Progressive Bayesian Particle Flows Based on Optimal Transport Map Sequences

Uwe D. Hanebeck

Intelligent Sensor-Actuator-Systems Laboratory (ISAS)
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology (KIT), Germany
Uwe.Hanebeck@kit.edu

***Abstract*—We propose a method for optimal Bayesian filtering with deterministic particles. In order to avoid particle degeneration, the filter step is not performed at once. Instead, the particles progressively flow from prior to posterior. This is achieved by splitting the filter step into a series of sub-steps. In each sub-step, optimal resampling is done by a map that replaces non-equally weighted particles with equally weighted ones. Inversions of the maps or monotonicity constraints are not required, greatly simplifying the procedure. The parameters of the mapping network are optimized w.r.t. a particle set distance. This distance is differentiable, and compares non-equally and equally weighted particles. Composition of the sequence of maps provides a final mapping from prior to posterior particles. Radial basis function neural networks are used as maps. It is important that no intermediate continuous density representation is required. The entire flow works directly with particle representations. This avoids costly density estimation.**

I. Introduction

We consider Bayesian filtering based on particle representations of probability density functions (PDFs). Particle filters are a very active field of research. Thousands of papers have been devoted to this topic. Most of them focus on large numbers of random particles and asymptotically valid methods. Necessarily finite numbers of particles lead to a variety of problems. This includes (i) particle degeneracy and (ii) non-reproducibility due to the randomness of the particles.

Standard approaches for coping with degeneracy employ *resampling*. By doing so, the weighted samples are replaced by a set of unweighted samples in order to reduce sample variance. However, standard resampling procedures simply delete small samples and replicate large ones. Their locations remain unchanged until the next prediction step. This is clearly suboptimal and does not allow for several consecutive filtering steps without intermediate prediction steps. In addition, proposal densities and importance sampling are used to alleviate degeneration effects. This allows to consider the current measurement during prediction and places the particles closer to the likelihood. However, this requires additional engineering and increases computational complexity.

We focus on methods that employ different forms of Bayesian filter updates. They replace the direct update by an indirect one that inherently avoids degeneration by using transport maps or flows.

II. State of the Art

The state of the art in calculating transport maps or flows will be investigated. The focus is on methods for converting the Bayesian filter step. However, some methods are only derived for the case of mapping between two given densities without explicitly considering the Bayes update.

A. Continuous Density Flow Filters

We start with flows of continuous densities as they have been developed earlier than particle flows. The first progressive Bayesian filter [1] employs a continuous density for representing the posterior. Progressively introducing the likelihood function leads to a homotopy continuation approach. A Gaussian mixture flow is derived that minimizes the squared integral deviation from the true posterior. This results in a system of explicit ordinary first-order differential equations solved over an artificial time from 0 to 1. The approach is generalized in [2] for the Kullback-Leibler divergence and the squared Hellinger distance.

B. Filters based on Transform Maps

Transformation of a random vector $\underline{\boldsymbol{x}}$ via a known nonlinear map $\underline{\boldsymbol{y}} = \underline{g}(\underline{\boldsymbol{x}})$ is a classic problem. The goal is to calculate the output density $f_y(\underline{y})$ given the input density $f_x(\underline{x})$. For that purpose, a different map is required that maps densities to densities, i.e., $f_y = G(f_x)$. It can be derived from the original mapping $\underline{g}(\cdot)$, its roots, and its Jacobian. The derivation is simplified for monotonic maps.

Here, we are faced with a more complex problem. Given two densities $f_x(\underline{x})$ and $f_y(\underline{y})$, we *want to find the map* $\underline{g}(\cdot)$ between their corresponding random vectors $\underline{\boldsymbol{x}}$ and $\underline{\boldsymbol{y}}$. When the two densities are continuous, a mapping exists, is unique, and monotonic [3]. Finding the map is challenging, especially under monotonicity constraints. For univariate $f_x(\cdot)$ and $f_y(\cdot)$, the map can be composed from the input cumulative distribution function (CDF) $F_x(\cdot)$ and the output quantile function $F_y^{-1}(\cdot)$ as

$$g(x) = F_y^{-1}\left(F_x(x)\right) \tag{1}$$

[4, p. 102, eq. 5-20]. However, calculation of cumulative distributions and quantile functions can be challenging with analytic expression only available in special cases. This is exacerbated in multivariate settings.

It gets even more challenging, when only samples of the two densities $f_x(\underline{y})$ and $f_y(\underline{y})$ are given. In that case, standard distance measures (such as the KL divergence) between densities cannot be employed. The most complicated case is the Bayes update, when only samples are given for $f_x(\underline{y})$. Then $f_y(\underline{y})$ is only given implicitly as the product of the samples and the likelihood and we can neither obtain values of the posterior density nor can we sample from it.

Filters based on Knothe-Rosenblatt Triangular Maps: In [5], a continuous prior density is assumed to be given from which samples can easily be drawn. In addition, a given likelihood is assumed that can be evaluated up to a constant. A map is characterized by the minimum of either the Hellinger metric or the KL divergence between prior and posterior, see [5, pp. 7818–7819]. The map is computed by solving a nonlinear, nonconvex minimization problem, where the distance is evaluated by Monte Carlo simulations from the prior density. Knothe-Rosenblatt rearrangements [6] are used as these are triangular and monotonic and thus easy to invert[1]. It is important to note that these maps are not optimal with respect to a transportation distance. To reduce complexity, decomposability of transport maps is investigated in [8], leading to sparse triangular maps. The construction of maps is generalized in [9], where two types of maps are considered. The first map[2] (direct transport) transforms a reference measure to a target measure [9, p. 6]. The reference density is known, the target density can be evaluated but is unnormalized. The second map (inverse transport) transforms the target measure to the given reference measure [9, p. 13]. This is useful when the target density is unknown and only samples of it are given. Finding the map is equivalent to maximum likelihood estimation and can be solved via convex optimization. These two maps are used in [10] to derive a nonlinear ensemble filter [10, p. 20] for the case when only samples of the prior density are available: An inverse map is constructed for transforming prior samples to a convenient reference measure. A direct map is used to transform the reference measure to the posterior.

Filters based on Normalizing Flows: Popularized in the machine learning community, normalizing flows are used in variational inference problems [11] to perform density estimation in order to model complex data distributions. A normalizing flow is a sequence of invertible transformations mapping a reference measure to a set of samples from a desired target density [12, p. 3]. The flows correspond to the inverse transport in [9] and are determined by maximum likelihood estimation.

C. Particle Flow Filters: Continuous Derivation

The first breed of particle flow filters derive a PDE or ODE based on a suitably parametrized posterior while assuming continuity of the involved densities. In a second step, the required discretizations are performed.

