expectations with respect to $p(x, y)$ can be computed efficiently. To allow for numerical integration, f must be such that all the dependences on θ can be moved outside of the integral over x and y.

Finally, after computing the integrals, we have to solve the system of equations (24) in order to find the optimal θ. Therefore, $f(\cdot)$ should be such that this solution can be found efficiently.

It is not clear how the most general $q(x|y, \theta)$ complying with the above desiderata can be found. Nevertheless, this discussion can guide the search for more general belief representations than the affine Gaussian, which leave the efficiency of the GF intact. The following section provides an example.

VI. The Feature Gaussian Filter

We propose to generalize the affine Gaussian approximate posterior of the GF by allowing for nonlinear features $\phi(y)$ of the measurement. More formally, we choose f in (22) as

$$f(x, y, \Gamma, \Sigma) = -\frac{1}{2}(x - \Gamma\phi(y))^T \Sigma^{-1} (x - \Gamma\phi(y)) \tag{25}$$

with parameters $\theta = (\Gamma, \Sigma)$ and ϕ an arbitrary feature function. This leads to an approximate distribution (22), which is Gaussian in x but can have nonlinear dependences on y,

$$q(x|y, \Gamma, \Sigma) = \mathcal{N}(x|\Gamma\phi(y), \Sigma). \tag{26}$$

In the following, we show that because this approximation complies with the desiderata from the previous section, the parameters can be optimized efficiently. We refer to the resulting filtering algorithm as the Feature Gaussian Filter (FGF). Finally, we show that the FGF is essentially equivalent to the standard GF using a virtual measurement, obtained by mapping the actual measurement through a nonlinear function.

A. *Finding* Γ

The derivative with respect to Γ is

$$\frac{\partial f(x, y, \Gamma, \Sigma)}{\partial \Gamma} = \Sigma^{-1}(x - \Gamma\phi(y))\phi(y)^T \tag{27}$$

and the corresponding analytic integral can readily be solved since the approximate distribution is Gaussian in x:

$$\int_x \frac{\partial f(x,y,\Gamma,\Sigma)}{\partial \Gamma} q(x|y,\Gamma,\Sigma) = 0. \tag{28}$$

Inserting these results into (24), we can solve for Γ

$$\boxed{\Gamma = E[x\phi(y)^T]E[\phi(y)\phi(y)^T]^{-1}.} \tag{29}$$

B. Finding Σ

The matrix Σ is constrained to be positive definite, such that the approximate distribution (26) is Gaussian. As it turns out, the unconstrained optimization yields a positive definite matrix. Thus, there is no need to take this constraint into account explicitly.

The derivative with respect to Σ^{-1} is

$$\frac{\partial f(x,y,\Gamma,\Sigma)}{\partial \Sigma^{-1}} = -\frac{1}{2}(x - \Gamma\phi(y))(x - \Gamma\phi(y))^T \tag{30}$$

and the corresponding analytic integral in x is

$$\int_x \frac{\partial f(x,y,\Gamma,\Sigma)}{\partial \Sigma^{-1}} q(x|y,\Gamma,\Sigma) = -\frac{1}{2}\Sigma. \tag{31}$$

Inserting these results into (24), we can solve for Σ

$$\boxed{\Sigma = E[(x - \Gamma\phi(y))(x - \Gamma\phi(y))^T].} \tag{32}$$

C. Connection to the Gaussian Filter

In the following we show that for a feature $\phi(y) = (c, \varphi(y)^T)^T$, which contains a constant $c \neq 0$ and an arbitrary sub-feature φ, the FGF is equivalent to the GF using $\hat{y} = \varphi(y)$ as the measurement. Inserting $\phi(y) = (c, \hat{y}^T)^T$ into (29), we obtain

$$\Gamma = \begin{pmatrix} \frac{\mu_x - \Sigma_{x\hat{y}}\Sigma_{\hat{y}\hat{y}}^{-1}\mu_{\hat{y}}}{c} & \Sigma_{x\hat{y}}\Sigma_{\hat{y}\hat{y}}^{-1} \end{pmatrix} \tag{33}$$

with the parameters $\mu_{(\cdot)}$ and $\Sigma_{(\cdot)}$ as defined in (16). The mean of the approximate posterior is

$$\Gamma\phi(y) = \mu_x + \Sigma_{x\hat{y}}\Sigma_{\hat{y}\hat{y}}^{-1}(\hat{y} - \mu_{\hat{y}}). \tag{34}$$

Inserting this result into (32), we obtain the covariance

$$\Sigma = \Sigma_{xx} - \Sigma_{x\hat{y}}\Sigma_{\hat{y}\hat{y}}^{-1}\Sigma_{x\hat{y}}^T. \tag{35}$$

Clearly, these equations correspond to the GF equations (17). This means that, if the feature vector $\phi(y)$ contains a constant, the FGF is equivalent to the GF using the virtual measurement $\hat{y} = \varphi(y)$ instead of y. In particular, with a feature $\phi(y) = (1, y^T)^T$, we retrieve the standard GF.

