Belief Condensation Filtering

Santiago Mazuelas, *Member, IEEE*, Yuan Shen, *Student Member, IEEE*, and Moe Z. Win, *Fellow, IEEE*

***Abstract*—Inferring a sequence of variables from observations is prevalent in a multitude of applications. Traditional techniques such as Kalman filters (KFs) and particle filters (PFs) are widely used for such inference problems. However, these techniques fail to provide satisfactory performance in many important nonlinear or non-Gaussian scenarios. In addition, there is a lack of a unified methodology for the design and analysis of different filtering techniques. To address these problems, in this paper, we propose a new filtering methodology called belief condensation (BC) filtering. First, we establish a general framework for filtering techniques and propose an optimality criterion that leads to BC filtering. We then propose efficient BC algorithms that can best represent the complex distributions arising in the filtering process. The performance of the proposed techniques is evaluated for two representative nonlinear/non-Gaussian problems, showing that the BC filtering can provide accuracy approaching the theoretical bounds and outperform existing techniques in terms of the accuracy versus complexity tradeoff.**

***Index Terms*— Nonlinear Filters, Filtering Algorithms, Inference Algorithms, Navigation.**

I. Introduction

INFERRING a sequence of variables from a sequence of observations (e.g., measurements of a time-evolving variable) is a prevalent task in many applications. Examples include navigation of mobile nodes [1]–[3], channel estimation and data detection [4], [5], speech recognition [6], [7], time series analysis in econometrics [8], [9], and power grid state estimation [10], [11]. Such inference tasks can be often described by a hidden Markov model (HMM), where the variables of interest are called hidden states and the observations are called measurements.[1] *Filtering* in HMMs is an inference process that aims to determine posterior distributions of hidden states given present and past measurements. The Markov assumption in HMMs together with Bayes' rule enable recursive filtering implementation, where the posterior distribution of the current state can be obtained from that of previous state using new measurements and the system models. These models characterize the evolution of states (dynamic model) and the relationship between the observations and the states (measurement model).

Inference using nonlinear and/or non-Gaussian system models is extremely important since the behavior of many practical scenarios cannot be captured by linear-Gaussian models [13], [14]. For example, nonlinear and non-Gaussian system models arise in navigation, which has attracted substantial research interest in recent years [15]–[21]. Navigation systems aim to obtain the positional states of agents from inter- and intra-node measurements. The relationship between these measurements and positional states is nonlinear. Moreover, the measurement errors cannot be accurately modeled by Gaussian distributions in many environments such as urban and indoors due to harsh propagation conditions.

Optimal algorithms for sequential Bayesian inference are only known for restricted cases: 1) when system models are linear and Gaussian [22], 2) when the number of states is finite [7], and 3) when nonlinear models belong to certain subclasses [23]. Corresponding algorithms for the three cases are the Kalman filter (KF), grid-based methods, and Beneš-Daum filters [2]. In general, the recursion given by Bayes' rule requires to propagate the complete posterior distribution, which in many cases cannot be described using a finite number of parameters. Therefore, for general continuous-state HMMs[2] with nonlinear/non-Gaussian models, practical filtering techniques must use approximations to track the posterior distributions [13], [24].

Conventional filtering techniques can be grouped into local and global methods [25], [26]. The former approximate posterior distributions in the neighborhood of a reference point, e.g., extended KF (EKF) [2], unscented KF (UKF) [13], cubature KF (CKF) [27], and quadrature KF (QKF) [28]; while the latter approximate posterior distributions over the region containing significant probability mass, e.g., particle filters (PFs),[3] Gaussian sum filters, and Rao-Blackwellized PFs [33]–[36]. Kalman-like filters have been widely applied for many important problems and were the default option until the PFs were proposed in the nineties [29]. PFs can achieve high accuracies with complexities comparable to EKF for low dimensional problems, but often require prohibitive complexities for high dimensional problems [14]. Other techniques such as Gaussian-sum filters and Rao-Blackwellized PFs have been developed to improve the accuracy versus complexity trade-off. Gaussian-sum filters [25], [37] use a bank of Kalman filters operating in parallel and approximate posterior distributions by mixtures of Gaussians. While

Manuscript received June 21, 2012; revised November 09, 2012 and April 12, 2013; accepted April 16, 2013. Date of publication May 07, 2013; date of current version August 14, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jitendra Tugnait. This research was supported, in part, by the Air Force Office of Scientific Research under Grant FA9550-12-0287 and the MIT Institute for Soldier Nanotechnologies. The material in this paper was presented, in part, at the IEEE International Conference on Communications, Kyoto, Japan, June 2011.

The authors are with the Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: mazuelas@mit.edu; shenyuan@mit.edu; moewin@mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2013.2261991

[1]HMMs have been widely used to model complex real-world problems with computational tractability [12]. For example, in the problem of navigation, the hidden states and observations often correspond to time-varying positions of agents and measurements related to the positions, respectively.

[2]HMMs with continuous state variables are referred to as state-space models [12]. Note that the observations can be either continuous or discrete.

[3]PFs include sampling importance resampling (SIR) filter [29], auxiliary SIR (ASIR) filter [30], adaptive PF (APF) [31], [32], and regularized PF (RPF) [33].

Gaussian-sum filters can outperform EKFs, their main limitation is that the number of components in the mixture increases exponentially with time [28]. Rao-Blackwellized PFs [36], also known as marginalized PFs [26], can obtain accuracies comparable to PFs but with smaller complexity; they require the assumption that the state variables can be partition into two conditionally independent sets. Finally, techniques based on variational methods (VM) and expectation propagation (EP) have been recently proposed to exploit the usage of exponential families for obtaining parsimonious approximations of complex distributions [26], [38].