Daum-Huang (DH) Particle Flows: Particle flows are derived in [13] from a log-homotopy relating prior and posterior. The flow is represented by a partial differential equation (Fokker-Planck) assuming continuous densities. Hence, estimating the required gradients from the particles is a challenging problem, see [14]. Several versions of DH flows have been proposed [15], many for coping with stiffness in the flow. However, they usually rely on some sort of nonlinear Kalman filter running in parallel that compromises performance [16].

Particle Flows based on Liouville Equation: In [17], a homotopy continuation approach similar to [1] is used. Assuming an ODE for moving particles from prior to posterior, the corresponding Liouville PDE is derived[3]. (A similar approach is pursued in [18, p. 5].) The desired ODE velocity field is then obtained as the solution of the Liouville PDE. Tractable solutions are obtained in the univariate case. The multivariate case requires a Gibbs approximation. This requires the full conditional distributions, which are numerically approximated. Finally, the mapped particles are "just" used as proposal distributions inside of sequential Monte Carlo samplers.

Particle Flows with Repulsion Kernels: It would be possible to employ density estimation to find an intermediate continuous representation to calculate gradients. However, density estimation usually is not differentiable. In [19], repulsion kernels [20] are used that represent the spread of probability mass around the particles. This leads to an ODE for the particle locations over an artificial time from 0 to 1.

D. Particle Flow Filters: Direct Discrete Derivation

The second breed of particle flow filters acknowledges that a computer implementation requires discretization anyway and directly derives a sequence of discrete updates.

The Bayesian particle flow proposed in this paper belongs to this category. It follows the procedure in [21]: The likelihood is adaptively split into several sub-likelihoods, each of which is easier to process. The sub-likelihoods are used to sequentially update prior particles. After each update, the weighted particles are optimally resampled with equally weighted ones. Resampling is performed by minimizing a suitable particle set distance [22] and [23]. As a result, the prior particles are moved to regions with high posterior density. The resampling step in this paper differs from the one in [21]: Transport maps are used for mapping weighted samples to unweighted ones instead of using direct optimization. This is advantageous as the number of map parameters is typically smaller than the number of parameters of the samples. In addition, the transport map allows to more easily incorporate smoothness.

III. Problem Formulation

We consider a dynamic system with a state $\underline{x} \in \mathbb{R}^N$ and state dimension N. A transition density describing the state evolution provides a forecast in the form of a prior PDF.

Given a continuous prior density $f_p^t(\underline{x})$ and a continuous likelihood function $f_L(\underline{x})$[4], the Bayesian filter step for calculating

[1] For general triangular transformations, see [7].

[2] This is the map used in [5].

[3] The Liouville PDE or continuity equation is a special case of the Fokker-Planck PDE for zero diffusivity.

[4] The likelihood function is usually obtained by plugging a specific measurement, say $\hat{\underline{y}}$, into the conditional density $f(\underline{y}|\underline{x})$ describing the relation between measurement $\underline{y}$ and state $\underline{x}$ such that $f_L(\underline{x}) = f(\hat{\underline{y}}|\underline{x})$.

the continuous posterior density $f_e^t(\underline{x})$ is

$$f_e^t(\underline{x}) \propto f_p^t(\underline{x}) \cdot f_L(\underline{x}) \ . \tag{2}$$

The symbol $\propto$ indicates that a normalization is required after the multiplication of $f_p^t(\cdot)$ and $f_L(\cdot)$.

For the important case that the prior density $f_p^t(\underline{x})$ is given as a set of weighted samples (or particles) it can be formally written as a Dirac mixture density

$$f_p(\underline{x}) = \sum_{i=1}^{L} w_{p,i} \cdot \delta(\underline{x} - \underline{x}_{p,i}) \ . \tag{3}$$

The weights $w_{p,i} > 0$, $\sum_{i=1}^{L} w_{p,i} = 1$, and sample locations $\underline{x}_{p,i}$ selected such that $f_p(\underline{x}) \approx f_p^t(\underline{x})$.

For a given Dirac mixture prior, the Bayesian filter step becomes

$$\begin{aligned} \tilde{f}_e(\underline{x}) &\propto f_L(\underline{x}) \cdot \sum_{i=1}^{L} w_{p,i} \cdot \delta(\underline{x} - \underline{x}_{p,i}) \\ &= \sum_{i=1}^{L} \underbrace{w_{p,i} \cdot f_L(\underline{x}_i)}_{\bar{w}_{e,i}} \cdot \delta(\underline{x} - \underline{x}_{p,i}) \ . \end{aligned} \tag{4}$$

Upon normalization, we obtain the posterior weights $\tilde{w}_{e,i} = \bar{w}_{e,i} / \sum_{i=1}^{L} \bar{w}_{e,i}$. The posterior Dirac mixture is now given as

$$\tilde{f}_e(\underline{x}) = \sum_{i=1}^{L} \tilde{w}_{e,i} \cdot \delta(\underline{x} - \tilde{\underline{x}}_{e,i}) \ . \tag{5}$$

The posterior sample locations do not change w.r.t. the prior samples, i.e., $\tilde{\underline{x}}_{e,i} = \underline{x}_{p,i}$ for $i = 1, \ldots, L$.

The posterior $\tilde{f}_e(\underline{x})$ in (5) is derived from the straightforward application of the Bayesian filter step to Dirac mixtures. This leads to a serious problem: The samples are not equally weighted anymore and do not equally contribute to the representation of the posterior. Often, some particle weights are (close to) zero, in fact dying out, leading to particle degeneracy mentioned above. A typical scenario is large system noise, which spreads the particles during the prediction step combined with low measurement noise leading to narrow likelihoods.

Many solutions, some systematic, many of heuristic nature, have been proposed to solve the degeneracy problem, which is a fundamental and difficult problem.

Remark III.1. *Our goal is to derive a Bayesian filter that inherently avoids degeneracy without any heuristic approaches. It should be easy to understand, simple to implement, numerically stable, and robust.*

Furthermore, we propose to use deterministic particles instead of random ones. This (i) reduces the required number of particles as the placement is more homogeneous and (2) ensures reproducibility. In this paper, we use the sampling method from [24] also used in [25].

IV. Optimal Resampling

We now develop an optimal resampling step. It replaces the non-equally weighted Dirac mixture $\tilde{f}_e(\underline{x})$ with its equally weighted approximation $f_e(\underline{x})$. $\tilde{f}_e(\underline{x})$ is composed of non-equal weights $\tilde{w}_{e,i}$ and locations $\tilde{\underline{x}}_{e,i} = \underline{x}_{p,i}$. $f_e(\underline{x})$ has equal weights, i.e., $w_{e,i} = w_{p,i} = 1/L$ and new locations $\underline{x}_{e,i}$.