Applying nonlinear transformations to the physical sensor measurements before feeding them into a GF is not uncommon in robotics and other applications (see [6, 20, 17] for example). The formal analysis herein provides insight into the effect of such nonlinear transformations and reveals why they are beneficial. Namely, they allow for a better fit of the conditional distribution. While these transformations are often motivated from physical insight or introduced heuristically, we provide a different interpretation of ϕ as a means of improving the fit of the posterior by allowing for more expressive nonlinear features. This shall be highlighted in the examples in Section VII, where we use monomials of increasing order as generic features.

D. Feature Selection

The above analysis shows that adding nonlinear features gives the approximate distribution more flexibility to fit the exact distribution. Overfitting is not possible since we are minimizing the KL-divergence to the exact distribution. It therefore makes sense to use as many features as the computational speed requirements allow.

Ideally, one would choose a feature which maps the measurement to a representation which relates to the state linearly. If this is not possible, then generic features such as monomials can be used.

E. Computational Complexity

The only cause of a difference in computational complexity between the standard GF and the FGF is the difference in the dimension of the measurement y and the feature $\phi(y)$. This means that the feature dimension has to be chosen such that the required computational speed is attained. The feature dimension can even be lower than the dimension of the actual measurement if the standard GF is too slow.

VII. Analysis and Simulation of the Feature Gaussian Filter

As the previous analysis suggests, it is beneficial to augment the measurement with nonlinear features since this gives the approximation more flexibility to fit the exact distribution, i.e. to achieve a lower KL-divergence (23). In this section, we illustrate this effect in more detail for two dynamical systems.

A. Estimation of Sensor Noise Magnitude

The measurement process (2) of a dynamical system can often be represented by a nonlinear observation model with additive noise

$$h(x, M, w) = \tilde{h}(x) + Mw \tag{36}$$

where $\tilde{h}$ is a nonlinear function of the system state, and the matrix M determines the magnitude of the sensor noise (recall that w is Gaussian with zero mean and unit variance). Often, the sensor accuracy (i.e. the matrix M) is not precisely known, or it may be time varying due to changing sensor properties and environmental conditions. It is then desirable to estimate the noise matrix M alongside the state x. In the following, we show that this is not possible with the standard GF, but can be achieved with the FGF.

We define an augmented state $\hat{x} := (x; m)$, where m is a column vector containing all the elements of the noise matrix M. The observation model in distributional form is $p(y|\hat{x}) = p(y|x,m) = \mathcal{N}(y|\tilde{h}(x), MM^T)$. The state x and the parameters m stem from independent processes, and we therefore have $p(\hat{x}) = p(x)p(m)$. Let us now apply the standard GF to this problem by computing the parameters in

(16). In particular, we compute the covariance between the augmented state and the measurement

$$\Sigma_{\hat{x}y} = \int_{x,m,y} \begin{pmatrix} x - \mu_x \\ m - \mu_m \end{pmatrix} (y - \mu_y)^T p(y|x,m)p(x)p(m). \quad (37)$$

The integral over y can be solved easily since $p(y|x,m)$ is Gaussian,

$$\Sigma_{\hat{x}y} = \int_{x,m} \begin{pmatrix} x - \mu_x \\ m - \mu_m \end{pmatrix} (\tilde{h}(x) - \mu_y)^T p(x)p(m). \quad (38)$$

Interestingly, the second factor does not depend on m. Therefore, the integral over m is solved easily and yields

$$\Sigma_{\hat{x}y} = \int_x \begin{pmatrix} x - \mu_x \\ \mu_m - \mu_m \end{pmatrix} (\tilde{h}(x) - \mu_y)^T p(x) = \begin{pmatrix} \Sigma_{xy} \\ 0 \end{pmatrix}. \quad (39)$$

As a result, there is no linear correlation between the measurement y and the parameters m. Inserting this result into (17) shows that the innovation corresponding to m is zero. The corresponding part of the covariance matrix does not change either. The measurement has hence no effect on the estimate of m. It will behave as if no observation had been made. This illustrates the failure of the GF to capture certain dependences in nonlinear dynamical systems.