The fundamental question related to sequential Bayesian filtering is how to design filters that can achieve high accuracy under computational complexity constraints in nonlinear/non-Gaussian scenarios. Existing techniques use different filtering approaches, lacking a systematic way to deal with the complexity versus accuracy trade-off. Hence, there is an essential need for a unifying methodology to guide the design and analysis of accurate and efficient nonlinear/non-Gaussian filters. Such a methodology will lead to new families of filtering techniques that can be employed in a wide range of modern applications.

In this paper we present a new methodology for nonlinear/non-Gaussian filtering called *belief condensation* (BC) filtering based on the idea of properly representing the complex distributions in the filtering process. Specifically, the main contributions of the paper are as follows.

- We establish a general framework for filtering and formulate an optimality criterion leading to the methodology of BC filtering (BCF).
- We propose algorithms for BC that best represent complex distributions by mixtures of exponential families (continuous BC) and by discrete distributions (discrete BC).
- We develop BCF techniques for nonlinear/non-Gaussian problems and compare them with existing filtering techniques under the proposed framework.
- We demonstrate the accuracy and complexity improvements of the BCFs over existing techniques in two representative nonlinear/non-Gaussian filtering problems.

The rest of the paper is organized as follows. Section II describes sequential Bayesian filtering in HMMs and presents a general framework for filtering, which leads to the methodology of BCF. We develop algorithms for continuous and discrete BC in Section III, and discuss existing filtering techniques in the proposed framework in Section IV. Section V describes the problem of navigation in harsh environments as a case study, and Section VI provides simulation results describing the performance of the proposed BCFs. Finally, conclusions are drawn in Section VII.

Notations: $\mathbf{x}_{1:k}$ denotes the sequence of random vectors (RVs) $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$, where $\mathbf{x}_i \in \mathbb{R}^n$ for $i = 1, 2, \ldots, k$; $\mathcal{P}$ denotes the set of all probability distributions in $\mathbb{R}^n$ for $n \in \mathbb{Z}_+$, $\mathcal{F}$ denotes a family of probability distributions, i.e., $\mathcal{F} \subset \mathcal{P}$, and depending on the context, we refer to a probability distribution either by its Radon-Nikodym derivative (RND) with respect to Lebesgue measure (e.g., the probability density function (PDF) for continuous RVs) or by its cumulative distribution function CDF; we denote RNDs by $f(\mathbf{x})$ and CDFs by $F(\mathbf{x})$; $\mathbb{E}_{f(\mathbf{x})}\{h(\mathbf{x})\}$ denotes the expectation of $h(\mathbf{x})$ with respect to $f(\mathbf{x})$; we denote a function g on $\mathbf{x}$ parameterized by $\boldsymbol{\theta}$ as $g(\mathbf{x};\boldsymbol{\theta})$ or simply $g_{\boldsymbol{\theta}}$ if no confusion is possible; $[\,\cdot\,]^{\mathrm{T}}$ denotes the transpose of its argument; if $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ we denote by $\mathbf{x} \leq \mathbf{y}$ the fact that $x_i \leq y_i$ for $i = 1, 2, \ldots, n$; $\mathbf{I}_n$ denotes the $n \times n$ identity matrix; and, finally, $\mathbb{1}_{\mathcal{A}}(\cdot)$ denotes the indicator function of the set $\mathcal{A}$.

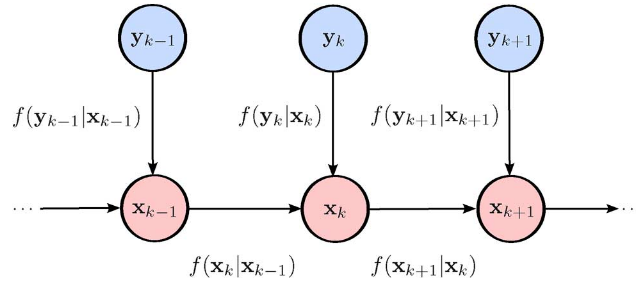


Fig. 1. Bayesian network representing a Hidden Markov Model.

II. Framework for Belief Condensation Filtering

In this section, we briefly review Bayesian inference in HMMs, and then present a general framework for filtering techniques as well as the methodology of BCF.

A. Sequential Bayesian Filtering in HMMs

A HMM is formed by a bivariate sequence of RVs $\{\mathbf{x}_k, \mathbf{y}_k\}_{k \in \mathbb{Z}_+}$, satisfying the following conditions: (i) $\{\mathbf{x}_k\}$ is a Markov chain, (ii) conditioned on $\{\mathbf{x}_k\}$, $\{\mathbf{y}_k\}$ is a sequence of independent RVs, and (iii) for each $k \in \mathbb{Z}_+$, the conditional distribution of $\mathbf{y}_k$ depends only on $\mathbf{x}_k$ [12] (see Fig. 1). The goal is to infer the variables of interest $\{\mathbf{x}_k\}$ called hidden-state variables, from the observations $\{\mathbf{y}_k\}$ called measurements.[4]

A direct consequence of the conditions in the HMM is that the joint distribution of all RVs can be factorized for $k > 1$ as

$$\begin{aligned} f(\mathbf{x}_{1:k}, \mathbf{y}_{1:k}) &= \prod_{i=1}^{k} f(\mathbf{x}_i|\mathbf{x}_{i-1}) \cdot f(\mathbf{y}_i|\mathbf{x}_i) \\ &= f(\mathbf{x}_{1:k-1}, \mathbf{y}_{1:k-1}) \cdot f(\mathbf{x}_k|\mathbf{x}_{k-1}) f(\mathbf{y}_k|\mathbf{x}_k) \end{aligned} \quad (1)$$

where $f(\mathbf{x}_1|\mathbf{x}_0) = f(\mathbf{x}_1)$. Hence, only two relationships are needed to completely describe a HMM:

- Dynamic model—the relationship between the state vector at time t_k and that at time t_{k-1}, i.e., $f(\mathbf{x}_k|\mathbf{x}_{k-1})$;
- Measurements model—the relationship between the measurements at time t_k and the state vector at time t_k, i.e., $f(\mathbf{y}_k|\mathbf{x}_k)$.