We collect the weights and locations in sets for $f_e(\underline{x})$

$$\mathcal{W}_e = \{w_{e,1}, \ldots, w_{e,L}\} \ , \ \mathcal{X}_e = \left\{\underline{x}_{e,1}, \ldots, \underline{x}_{e,L}\right\} \ , \tag{6}$$

and for $\tilde{f}_e(\underline{x})$

$$\tilde{\mathcal{W}}_e = \{\tilde{w}_{e,1}, \ldots, \tilde{w}_{e,L}\} \ , \ \tilde{\mathcal{X}}_e = \left\{\tilde{\underline{x}}_{e,1}, \ldots, \tilde{\underline{x}}_{e,L}\right\} \ , \tag{7}$$

instead of vectors and matrices to underline that there is no inherent order.

Remark IV.1. *We first consider the case that no degeneration of $\tilde{f}_e(\underline{x})$ occurred. This means that the weight variance in $\tilde{w}_{e,i}$ is small and all samples contribute to the density representation. Degeneration will be treated in Sec. V.*

The key idea to finding $\underline{x}_{e,i}$ is to use an optimal map $\underline{M}(.)$ that transforms the prior random vector $\underline{\boldsymbol{x}}_p$ to the posterior random vector $\underline{\boldsymbol{x}}_e$

$$\underline{\boldsymbol{x}}_e = \underline{M}(\underline{\boldsymbol{x}}_p) \ . \tag{8}$$

This map is used to map prior samples $\underline{x}_{p,i}$ to posterior samples $\underline{x}_{e,i}$. Mapping samples does not change their weights but their locations. This guarantees equally weighted posterior samples $\underline{x}_{e,i}$. We now have to find a map $\underline{M}(.)$ that leads to posterior samples that fulfill $f_e(\underline{x}) \approx \tilde{f}_e(\underline{x})$.

The map generation is shown in Fig. 1. $\tilde{f}_e(\underline{x})$ is obtained by a Bayes update, i.e., by multiplying $f_p(\underline{x})$ with the likelihood $f_L(\underline{x})$ (upper path). It serves as the reference density. Its locations $\tilde{\underline{x}}_{e,i}$ are identical to those of $f_p(\underline{x})$, only its weights $\tilde{w}_{e,i}$ are changed. In the lower path, the map $\underline{M}(.)$ propagates $f_p(\underline{x})$ to $f_e(\underline{x})$. $f_e(\underline{x})$ has identical weights as $f_p(\underline{x})$. Its locations $\underline{x}_{e,i}$ have changed due to the mapping though. We desire $f_e(\underline{x})$ to be close to $\tilde{f}_e(\underline{x})$ w.r.t. an appropriate distance measure D. The map $\underline{M}(.)$ is adjusted accordingly by minimizing $D(f_e(\underline{x}), \tilde{f}_e(\underline{x}))$.

A. Properties of Map

It would be sufficient to use a discrete map for mapping the L samples individually. However, in that case the number of map parameters would be equal to the number of samples. As a result, we would not have a complexity gain compared to a direct reapproximation. In addition, there would be no smoothing.

Hence, we use a continuous map, preferably one with as few parameters as possible to reduce complexity. This automatically allows interpolation between samples. In summary, a few prior samples economically produce a smooth mapping that can be used to map many samples.

Remark IV.2. *Interpolation can be used to increase the number of posterior samples $\underline{x}_i^e$. This can be done after the map has been generated.*

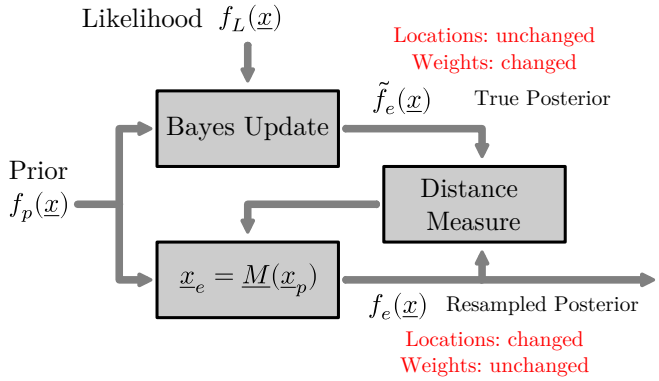

Fig. 1: Block diagram of generating a resampling map given prior $f_p(\underline{x})$ and likelihood $f_L(\underline{x})$. $f_e(\underline{x})$ is the non-equally weighted posterior resulting from the base update. $f_e(\underline{x})$ is the resampled equally weighted posterior with $f_e(\underline{x}) \approx \tilde{f}_e(\underline{x})$.

The map $M(.)$ has the following properties:

Property 1 Might be non-monotonic.
Property 2 No inverse required.
Property 3 Only used for mapping sample values.
Property 4 Differentiability w.r.t. parameters required.
Property 5 Differentiability not required w.r.t. $\underline{x}$.

B. Specific Map

According to Remark IV.1, in this section we assume that the change from $f_p(\underline{x})$ to $\tilde{f}_e(\underline{x})$ is small. Hence, the map $\underline{M}(.)$ is close to the identity mapping.

Here, we propose a combination of an affine base map combined with a radial basis function nonlinearity. For a single output it is given by

$$M_i(\underline{x}) = \underbrace{\underline{a}_i^T \cdot \underline{x} + b_i}_{\text{affine part}} + \underbrace{\sum_{r=1}^{R} v_{r,i} \cdot \mathrm{RBF}_r(\underline{x} - \underline{x}_r)}_{\text{radial basis function part}} \ , \tag{9}$$

with weights $v_{r,i} \in \mathbb{R}$, locations $\underline{x}_r$, and radial basis kernels $\mathrm{RBF}_r(.)$. The complete vector-valued map $\underline{M}(.)\colon \mathbb{R}^N \to \mathbb{R}^N$ is given by

$$\underline{M}(\underline{x}) = \big[M_1(\underline{x}), M_2(\underline{x}), \ldots, M_N(\underline{x})\big]^T \ . \tag{10}$$

For the optimization, initial values can simply be set to $\underline{a}_i = \underline{1}$, $b_i = 0$, and $v_{r,i} = 0$ for $r = 1, \ldots, R$, $i = 1, \ldots, N$. The RBF locations $\underline{x}_r$ could be set to fixed a priori locations for $r = 1, \ldots, R$.