In contrast, if a nonlinear feature in the measurement y is used, the integral over y in (37) will not yield $\tilde{h}(x)$, but instead some function depending on both x and m. This dependence allows the FGF to infer the desired parameters.

Numerical example: For the purpose of illustrating the theoretical argument above, we use a small toy example. We consider a single sensor, where all quantities in (36), including the standard deviation M, are scalars. Since we are only interested in the estimate of M, we choose $\tilde{h}(x) = 0$. The observation model (36) simplifies to

$$h(M_2, w_2) = M_2 w_2. \quad (40)$$

Note that we have reintroduced time indices. Picking a simple process model and an initial distribution

$$g(M_1, v_2) = M_1 + 0.1 v_2 \quad (41)$$
$$p(M_1) = \mathcal{N}(M_1|5, 1) \quad (42)$$

the dynamical system (1), (2) is fully defined. This example captures the fundamental properties of the FGF as pertaining to the estimation of sensor noise intensity M. The same qualitative effects hold for multivariate systems (36) for the reasons stated above.

In Figure 3, we plot the exact conditional distribution $p(M_2|y_2)$ implied by this system in grayscale. This distribution was computed numerically for the purpose of comparison. It would, of course, be too expensive to use in a filtering algorithm. The overlaid orange contour lines show the approximate conditional distribution $q(M_2|y_2)$ obtained with the standard GF. No matter what measurement y_2 is obtained, the posterior $q(M_2|y_2)$ is the same. The GF does not react to the measurements at all.

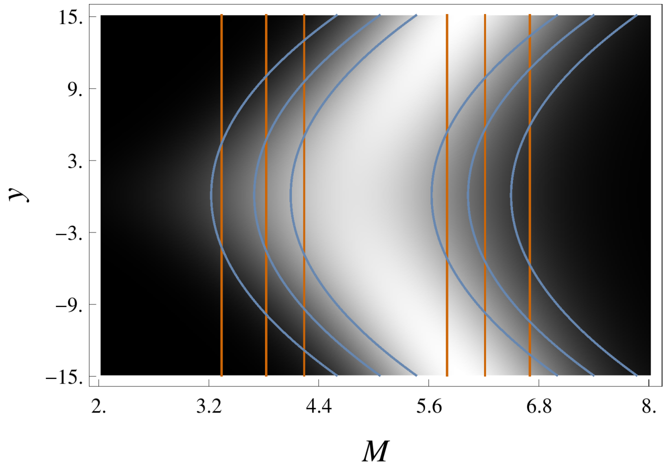

Figure 3. Estimation of sensor noise magnitude: Density plot of the true conditional distribution $p(M_2|y_2)$ with overlaid contour lines of the approximate conditional distribution $q(M_2|y_2)$ of the GF in orange and of the FGF in blue.

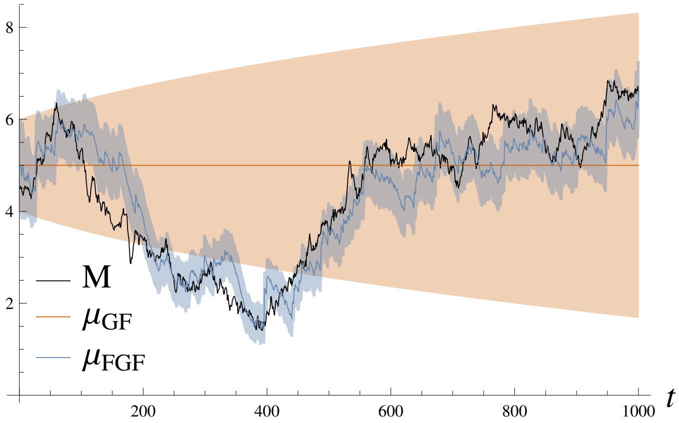

Figure 4. Estimation of sensor noise magnitude: The simulated noise parameter M is shown in black, together with the mean and standard deviation of the estimates obtained with the GF (orange) and the FGF (blue).

The true conditional distribution $p(M_2|y_2)$ depends on y_2, which means that the measurement does in fact contain information about the state M_2. However, the approximation $q(M_2|y_2)$ made by the GF is not expressive enough to capture this information, which results in a very poor fit to $p(M_2|y_2)$.

The standard GF is the special case of the FGF with the feature $\phi(y) = (1, y)^T$. Let us take the obvious next step and add a quadratic term to the feature $\phi(y) = (1, y, y^2)^T$. The resulting approximation is represented by the blue contour lines in Figure 3. Clearly, $q(x_2|y_2)$ now depends on the measurement y_2, which allows the FGF to exploit the information about the state x_2 contained in the measurement. The approximation $q(x_2|y_2)$ of the FGF has a more flexible form, which allows for a better fit of the true posterior.