The task of inferring $\mathbf{x}_k$ from $\mathbf{y}_{1:k}$, i.e., the set of observations up to time t_k, is referred to as filtering. In a Bayesian framework, this corresponds to determining the posterior distribution $f(\mathbf{x}_k|\mathbf{y}_{1:k})$. By using Bayes' rule and marginalization, the posterior distribution for $k \in \mathbb{Z}_+$ can be written as

$$f(\mathbf{x}_k|\mathbf{y}_{1:k}) = \frac{f(\mathbf{y}_k|\mathbf{x}_k) \int f(\mathbf{x}_k|\mathbf{x}_{k-1}) f(\mathbf{x}_{k-1}|\mathbf{y}_{1:k-1}) d\mathbf{x}_{k-1}}{f(\mathbf{y}_k|\mathbf{y}_{1:k-1})}. \quad (2)$$

[4]In this paper we consider continuous-state HMM, where the state vector takes values in $\mathbb{R}^n$ and its probability distribution is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^n$.

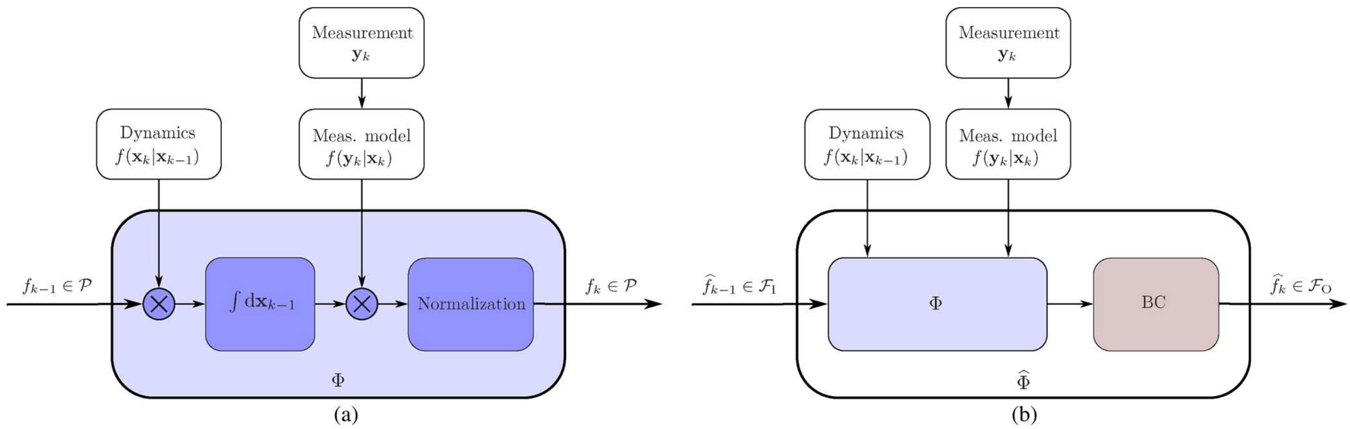


Fig. 2. Block diagrams representing Bayesian filtering and the optimal tractable architecture of BCF. (a) Block diagram representing the recursion for Bayesian filtering. (b) block diagram representing a BCF.

Once this posterior distribution is determined, point estimates can be obtained by taking the expectation or by obtaining the mode, corresponding respectively to minimum mean squared error (MMSE) or maximum a posteriori (MAP) estimators.

This paper focuses on recursive filtering techniques that obtain the posterior distribution at time t_k, i.e., $f(\mathbf{x}_k|\mathbf{y}_{1:k})$, from the previous posterior at time t_{k-1}, the new measurements $\mathbf{y}_k$, and dynamic and measurements models. Most of existing filtering techniques are recursive in this sense, for instance Kalman-like filters, Gaussian-sum filters, and PFs.[5]

B. Framework for Sequential Bayesian Filtering

In this section we present a general framework for filtering techniques based on the characterization of the filtering process as a mapping that transforms probability distributions. Under this framework each filtering technique can be viewed as a mapping that aims to approximate the exact mapping provided by (2). Then, we define a relationship of dominance (partial order) between filtering techniques that enables the formulation of an optimality criterion leading to the methodology of BCF.

The recursion given by (2) can be viewed as a mapping $\Phi : \mathcal{P} \to \mathcal{P}$ depicted in Fig. 2(a) that maps the posterior distribution $f_{k-1} = f(\mathbf{x}_{k-1}|\mathbf{y}_{1:k-1})$ to $f_k = f(\mathbf{x}_k|\mathbf{y}_{1:k})$, i.e., $f_k = \Phi(f_{k-1})$ for $k \in \mathbb{Z}_+$.[6] When the system models are linear-Gaussian and f_{k-1} is Gaussian, then $f_k = \Phi(f_{k-1})$ is also Gaussian. In such cases, f_k can be obtained in a closed form and the recursion given by Φ can be easily implemented, leading to the KF [22], [39]–[41]. However, in general when either dynamic or measurements models are nonlinear or are non-Gaussian, Φ cannot be implemented due to the lack of closed-form solutions for (2),[7] and practical nonlinear/non-Gaussian filtering techniques must resort to approximations.