C. Distance Measure

We have to compare the weighted Dirac mixture $\tilde{f}_e(\underline{x})$ with its unweighted counterpart $f_e(\underline{x})$. A suitable distance measure should satisfy the following requirements:

Requirement 1: Handle Dirac mixture densities[5].
Requirement 2: Handle non-equal weights.
Requirement 3: Handle non-equal supports.
Requirement 4: Be differentiable w.r.t. locations (and weights).

The first three properties cannot be handled with standard distances that require continuous densities. This includes the KL-divergence and integral squared distances. On the other hand, Wasserstein distances could be used. However, they suffer from large complexity. In addition, property 4 is not fulfilled. Here, we propose the use of the Cramér-von Mises distance [22] based on Localized Cumulative Distributions [23].

For two Dirac mixture densities f_x with L components, weights $w_{x,1}, w_{x,2}, \ldots, w_{x,L}$, locations

$$\underline{x}_i = \big[x_{1,i}, x_{2,i}, \ldots, x_{N,i}\big]^T \in \mathbb{R}^N \tag{11}$$

for $i = 1, \ldots, L$ and f_y with M components, weights $w_{y,1}, w_{y,2}, \ldots, w_{y,M}$, locations

$$\underline{y}_j = \big[y_{1,j}, y_{2,j}, \ldots, y_{N,j}\big]^T \in \mathbb{R}^N \tag{12}$$

for $j = 1, \ldots, M$, the Cramér-von Mises distance D is given by

$$D = D_{yy} - 2D_{xy} + D_{xx} + c\, D_E \ , \tag{13}$$

with

$$D_{yy} = \sum_{i=1}^{M}\sum_{j=1}^{M} w_{y,i} \cdot w_{y,j} \cdot \mathrm{xlog}\left(\sum_{d=1}^{D} (y_{d,i} - y_{d,j})^2\right) \ , \tag{14}$$

$$D_{xy} = \sum_{i=1}^{L}\sum_{j=1}^{M} w_{x,i} \cdot w_{y,j} \cdot \mathrm{xlog}\left(\sum_{d=1}^{D} (x_{d,i} - y_{d,j})^2\right) \ , \tag{15}$$

$$D_{xx} = \sum_{i=1}^{L}\sum_{j=1}^{L} w_{x,i} \cdot w_{x,j} \cdot \mathrm{xlog}\left(\sum_{d=1}^{D} (x_{d,i} - x_{d,j})^2\right) \ , \tag{16}$$

and $\mathrm{xlog}(z) = z \cdot \log(z)$. D_E can be viewed as a penalty term (with weight c) that ensures equal means and is given by

$$D_E = \sum_{d=1}^{N}\left(\sum_{i=1}^{L} w_{x,i} \cdot x_{d,i} - \sum_{i=1}^{M} w_{x,i} \cdot y_{d,i}\right)^2 \ . \tag{17}$$

When only the minimizer w.r.t. parameters of f_x is desired, but not the corresponding value of D, D_{yy} can be neglected.

D. Map Optimization

In this paper, L and M are assumed to be equal, with

$$w_{x,i} = w_{e,i} \ , \quad w_{y,i} = \tilde{w}_{e,i} \ , \quad \underline{x}_i = \underline{x}_{e,i} \ , \quad \underline{y}_i = \underline{\tilde{x}}_{e,i} \ , \tag{18}$$

for $i = 1, \ldots, L$. This leads to the following dependency of D

$$D = D\big(\mathcal{W}_e, \mathcal{X}_e, \tilde{\mathcal{W}}_e, \tilde{\mathcal{X}}_e\big) \ . \tag{19}$$

As $\mathcal{X}_e$ is obtained from $\tilde{\mathcal{X}}_e$ via the map $\mathcal{X}_e = \underline{M}(\tilde{\mathcal{X}}_e)$, this can be rewritten as

$$D = D\big(\mathcal{W}_e, \underline{M}(\tilde{\mathcal{X}}_e), \tilde{\mathcal{W}}_e, \tilde{\mathcal{X}}_e\big) \ . \tag{20}$$

The optimal map $\underline{M}^*$ is now found by minimization

$$\underline{M}^* = \operatorname*{arg\,min}_{\underline{M} \in \mathcal{M}} D\big(\mathcal{W}_e, \underline{M}(\tilde{\mathcal{X}}_e), \tilde{\mathcal{W}}_e, \tilde{\mathcal{X}}_e\big) \ , \tag{21}$$

where $\mathcal{M}$ is the set of viable maps.

The gradient of D in (13) is available in closed form [22]. When the likelihood is given in analytic form and differentiable, the gradient with respect to the map parameters can be derived. A BFGS quasi-Newton method is used for optimization.

[5] i.e., discrete densities on a continuous domain.

V. Progressive Processing

In the previous section, we assumed that reweighting of particles with the likelihood $f_L(\underline{x})$ kept all particles "alive". Usually, however, performing the Bayes update in one step leads to particle degeneration. Only a few samples stay "alive", the remaining ones are close to zero. In that case, resolution is lost as not all particles contribute to the density representation.

A proven remedy for keeping particles "alive" is to perform progressive processing [21]. The likelihood is decomposed into a product of sub-likelihoods, each of which is carefully selected to avoid degeneration

$$f_L(\underline{x}) = f_L^{(1)}(\underline{x}) \cdot f_L^{(2)}(\underline{x}) \cdots f_L^{(K)}(\underline{x}) = \prod_{k=1}^{K} f_L^{(k)}(\underline{x}) \quad . \quad (22)$$

As we use a product decomposition, each sub-likelihood is intuitively "wider" than the original one. Sequential Bayes sub-updates with the sub-likelihoods then provide the desired posterior:

$$f_e(\underline{x}) \propto f_p(\underline{x}) \cdot f_L(\underline{x}) = \underbrace{f_p(\underline{x}) \cdot f_L^{(1)}(\underline{x})}_{\text{non-equally weighted } \tilde{f}_e^{(1)}(\underline{x})}$$

Resampling RS

$$\underbrace{f_e^{(1)}(\underline{x}) \cdot f_L^{(2)}(\underline{x})}_{\text{non-equally weighted } \tilde{f}_e^{(2)}(\underline{x})}$$

Resampling RS

$$\underbrace{f_e^{(2)}(\underline{x}) \cdot f_L^{(3)}(\underline{x})}_{\text{non-equally weighted } \tilde{f}_e^{(K)}(\underline{x})}$$

Resampling RS

$$f_e^{(K)}(\underline{x}) \approx f_e(\underline{x})$$

After every sub-update, we obtain an non-equally weighted sub-posterior $\tilde{f}_e^{(k)}(\underline{x})$. In order to prepare for the next sub-update, the optimal resampling method from Sec. IV is used. This involves mapping samples $\tilde{\underline{x}}_{e,i}^{(k)}$ to $\underline{x}_{e,i}^{(k)}$ with sub-mapping $\underline{M}^{(k)}(.)$. This results in an equally weighted sub-posterior $f_e^{(k)}(\underline{x}) \approx \tilde{f}_e^{(k)}(\underline{x})$. After K sub-update steps, the result is an equally weighted sub-posterior $f_e^{(K)}(\underline{x})$, which is equal to $f_e(\underline{x})$.