To analyse actual filtering performance, we simulate the dynamical system and the two filters for 1000 time steps. The results are shown in Figure 4. As expected, the standard GF does not react in any way to the incoming measurements. The FGF, on the other hand, is capable of inferring the state M from the measurement y, as suggested by the theoretical

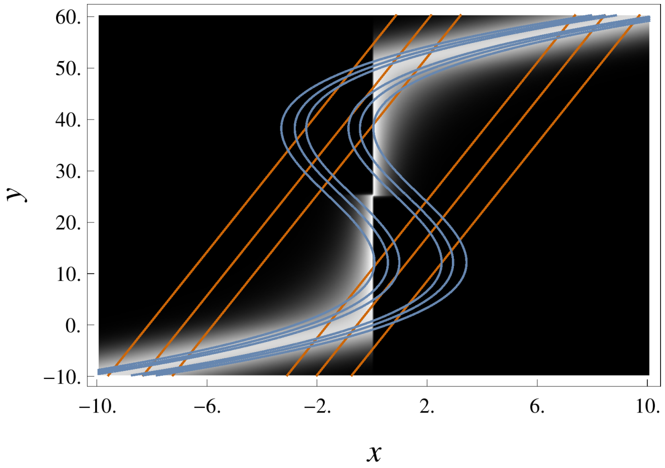

Figure 5. Nonlinear observation model: Density plot of the true conditional distribution $p(x_2|y_2)$ with overlaid contour lines of the approximate conditional distribution $q(x_2|y_2)$ of the GF in orange and of the FGF in blue.

analysis above.

B. Nonlinear Observation Model

In this section, we investigate how the theoretical benefit of adding nonlinear features translates into improved filtering performance for systems with nonlinear observation models. To clearly illustrate the difference of GF and FGF, we choose a simple system with a strong nonlinearity (step function). Given the theoretical analysis herein, it is to be expected that the insights gained from this artificial example extend to more realistic nonlinear problems in robotics and other applications.

The process model, the observation model, and the initial state distribution are given by

$$g(x_1, v_2) = x_1 + v_2 \tag{43}$$

$$h(x_2, w_2) = x_2 + w_2 + 50H(x_2) \tag{44}$$

$$p(x_1) = \mathcal{N}(x_1|0, 5) \tag{45}$$

where $H(\cdot)$ is the Heaviside step function.

In Figure 5, we plot the true conditional density $p(x_2|y_2)$ with overlaid orange contour lines of the approximate conditional distribution $q(x_2|y_2)$ obtained using the standard GF. The contour lines reflect the estimator structure of the GF in (17). The mean of the approximate density $q(x_2|y_2)$ is an affine function of the measurement y_2. For nonlinear observation models, this coarse approximation can lead to loss of valuable information contained in the measurement y_2.

The approximate density $q(x_2|y_2)$ obtained using a feature $\phi(y) = (1, y, y^2, y^3)^T$, which is represented by the blue contour lines in Figure 5, fits the true posterior much better. This illustrates that nonlinear features allow for approximate posteriors with much more elaborate dependences on y.

Figure 6 shows how this difference translates to filtering performance. When x is far away from zero, the nonlinearity has no effect: the system behaves like a linear system. The density plot in this regime would be centered at a linear part of the distribution, and both filters would achieve a perfect fit. Both the standard GF and the FGF are therefore optimal in that case. When the state is close to zero, however, the advantage of the FGF becomes apparent. Its tracking performance is good even when the state is close to the nonlinearity of the observation model, due to more flexibility in y_2 of the posterior approximation $q(x_2|y_2)$.

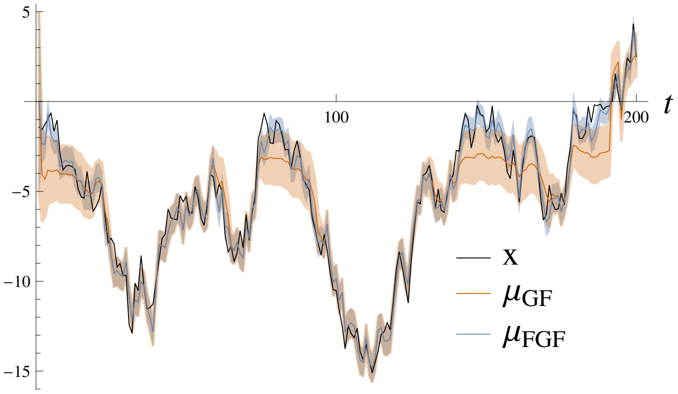

Figure 6. Nonlinear observation model: We plot the simulated state x (black) and the means and standard deviations of the estimates obtained with the GF (orange) and the FGF (blue).