Each recursive filtering technique can be viewed as a mapping $\widehat{\Phi}$ that approximates the exact mapping Φ under tractability constraints. The implementation constraints require that inputs and outputs of such mapping $\widehat{\Phi}$ belong to tractable families of distributions $\mathcal{F}_{\mathrm{I}}, \mathcal{F}_{\mathrm{O}}$, i.e., $\widehat{\Phi} : \mathcal{F}_{\mathrm{I}} \to \mathcal{F}_{\mathrm{O}}$. In the following we establish a hierarchy of filtering techniques, i.e., mappings $\widehat{\Phi}$, according to the closeness of the corresponding $\widehat{\Phi}$ to Φ. This hierarchy enables the comparison of different filtering techniques and the formulation of an optimality criterion.

Given a discrepancy D between probability distributions,[8] one can define a partial ordering of mappings $\widehat{\Phi}$ and the optimality criterion based on this ordering as follows.

Definition 1: A $(\mathcal{F}_{\mathrm{I}}, \mathcal{F}_{\mathrm{O}})$-filter $\widehat{\Phi}$ is said to D-*dominate* a $(\mathcal{F}'_{\mathrm{I}}, \mathcal{F}'_{\mathrm{O}})$-filter $\widehat{\Phi}'$ if

1) $\mathcal{F}'_{\mathrm{I}} \subseteq \mathcal{F}_{\mathrm{I}}, \mathcal{F}'_{\mathrm{O}} \subseteq \mathcal{F}_{\mathrm{O}}$, and
2) $D(\Phi(g), \widehat{\Phi}(g)) \leq D(\Phi(g), \widehat{\Phi}'(g)), \forall g \in \mathcal{F}'_{\mathrm{I}}$.

Remark 1: Given a discrepancy measure, this definition provides a partial order for filtering techniques: a filter $\widehat{\Phi}$ dominates another filter $\widehat{\Phi}'$ if the outputs of $\widehat{\Phi}$ are always closer to the outputs of the exact filter Φ.

Definition 2: A $(\mathcal{F}_{\mathrm{I}}, \mathcal{F}_{\mathrm{O}})$-filter is called D-*optimal* if it D-dominates any $(\mathcal{F}'_{\mathrm{I}}, \mathcal{F}'_{\mathrm{O}})$-filter with $\mathcal{F}'_{\mathrm{I}} \subseteq \mathcal{F}_{\mathrm{I}}$, and $\mathcal{F}'_{\mathrm{O}} \subseteq \mathcal{F}_{\mathrm{O}}$.

We next show that better filters can be obtained by using more general family of distributions, which agrees with the intuition that filters employing more general distributions to represent posterior distributions result in better performances.

Proposition 1: A D-optimal $(\mathcal{F}_{\mathrm{I}}, \mathcal{F}_{\mathrm{O}})$-filter D-dominates a D-optimal $(\mathcal{F}'_{\mathrm{I}}, \mathcal{F}'_{\mathrm{O}})$-filter if $\mathcal{F}'_{\mathrm{I}} \subseteq \mathcal{F}_{\mathrm{I}}$ and $\mathcal{F}'_{\mathrm{O}} \subseteq \mathcal{F}_{\mathrm{O}}$.

Proof: If $\widehat{\Phi}$ and $\widehat{\Phi}'$ are D-optimal $(\mathcal{F}_{\mathrm{I}}, \mathcal{F}_{\mathrm{O}})$- and $(\mathcal{F}'_{\mathrm{I}}, \mathcal{F}'_{\mathrm{O}})$-filters, respectively. Then, $\widehat{\Phi}$ D-dominates $\widehat{\Phi}'$ because for any $g \in \mathcal{F}'_{\mathrm{I}}$ we have that $D(\Phi(g), \widehat{\Phi}(g)) \leq D(\Phi(g), \widehat{\Phi}'(g))$ since $g \in \mathcal{F}_{\mathrm{I}}$. □

C. Optimal Filtering via Belief Condensation

In this section, we provide the process to construct D-optimal filters. We first define the concept of BC that refers to the

[5]The prevalent choice in the PFs literature is to use importance densities that enable recursive implementation, i.e., importance densities of the form $q(\mathbf{x}_k|\mathbf{x}_{1:k-1}, \mathbf{y}_{1:k}) = q(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_k)$.

[6]Note that the arguments of this mapping Φ include both the previous posterior distribution and the current measurement. In this paper, we suppress the measurements in the notation for convenience.

[7]In this case, to evaluate the posteriors $f_{k+r} = \Phi^r(f_k)$ for $r > 1$ requires all of the measurements $\mathbf{y}_{k+1:k+r}$ and the evaluation of nested integrals of order r.

[8]Examples of discrepancy functions include Kullback-Leibler divergence, Kolmogorov-Smirnov distance, L2 distance, etc.

representation of a general distribution f by a closest one in a tractable family $\mathcal{F}$, and then show that a filter is optimal if and only if it is a BC-based filter.

Definition 3: The mapping $\psi : \mathcal{P} \to \mathcal{F}$ is called a $(\mathcal{F}, D)$-*condensation* if for all f in the domain of ψ

$$D\left(f, \psi(f)\right) = \min_{\widetilde{f} \in \mathcal{F}} D(f, \widetilde{f}). \tag{3}$$

Definition 4: A $(\mathcal{F}_{\mathrm{I}}, \mathcal{F}_{\mathrm{O}})$-filter $\widehat{\Phi}$ is a *BCF* for Φ if it is formed by concatenating Φ with a $(\mathcal{F}_{\mathrm{O}}, D)$-condensation ψ, that is, $\widehat{\Phi} = \psi \circ \Phi$.