The total map from prior samples $\underline{x}_{p,i}$ to posterior samples $\underline{x}_{e,i}$ is given by composition of the sequence of individual mappings as

$$\underline{M}(\underline{x}) = \underline{M}^{(K)}\left(\underline{M}^{(K-1)}\left(\ldots \underline{M}^{(2)}\left(\underline{M}^{(1)}(\underline{x})\ldots\right)\right)\right) \quad (23)$$

or

$$\underline{M} = \underline{M}^{(K)} \circ \underline{M}^{(K-1)} \circ \cdots \circ \underline{M}^{(2)} \circ \underline{M}^{(1)} \quad . \quad (24)$$

VI. Numerical Results

A. Sanity Check: Linear System

As a first example, we consider the simplest scalar linear measurement equation $y = x + v$. x is the desired state, y is the measurement, and v is Gaussian measurement noise with $v \sim f_v(v) = \mathrm{N}(v; 0, 1)$[6]. The true prior is Gaussian and given by $f_p^t(x) = \mathrm{N}(x; 0, 1)$. For $y = \hat{y}$, the likelihood is given by $f_L(x) = f(\hat{y}|x) = f_v(\hat{y} - x) = \mathrm{N}(\hat{y}; x, 1)$.

We assume that only samples $x_{p,i}$ of the prior density $f_p^t(x)$ are available. This gives the prior Dirac mixture density $f_p(x)$ in (3). The proposed method is used to perform the Bayesian filter step with prior $f_p(x)$ and likelihood $f_L(x)$. This produces samples $x_{e,i}$ of the posterior $f_e(x)$.

In this simple case, we can use the analytic prior and the likelihood to calculate an analytic true posterior and its CDF. The true continuous posterior is given by $f_e^t(x) = \mathrm{N}(x; \hat{y}/2, 1/2)$. Samples $x_{e,i}^t$ from $f_e^t(x)$ can now be used as reference for the samples $x_{e,i}$. Note that the analytic prior and the reference samples are not known to the estimator.

a) Comparison with Naive Bayes Update: We use $L = 10$ particles from a prior Gaussian with standard deviation $\sigma_p = 1$. Three noise standard deviations $\sigma_v = 1$, $\sigma_v = 0.75$, and $\sigma_v = 0.5$ are considered. In Fig. 2, the true continuous posterior $f_e^t(x)$ is shown in green for the three cases. In black, the result of naively multiplying $f_p(x)$ with the likelihood $f_L(x)$ is shown, yielding the non-equally weighted Dirac mixture $\tilde{f}_e(x)$. The outer samples of $\tilde{f}_e(x)$ are close to zero and do not contribute a lot. In contrast, the samples $x_{e,i}$ of $f_e(x)$ generated by the proposed Bayesian particle flow are shown in red. $f_e(x)$ approximates the true continuous posterior $f_e^t(x)$ well, but for smaller noise variances, the approximation gets slightly worse. This is expected due to the finite resolution of the prior density with only $L = 10$ particles.

b) Comparison with reference samples: In Fig. 3, the results of the same setup are shown for $L = 30$ particles. The true continuous posterior is again shown in green. It is almost perfectly approximated by the samples generated by the proposed Bayesian particle flow (in red). For smaller noise variances, the approximation gets only slightly worse. As a comparison, the reference samples $x_{e,i}^t$ are shown in black.

c) Comparison of Maps: The corresponding total map $M(.)$ from prior $\boldsymbol{x}_p$ to posterior $\boldsymbol{x}_e$ is shown in Fig. 6 (in red). It is compared with the true linear map (in green) calculated analytically from the CDFs $F_p^t(x)$ and $F_e^t(x)$ with (1) as $M^t(x) = Q^t(F_p^t(x))$ with $Q^t(y) = (F_e^t)^{-1}(y)$. The maps are almost identical in the relevant region.

B. Cubic Sensor Problem

We now consider the (in-)famous cubic sensor problem with measurement equation $y = x^3 + v$. Again, x is the state, y the measurement, and v Gaussian measurement noise with $v \sim f_v(v) = \mathrm{N}(v; 0, \sigma_v)$. The prior is Gaussian and given by $f_p(x) = \mathrm{N}(x; 0, 1)$. For $y = \hat{y}$, the likelihood is given by

[6] $\mathrm{N}(x; m, \sigma)$ is a Gaussian density over realizations x with mean m and standard deviation σ.

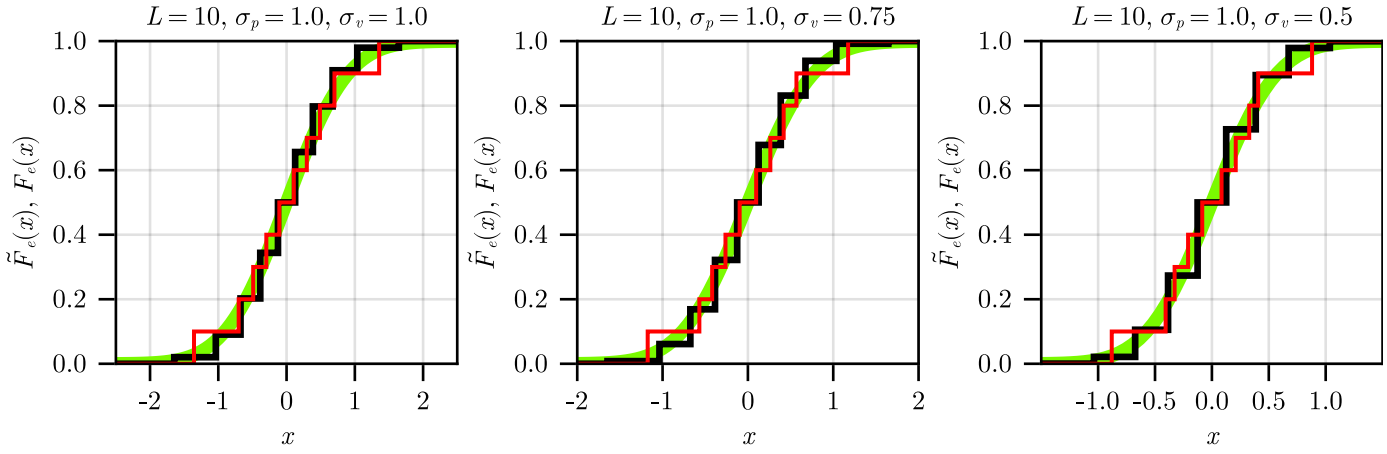

Fig. 2: Results of a linear update with $L = 10$ samples. The true continuous posterior is shown in green. The non-equally weighted Dirac mixture $\tilde{f}_e(x)$ after multiplication of $f_p(x)$ with the likelihood $f_L(x)$ is shown in black. The equally weighted Dirac mixture $f_e(x)$ produced by the proposed Bayesian particle flow is shown in red. Please note the different x-axes scales.