VIII. Conclusion

We showed that the GF can be understood as an optimal approximation to the exact distribution, subject to the constraint that the form of the belief $q(x|y)$ be Gaussian in x and affine in y. Theoretical analysis and simulations showed that this form can be too restrictive to accurately represent the belief in nonlinear systems. We discussed how this constraint can be relaxed while maintaining the efficiency of the GF. This analysis served as a basis for potential generalizations of the GF.

We proposed one such generalization, the Feature Gaussian Filter (FGF). The name is motivated by the fact that the FGF is equivalent to a GF that uses a virtual measurement, or feature, which is obtained by applying a nonlinear function to the actual measurement. We showed both theoretically and in simulation that using nonlinear features can significantly improve the performance of the GF. For instance, the practically relevant problem of estimating the sensor noise magnitude alongside the state cannot be tackled by the standard GF because the expressive power of its belief is too limited. We showed that this issue can be resolved by the FGF.

The results obtained in the simulation examples herein are promising and suggest that the FGF may yield superior filtering performance for nonlinear problems in robotics and other applications. Analysing the performance of the FGF in a more realistic, high dimensional scenario remains future work.

Whether it is possible to find an approximate posterior of a more general form than in the FGF, while complying with the requirements derived in Section V, is another open question.

References

[1] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[2] V.E. Beneš. Exact finite-dimensional filters for certain diffusions with nonlinear drift. *Stochastics: An International Journal of Probability and Stochastic Processes*, 1981.

[3] P. Bickel, B. Li, and T. Bengtsson. Sharp failure rates for the bootstrap particle filter in high dimensions. In *IMS Collections: Pushing the Limits of Contemporary Statistics*. 2008.

[4] O. Cappe, S.J. Godsill, and E. Moulines. An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, 2007.

[5] F.E. Daum. Exact finite-dimensional nonlinear filters. *IEEE Transactions on Automatic Control*, 1986.

[6] F.E. Daum and R.J. Fitzgerald. Decoupled Kalman filters for phased array radar tracking. *IEEE Transactions on Automatic Control*, 1983.

[7] N.J. Gordon, D.J. Salmond, and A.F.M. Smith. Novel approach to nonlinear/non-Gaussian bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 1993.

[8] K. Ito and K. Xiong. Gaussian filters for nonlinear filtering problems. *IEEE Transactions on Automatic Control*, 2000.

[9] S.J. Julier and J.K. Uhlmann. A new extension of the Kalman filter to nonlinear systems. In *Proceedings of AeroSense: The 11th Int. Symp. on Aerospace/Defense Sensing, Simulations and Controls*, 1997.

[10] R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 1960.

[11] H.J. Kushner. Approximations to optimal nonlinear filters. *IEEE Transactions on Automatic Control*, 1967.

[12] H.J. Kushner and A.S. Budhiraja. A nonlinear filtering algorithm based on an approximation of the conditional distribution. *IEEE Transaction on Automatic Control*, 2000.

[13] B. Li, T. Bengtsson, and P. Bickel. Curse-of-dimensionality revisited: Collapse of importance sampling in very large scale systems. Technical report, Department of Statistics, UC-Berkeley, 2005.

[14] M.R. Morelande and A.F. Garcia-Fernandez. Analysis of Kalman filter approximations for nonlinear measurements. *IEEE Transactions on Signal Processing*, 2013.

[15] M. Nørgaard, N.K. Poulsen, and O. Ravn. New developments in state estimation for nonlinear systems. *Automatica*, 2000.

[16] A.B. Owen. Monte carlo theory, methods and examples (book draft), 2013. URL http://statweb.stanford.edu/~owen/mc/.

[17] N. Rotella, M. Bloesch, L. Righetti, and S. Schaal. State estimation for a humanoid robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014.

[18] S. Särkkä. *Bayesian filtering and smoothing*. Cambridge University Press, 2013.

[19] H.W. Sorenson. *Kalman Filtering: Theory and Application*. IEEE Press selected reprint series. IEEE Press, 1960.

[20] J. Vaganay, M.J. Aldon, and A. Fournier. Mobile robot attitude estimation by fusion of inertial data. In *IEEE International Conference on Robotics and Automation*, 1993.

[21] Y. Wu, D. Hu, M. Wu, and X. Hu. A numerical-integration perspective on Gaussian filters. *IEEE Transactions on Signal Processing*, 2006.