Proposition 2: A $(\mathcal{F}_{\mathrm{I}}, \mathcal{F}_{\mathrm{O}})$-filter is D-optimal if and only if it is a BCF for Φ given by a $(\mathcal{F}_{\mathrm{O}}, D)$-condensation.

Proof (Sufficiency): If $\widehat{\Phi} : \mathcal{F}_{\mathrm{I}} \to \mathcal{F}_{\mathrm{O}}$ is D-optimal, then $\widehat{\Phi}$ D-dominates a BCF given by a $(\mathcal{F}_{\mathrm{O}}, D)$-condensation ψ, i.e.,

$$D\big(\Phi(g), \widehat{\Phi}(g)\big) \leq D\left(\Phi(g), \psi \circ \Phi(g)\right) = \min_{\widetilde{f} \in \mathcal{F}_{\mathrm{O}}} D\big(\Phi(g), \widetilde{f}\big).$$

Since $\widehat{\Phi}(g) \in \mathcal{F}_{\mathrm{O}}$, it follows

$$\min_{\widetilde{f} \in \mathcal{F}_{\mathrm{O}}} D\big(\Phi(g), \widetilde{f}\big) \leq D\big(\Phi(g), \widehat{\Phi}(g)\big)$$

and hence the inequality becomes equality. Therefore, $\widehat{\Phi}$ is a BCF given by a $(\mathcal{F}_{\mathrm{O}}, D)$-condensation.

Proof (Necessity): Suppose $\widehat{\Phi} : \mathcal{F}_{\mathrm{I}} \to \mathcal{F}_{\mathrm{O}}$ is a BCF for Φ given by a $(\mathcal{F}_{\mathrm{O}}, D)$-condensation, and that $\widehat{\Phi}$ is not D-optimal. Then, there exists a mapping $\widetilde{\Phi} : \mathcal{F}'_{\mathrm{I}} \to \mathcal{F}'_{\mathrm{O}}$ with $\mathcal{F}'_{\mathrm{I}} \subseteq \mathcal{F}_{\mathrm{I}}$ and $\mathcal{F}'_{\mathrm{O}} \subseteq \mathcal{F}_{\mathrm{O}}$ that is not D-dominated by $\widehat{\Phi}$. This implies that there exists a distribution $g \in \mathcal{F}'_{\mathrm{I}} \subseteq \mathcal{F}_{\mathrm{I}}$ with

$$D\left(\Phi(g), \widetilde{\Phi}(g)\right) < D\left(\Phi(g), \widehat{\Phi}(g)\right)$$

which contradicts that $\widehat{\Phi}$ is a BCF for Φ given by a $(\mathcal{F}_{\mathrm{O}}, D)$-condensation. □

Proposition 2 shows that the BCF methodology depicted in Fig. 2(b) is the best approach to perform filtering in terms of the above optimality criterion, where specific BCFs depend on the choice of family and discrepancy. Given a family $\mathcal{F}$ and a discrepancy D, one can further classify the $(\mathcal{F}, D)$-BC filters depending on the algorithm used to perform the minimization in (3).

Notice that finding a global minimum in (3) can be difficult and hence the accuracy and complexity of BC filters is determined by that of the specific algorithm used to perform BC. In the following section, we present efficient algorithms for BC that use tractable families of distributions.

III. Algorithms for Belief Condensation

In this section, we describe suitable families of distributions and then propose BC algorithms to represent a complex distribution by a member of such families.

A. Choice of Probability Distributions

The family of distributions for filtering must be capable of representing the true posterior distributions with high accuracy, meanwhile amenable to implementation with reasonable complexity. Both mixtures of exponential families and discrete distributions are appropriate choices. In particular, mixtures of exponential families can offer more efficient representations and discrete distributions can offer simpler implementations.[9] In the following we describe algorithms to perform BC with mixtures of exponential families (continuous BC) as well as with discrete distributions (discrete BC).

B. Continuous Belief Condensation

In this section we consider families of distributions $\mathcal{F}_{\Xi^m}$ formed by mixtures of m distributions belonging to exponential families $\mathcal{F}_{\Theta_1}, \mathcal{F}_{\Theta_2}, \ldots, \mathcal{F}_{\Theta_m}$. That is, each member of $\mathcal{F}_{\Theta_i}$ has the form

$$g_i(\mathbf{x}; \boldsymbol{\theta}_i) = q_i(\mathbf{x}) \exp\left\{\boldsymbol{\theta}_i^{\mathrm{T}} \mathbf{t}_i(\mathbf{x}) - A_i(\boldsymbol{\theta}_i)\right\}$$

for $i = 1, 2, \ldots, m$, where $\boldsymbol{\theta}_i \in \Theta_i$, $\mathbf{t}_i(\mathbf{x})$, and $A_i(\boldsymbol{\theta}_i)$ are the natural parameters, sufficient statistics, and log-partition function of $\mathcal{F}_{\Theta_i}$ [42]. We say that a PDF $g(\mathbf{x})$ belongs to $\mathcal{F}_{\Xi^m}$, if there exists $\boldsymbol{\xi} = (\alpha_1, \boldsymbol{\theta}_1, \alpha_2, \boldsymbol{\theta}_2, \ldots, \alpha_m, \boldsymbol{\theta}_m) \in \Xi^m$ such that

$$g(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i g_i(\mathbf{x}; \boldsymbol{\theta}_i) \triangleq g(\mathbf{x}; \boldsymbol{\xi})$$

where $\alpha_1, \alpha_2 \ldots, \alpha_m \in \mathbb{R}_+$ and $\sum_{i=1}^{m} \alpha_i = 1$.