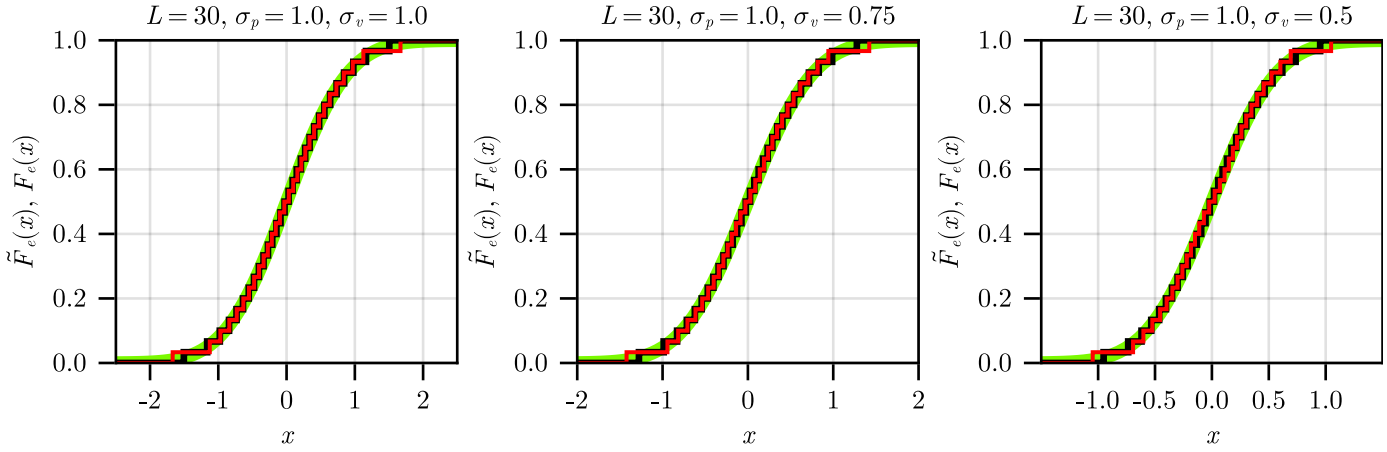

Fig. 3: Results of a linear update with $L = 30$ samples. The true continuous posterior $f_e^t(x)$ is shown in green. The equally weighted Dirac mixture $f_e(x)$ produced by the proposed Bayesian particle flow is shown in red. The reference Dirac mixture obtained by directly sampling from the true continuous posterior $f_e^t(x)$ is shown in black. Please note the different x-axes scales.

$f_L(x) = f_v(\hat{y} - x^3) = \mathrm{N}\left(\hat{y}; x^3, \sigma_v\right)$. The true posterior is given by

$$\tilde{f}_e(x) \propto f_p(x) \cdot f_L(x) = \mathrm{N}(0,1) \cdot \mathrm{N}\left(\hat{y}; x^3, \sigma_v\right) \quad . \tag{25}$$

Remark VI.1. *Sampling from $\tilde{f}_e(x)$ is difficult. Calculating the required CDF and its inverse can only be done by numerical integration.*

We are given only samples $x_{p,i}$ of the prior density $f_p(x)$ and the analytic likelihood $f_L(x)$. Samples $x_{e,i}$ of the posterior $f_e(x)$ are calculated with the proposed Bayesian flow. Fig. 4 (4, 5) show the prior PDF $f_p(x)$, the prior CDF $F_p(x)$, and its samples $x_{p,i}$. Fig. 4 (3) shows the flow of the particles. Fig. 4 (1, 2) show the posterior CDF $F_e(x)$, the posterior PDF $f_e(x)$, and its samples $x_{e,i}$. The total mapping from $\boldsymbol{x}_p$ to $\boldsymbol{x}_e$ is shown in Fig. 7 (in red) compared to a reference map $M^t(.)$ (in green). $M^t(.)$ is calculated from $F_p^t(.)$ and a numerically calculated $Q^t(.) = (F_e^t)^{-1}(.)$ with (1). Again, the maps are almost identical in the relevant region.

C. Comparison with Particle Filter

The proposed Bayesian particle flow will now be compared with the standard particle filter. The underlying continuous prior is given by $f_p^t(x) = \mathrm{N}(x; 0, 1)$ and shown in blue in Fig. 8. We only have a Dirac mixture approximation $f_p(x) \approx f_p^t(x)$ as in (3) available. The likelihood

$$f_L(x) = \exp\left(-\frac{1}{2}\big((x-1.2)(x-1.5)(x+1.2)(x+1.5)\big)^2\right) \tag{26}$$

is shown in green. The true continuous posterior $f_e^t(x)$ (in red) is unknown to the estimator.

The results are shown in Fig. 5. The posterior estimate of the proposed Bayesian particle flow for $L = 50$ particles in Fig. 5 (1) is very close to the true posterior. Results of the standard particle filter are shown in Fig. 5 (2)–(5) for different L with ten runs each. For $L = 50$, the results are of low resolution at the peaks of the posterior. This is due to the fact that the filter step of the particle filter produces weighted samples. In addition, there is large variability between different

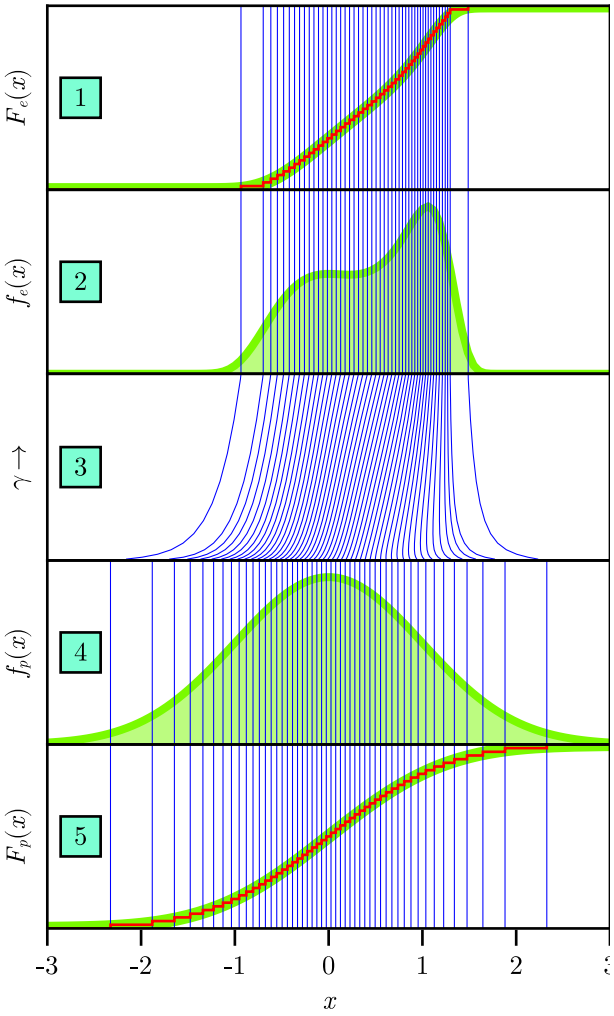

Fig. 4: Cubic sensor problem: Results of update. (1,2) Posterior PDF $f_e(x)$, CDF $F_e(x)$, and samples $x_{e,i}$. (3) Flow of particles. (4,5) Prior PDF $f_p(x)$, CDF $F_p(x)$, and samples $x_{p,i}$.

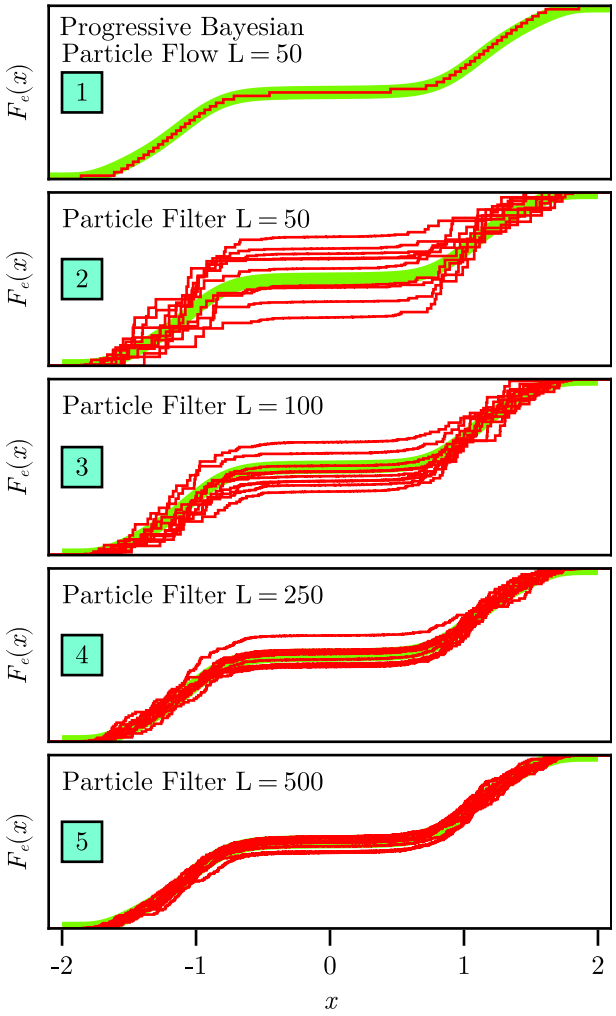

Fig. 5: Comparing the Bayesian particle flow with standard particle filter for single measurement update with likelihood shown in Fig. 8. (1) Result of proposed filter for $L = 50$. (2)–(5) Results of particle filter for different L with 10 runs each.

runs. For increasing L, the resolution degradation becomes less pronounced and the variability between runs decreases. However, even for $L = 500$ the variability is still clearly present.

VII. Conclusions

A new Bayesian particle flow has been derived that deterministically guides particles from prior to posterior. It does not require any continuous density representations, neither in its derivation nor in its implementation. It is composed of a finite sequence of potentially non-monotonic maps for propagating particles. The filter works with arbitrary nonlinear measurement equations. However, it is assumed that it has already been converted to a likelihood function that can be evaluated. The method is easy to understand and its implementation is straightforward.

Several distance measures could be employed for map optimization, see Subsec. IV-A. However, we found the distance derived in [22] most useful as it has low complexity and is differentiable. Its complexity is quadratic in the number of particles L and linear in the number of dimensions D.

Several aspects have been omitted in this paper due to space restrictions. (i) We did not give details on how to find appropriate sub-likelihoods that keep particles "alive" and their number. For that purpose, we use the method described in [26]. (ii) We assumed that the number of particles L is constant. This is not necessary. The number of samples can be adapted to the complexity of the underlying density. Methods for an efficient adaptation will be developed. (iii) We did not discuss the prediction step, i.e., propagating particles through the system model. For deterministic particles, the prediction step is significantly different from the random case in terms of combining state and noise samples. In order to avoid a full Cartesian product, the method from [27] can be used.

References

[1] Uwe D. Hanebeck, Kai Briechle, and Andreas Rauh. "Progressive Bayes: A New Framework for Nonlinear State Estimation". *Proceedings of SPIE, AeroSense Symposium*. Vol. 5099. Orlando, Florida, USA, May 2003, pp. 256–267.

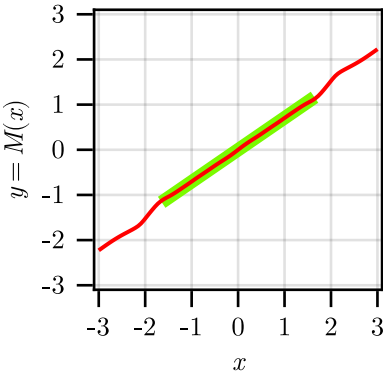

Fig. 6: Map $M(.)$ for linear filter step.

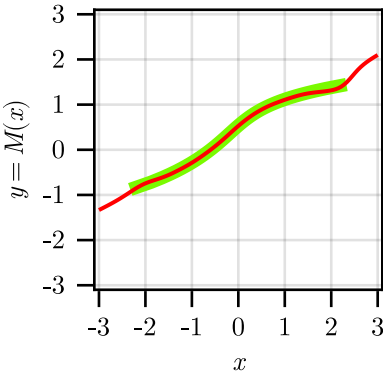

Fig. 7: Map $M(.)$ for cubic sensor problem.