For continuous BC, we measure the discrepancy between the probability distributions $f(\mathbf{x}) \in \mathcal{P}$ and $g(\mathbf{x}; \boldsymbol{\xi}) \in \mathcal{F}_{\Xi^m}$ by using the Kullback-Leibler (KL) divergence D_{KL}, that is

$$D_{\mathrm{KL}}(f, g_{\boldsymbol{\xi}}) = \mathbb{E}_f\left\{\log \frac{f}{g_{\boldsymbol{\xi}}}\right\}.$$

In the following, we give a theorem describing a recursive process that condenses a continuous probability distribution $f(\mathbf{x})$, satisfying the following regularity conditions,[10] into a mixture of exponential families.

(A1) The differential entropy of f is finite

(A2) $\mathbb{E}_f\{|\log q_i(\mathbf{x})|\}$ is finite for each $i \in \{1, 2, \ldots, m\}$

(A3) $\mathbb{E}_f\{|t_{i,j}(\mathbf{x})|\}$ is finite, where $t_{i,j}(\mathbf{x})$ is the j-th component of the sufficient statistic $\mathbf{t}_i(\mathbf{x})$, for $i \in \{1, 2, \ldots, m\}$

(A4) The set $U_f = \{\boldsymbol{\xi} \in \Xi^m : D_{\mathrm{KL}}(f, g_{\boldsymbol{\xi}}) < \infty\}$ is open and non-empty.

Theorem 1 (Continuous BC): If a continuous probability distribution $f(\mathbf{x})$ satisfies the regularity conditions (A1-A4), then the sequence $\{D_{\mathrm{KL}}(f, g_{\boldsymbol{\xi}^{[l]}})\}_{l \in \mathbb{Z}_+}$ is monotonically decreasing, where the sequence $\{\boldsymbol{\xi}^{[l]}\}_{l \in \mathbb{Z}_+} \subset \Xi^m$ is recursively determined by

$$\alpha_i^{[l+1]} = \alpha_i^{[l]} \cdot \mathbb{E}_{g_i\left(\mathbf{x}; \boldsymbol{\theta}_i^{[l]}\right)}\left\{\frac{f(\mathbf{x})}{g(\mathbf{x}; \boldsymbol{\xi}^{[l]})}\right\}, \quad i = 1, 2, \ldots, m \tag{4}$$

[9]PFs use discrete distributions and can obtain better accuracies than Kalman-like filters when the true posterior distributions are far from Gaussian.

[10]Notice that these regularity conditions are very mild conditions regarding the tractability of f and the suitability of the family of distributions $\mathcal{F}_{\Xi^m}$. These conditions are satisfied in well-posed filtering problems. For instance, they are satisfied if f has finite differential entropy, first moment and second moment, and $\mathcal{F}_{\Xi^m}$ is a mixture of Gaussians.

and $\boldsymbol{\theta}_i^{[l+1]}$ satisfying

$$\mathbb{E}_{g_i\left(\mathbf{x};\boldsymbol{\theta}_i^{[l+1]}\right)}\{\mathbf{t}_i(\mathbf{x})\} = \frac{\mathbb{E}_{g_i\left(\mathbf{x};\boldsymbol{\theta}_i^{[l]}\right)}\left\{\frac{f(\mathbf{x})}{g\left(\mathbf{x};\boldsymbol{\xi}^{[l]}\right)}\mathbf{t}_i(\mathbf{x})\right\}}{\mathbb{E}_{g_i\left(\mathbf{x};\boldsymbol{\theta}_i^{[l]}\right)}\left\{\frac{f(\mathbf{x})}{g\left(\mathbf{x};\boldsymbol{\xi}^{[l]}\right)}\right\}}, \quad i = 1,2,\ldots,m \tag{5}$$

for any initial parameter

$$\boldsymbol{\xi}^{[0]} = \left(\alpha_1^{[0]}, \boldsymbol{\theta}_1^{[0]}, \alpha_2^{[0]}, \boldsymbol{\theta}_2^{[0]}, \ldots, \alpha_m^{[0]}, \boldsymbol{\theta}_m^{[0]}\right) \in U_f.$$

Moreover, if $m = 1$ the sequence converges to the global minimum of $D_{\mathrm{KL}}(f, g_{\boldsymbol{\xi}})$ in $l = 1$.

Proof: See Appendix A. □

Remark 2: Notice that expectation maximization (EM) [43] is a special case of the Continuous BC Theorem in which the expectations are evaluated as arithmetic averages using a set of samples. Moreover, EP [38] is also a special case of the Continuous BC Theorem in which the mixture has only one component, i.e., $m = 1$.

In the case where the exponential families are Gaussian, the regularity condition (A3) is equivalent to f having first and second absolute moments. In this case, $\boldsymbol{\theta}_i^{[l+1]}$ in (5) can be obtained in a closed form as shown in the following corollary.