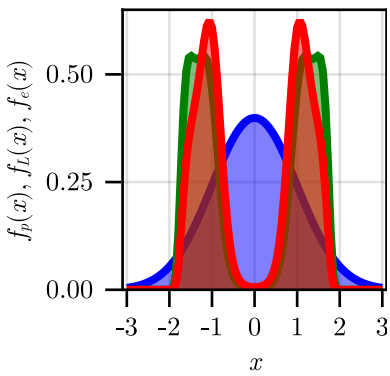

Fig. 8: Green: Likelihood $f_L(x)$ in (26). Blue: $f_p^t(x)$. Red: $f_e^t(x)$.

[2] J. Hagmar et al. "Optimal Parameterization of Posterior Densities Using Homotopy". *Proceedings of the 14th International Conference on Information Fusion (Fusion 2011)*. Chicago, Illinois, July 2011.

[3] Robert J. McCann. "Existence and Uniqueness of Monotone Measure-Preserving Maps". *Duke Mathematical Journal* 80.2 (Nov. 1, 1995).

[4] Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes*. 3rd ed. McGraw-Hill Series in Electrical Engineering. New York: McGraw-Hill, 1991.

[5] Tarek A. El Moselhy and Youssef M. Marzouk. "Bayesian Inference with Optimal Maps". *Journal of Computational Physics* 231.23 (Oct. 1, 2012), pp. 7815–7850.

[6] Herbert Knothe. "Contributions to the Theory of Convex Bodies." *Michigan Mathematical Journal* 4.1 (Jan. 1957), pp. 39–52.

[7] V. I. Bogachev, A. V. Kolesnikov, and K. V. Medvedev. "Triangular Transformations of Measures". *Sbornik: Mathematics* 196.3 (Apr. 30, 2005), p. 309.

[8] Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. "Inference Via Low-Dimensional Couplings". *The Journal of Machine Learning Research* 19.1 (Jan. 1, 2018), pp. 2639–2709.

[9] Youssef Marzouk et al. "Sampling via Measure Transport: An Introduction". *Handbook of Uncertainty Quantification*. Ed. by Roger Ghanem, David Higdon, and Houman Owhadi. Cham: Springer International Publishing, 2016, pp. 1–41.

[10] Alessio Spantini, Ricardo Baptista, and Youssef Marzouk. *Coupling Techniques for Nonlinear Ensemble Filtering*. June 30, 2019. arXiv: 1907.00389v1 [stat.ME]. preprint.

[11] Danilo Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows". *International Conference on Machine Learning*. PMLR, 2015, pp. 1530–1538.

[12] George Papamakarios et al. "Normalizing Flows for Probabilistic Modeling and Inference". *The Journal of Machine Learning Research* 22.1 (Jan. 1, 2021), 57:2617–57:2680.

[13] Fred Daum and Jim Huang. "Particle Flow for Nonlinear Filters with Log-Homotopy". *Proc. SPIE, Signal and Data Processing of Small Targets*. Vol. 6969. 2008.

[14] Frederick Daum et al. "Gradient Estimation for Particle Flow Induced by Log-Homotopy for Nonlinear Filters". *Proc. SPIE, Signal Processing, Sensor Fusion, and Target Recognition XVIII*. Vol. 7336. 2009.

[15] Fred Daum and Jim Huang. "Seven Dubious Methods to Mitigate Stiffness in Particle Flow with Non-Zero Diffusion for Nonlinear Filters, Bayesian Decisions, and Transport". *Signal and Data Processing of Small Targets 2014*. SPIE Defense + Security. Vol. 9092. Baltimore, MD, 2014.

[16] Tao Ding and M.J. Coates. "Implementation of the Daum-Huang Exact-Flow Particle Filter". *2012 IEEE Statistical Signal Processing Workshop (SSP)*. 2012 IEEE Statistical Signal Processing Workshop (SSP). Aug. 2012, pp. 257–260.

[17] Jeremy Heng, Arnaud Doucet, and Yvo Pokern. "Gibbs Flow for Approximate Transport with Applications to Bayesian Computation". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83.1 (Feb. 1, 2021), pp. 156–187.

[18] De Melo et al. *Stochastic Particle Flow for Nonlinear High-Dimensional Filtering Problems*. Nov. 4, 2015. arXiv: 1511.01448v3 [stat.ME]. preprint.

[19] Uwe D. Hanebeck. "FLUX: Progressive State Estimation Based on Zakai-type Distributed Ordinary Differential Equations". *arXiv preprint: Systems and Control (cs.SY)* (Aug. 2018).

[20] Uwe D. Hanebeck. "Kernel-Based Deterministic Blue-noise Sampling of Arbitrary Probability Density Functions". *Proceedings of the 48th Annual Conference on Information Sciences and Systems (CISS 2014)*. Princeton, New Jersey, USA, Mar. 2014.

[21] Patrick Ruoff, Peter Krauthausen, and Uwe D. Hanebeck. "Progressive Correction for Deterministic Dirac Mixture Approximations". *Proceedings of the 14th International Conference on Information Fusion (Fusion 2011)*. Chicago, Illinois, USA, July 2011.

[22] Uwe D. Hanebeck. "Optimal Reduction of Multivariate Dirac Mixture Densities". *at – Automatisierungstechnik, Oldenbourg Verlag* 63.4 (Apr. 2015), pp. 265–278.

[23] Uwe D. Hanebeck and Vesa Klumpp. "Localized Cumulative Distributions and a Multivariate Generalization of the Cramér-von Mises Distance". *Proceedings of the 2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2008)*. Seoul, Republic of Korea, Aug. 2008, pp. 33–39.

[24] Uwe D. Hanebeck, Marco F. Huber, and Vesa Klumpp. "Dirac Mixture Approximation of Multivariate Gaussian Densities". *Proceedings of the 2009 IEEE Conference on Decision and Control (CDC 2009)*. Shanghai, China, Dec. 2009.

[25] Jannik Steinbring, Martin Pander, and Uwe D. Hanebeck. "The Smart Sampling Kalman Filter with Symmetric Samples". *Journal of Advances in Information Fusion* 11.1 (June 2016), pp. 71–90.

[26] Jannik Steinbring and Uwe D. Hanebeck. "Progressive Gaussian Filtering Using Explicit Likelihoods". *Proceedings of the 17th International Conference on Information Fusion (Fusion 2014)*. Salamanca, Spain, July 2014.

[27] Henning Eberhardt, Vesa Klumpp, and Uwe D. Hanebeck. "Optimal Dirac Approximation by Exploiting Independencies". *Proceedings of the 2010 American Control Conference (ACC 2010)*. Baltimore, Maryland, USA, June 2010.