Corollary 1: Let $\mathcal{F}_{\Xi^m}$ be the family formed by mixtures of m Gaussian distributions, where each component of the mixture is determined by a mean $\boldsymbol{\mu}_i$ and a variance matrix $\boldsymbol{\Sigma}_i$. If $f(\mathbf{x})$ is a continuous probability distribution satisfying the regularity conditions (A1-A4), then the natural parameters $\boldsymbol{\theta}_i^{[l+1]}$ in (5) of Theorem 1 can be obtained explicitly, since $\boldsymbol{\theta}_i^{[l+1]}$ consists of $(\boldsymbol{\Sigma}_i^{[l+1]})^{-1}\boldsymbol{\mu}_i^{[l+1]}$ and $-\frac{1}{2}(\boldsymbol{\Sigma}_i^{[l+1]})^{-1}$, with

$$\boldsymbol{\mu}_i^{[l+1]} = \frac{\mathbb{E}_{g_i\left(\mathbf{x};\boldsymbol{\theta}_i^{[l]}\right)}\left\{\frac{f(\mathbf{x})}{g\left(\mathbf{x};\boldsymbol{\xi}^{[l]}\right)}\mathbf{x}\right\}}{\mathbb{E}_{g_i\left(\mathbf{x};\boldsymbol{\theta}_i^{[l]}\right)}\left\{\frac{f(\mathbf{x})}{g\left(\mathbf{x};\boldsymbol{\xi}^{[l]}\right)}\right\}} \tag{6}$$

$$\boldsymbol{\Sigma}_i^{[l+1]} = \frac{\mathbb{E}_{g_i\left(\mathbf{x};\boldsymbol{\theta}_i^{[l]}\right)}\left\{\frac{f(\mathbf{x})}{g\left(\mathbf{x};\boldsymbol{\xi}^{[l]}\right)}\mathbf{x}\mathbf{x}^{\mathrm{T}}\right\}}{\mathbb{E}_{g_i\left(\mathbf{x};\boldsymbol{\theta}_i^{[l]}\right)}\left\{\frac{f(\mathbf{x})}{g\left(\mathbf{x};\boldsymbol{\xi}^{[l]}\right)}\right\}} - \boldsymbol{\mu}_i^{[l+1]}\left(\boldsymbol{\mu}_i^{[l+1]}\right)^{\mathrm{T}}. \tag{7}$$

Proof: In this case, $\mathbf{t}_1(\mathbf{x}) = \mathbf{x}$ and $\mathbf{t}_2(\mathbf{x}) = \mathbf{x}\mathbf{x}^{\mathrm{T}}$. Thus, the result follows from the previous theorem since

$$\mathbb{E}_{g_i\left(\mathbf{x};\boldsymbol{\theta}_i^{[l+1]}\right)}\{\mathbf{x}\} = \boldsymbol{\mu}_i^{[l+1]}$$
$$\mathbb{E}_{g_i\left(\mathbf{x};\boldsymbol{\theta}_i^{[l+1]}\right)}\{\mathbf{x}\mathbf{x}^{\mathrm{T}}\} = \boldsymbol{\Sigma}_i^{[l+1]} + \boldsymbol{\mu}_i{}^{[l+1]}\left(\boldsymbol{\mu}_i^{[l+1]}\right)^{\mathrm{T}}.$$

□

Remark 3: The values for means and covariances in (6) and (7) are obtained as averaged values of first and second moments of $f(\mathbf{x})$ weighted by $g_i(\mathbf{x};\boldsymbol{\theta}_i^{[k]})/g(\mathbf{x};\boldsymbol{\xi}^{[k]})$ for each $i = 1,2,\ldots,m$. Therefore, in the case $m = 1$, they become just the mean and covariance of $f(\mathbf{x})$, and in the general case they are weighted averages, with more emphasis near the modes of $g_i(\mathbf{x};\boldsymbol{\theta}_i^{[k]})$ for each $i = 1,2,\ldots,m$.

The recursions in the Continuous BC Theorem and its corollary are also applicable when the distribution $f(\mathbf{x})$ is known only up to a multiplicative constant. This fact is useful in the implementation of continuous-BC filters since the normalization step in Φ (see (2) and Fig. 2(a)) is not required.

Remark 4: The main complexity of continuous BC lies in evaluating expectations of the form $\mathbb{E}_{g(\cdot)}\{f(\cdot)h(\cdot)\}$, where $f(\cdot)$ is the distribution to condense, $g(\cdot)$ is a distribution in an exponential family, and $h(\cdot)$ is an elementary function. The fact that the expectations are taken with respect to a member of an exponential family can be exploited for efficient numerical computation. For example, if $g(\cdot)$ is a Gaussian distribution, efficient quadrature rules are known, where for state vectors of dimension n, only $2n$ or $n^2 + 1$ point-wise evaluations are needed to obtain cubature formulae of degree 3 or 5, respectively [27], [44].

C. *Discrete Belief Condensation*

In this section we consider families of distributions $\mathcal{F}_{\Xi^m}$ formed by discrete distributions with m support points. We say that a CDF $G(\mathbf{x})$ belongs to $\mathcal{F}_{\Xi^m}$, if there exists $\boldsymbol{\xi} = (\alpha_1, \boldsymbol{\theta}_1, \alpha_2, \boldsymbol{\theta}_2, \ldots, \alpha_m, \boldsymbol{\theta}_m) \in \Xi^m$ such that $G(\mathbf{x}) = \sum_{i=1}^m \alpha_i \mathbb{1}_{\{\mathbf{z}:\mathbf{z}\geq\boldsymbol{\theta}_i\}}(\mathbf{x}) \triangleq G(\mathbf{x};\boldsymbol{\xi})$, where $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_m \in \mathbb{R}^n$, $\alpha_1, \alpha_2, \ldots, \alpha_m \in \mathbb{R}_+$, and $\sum_{i=1}^m \alpha_i = 1$.[11]

For discrete BC, we measure the discrepancy between the probability distribution $F(\mathbf{x}) \in \mathcal{P}$ and $G(\mathbf{x};\boldsymbol{\xi}) \in \mathcal{F}_{\Xi^m}$ by the Kolmogorov-Smirnov (KS) distance D_{KS},[12] that is

$$D_{\mathrm{KS}}(F, G_{\boldsymbol{\xi}}) = \sup_{\mathbf{x}\in\mathbb{R}^n} |F(\mathbf{x}) - G(\mathbf{x};\boldsymbol{\xi})|.$$

In the following, we derive the lower bound of the KS distance and show how to find $G(\mathbf{x};\boldsymbol{\xi})$ that achieve the lower bound for the one-dimensional case.

Proposition 3: Let $F(\mathbf{x})$ be the CDF of a continuous RV and $\mathcal{F}_{\Xi^m}$ be the family of discrete distributions with m support points. Then

$$D_{\mathrm{KS}}(F, G_{\boldsymbol{\xi}}) \geq \frac{1}{2m}, \qquad \forall G_{\boldsymbol{\xi}} \in \mathcal{F}_{\Xi^m}.$$

Proof: See Appendix B. □

Proposition 4: Let $F(x)$ be the CDF of a continuous one-dimensional RV and $\theta_i \in \mathbb{R}$ satisfy $F(\theta_i) = \frac{2i-1}{2m}$ for $i = 1,2,\ldots,m$, then

$$D_{\mathrm{KS}}(F, G_{\boldsymbol{\xi}}) = \frac{1}{2m}$$

where $\boldsymbol{\xi} = (\alpha_1, \theta_1, \alpha_2, \theta_2, \ldots, \alpha_m, \theta_m)$ with $\alpha_i = \frac{1}{m}$ for $i = 1,2,\ldots,m$.

[11]Notice that the PDF corresponding to $G(\mathbf{x};\boldsymbol{\xi})$ is a sum of delta functions, i.e., $g(\mathbf{x};\boldsymbol{\xi}) = \sum_{i=1}^m \alpha_i \delta(\mathbf{x} - \boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_m$ are known as support points and the corresponding probabilities $\alpha_1, \alpha_2, \ldots, \alpha_m$ are known as weights.

[12]Note that the KL divergence cannot be defined since the distribution F is continuous whereas the support of the distributions $G(\mathbf{x};\boldsymbol{\xi}) \in \mathcal{F}_{\Xi^m}$ is finite. Instead, KS distance is suitable for measuring discrepancies between continuous and discrete distributions and has been widely used, for instance, to determine if a set of samples follow a specific distribution through the KS hypothesis test.

TABLE I
REPRESENTATIVE FILTERING TECHNIQUES IN TERMS OF FAMILIES OF DISTRIBUTIONS AND FILTERING APPROACHES

Filtering Techniques	Families of Distributions	Filtering Approaches
EKF [2]	Gaussian distributions	Linearize the models by Taylor expansions and use KF recursions
UKF [13] CKF [27] QKF [28]	Gaussian distributions	Approximate KF recursions by numerical integration with different quadrature rules
SIR PF [29] ASIR PF [30] APF [31], [32] RPF [33]	Discrete distributions	Approximate complex distributions by random sampling with different proposal densities
VM-based [26]	Separable distributions[13]	Approximate complex distributions by left sided KL minimization
EP-based [38]	Exponential families	Approximate complex distributions by right sided KL minimization
Gaussian sum [25], [37]	Mixtures of Gaussians	Use several Kalman-like filters operating in parallel

Proof: Let A_i be the intervals $A_0 = (-\infty, \theta_1)$, $A_i = [\theta_i, \theta_{i+1}) \subset \mathbb{R}$ for $i = 1, 2, \dots, m-1$, and $A_m = [\theta_m, \infty)$. The sets $A_i, i = 0, 1, \dots, m$ form a partition of $\mathbb{R}$ and hence

$$D_{\mathrm{KS}}(F, G_{\boldsymbol{\xi}}) = \max_i \left\{ \sup_{x \in A_i} |F(x) - G_{\boldsymbol{\xi}}(x)| \right\}.$$

Then the result follows by observing that $G_{\boldsymbol{\xi}}(x) = \frac{i}{m}$ for all $x \in A_i$ and

$$\sup_{x \in A_i} F(x) = \frac{2(i+1)-1}{2m}$$
$$\inf_{x \in A_i} F(x) = \frac{2i-1}{2m}$$

since F is a continuous function of x. □

Remark 5: Notice that finding $G(\mathbf{x}; \boldsymbol{\xi})$ that achieves the lower bound of $\frac{1}{2m}$ for multidimensional cases is complicated. Suboptimal approaches such as random sampling can be used for any dimension. Random sampling provides the distribution as $G_{\boldsymbol{\xi}} = \sum_{i=1}^{m} \frac{1}{m} \mathbb{1}_{\{\mathbf{z}:\mathbf{z} \geq \boldsymbol{\theta}_i\}}(\mathbf{x})$, where $\{\boldsymbol{\theta}_i : i = 1, 2, \dots, m\}$ are i.i.d. samples drawn from the target distribution F. These approaches are currently used by PFs to approximate complex posterior distributions. Glivenko-Cantelli Theorems [45] show that the accuracy of the approximations in terms of KS distance increases with the number of samples at a suboptimal rate on the order of $\frac{1}{\sqrt{m}}$.

In this section, we have presented algorithms for both continuous and discrete BC where we use mixtures of exponential families with KL divergence and discrete distributions with KS distance, respectively. These efficient algorithms yield cost-effective BCF techniques that are optimal in the sense described in Section II.

IV. RELATIONSHIP TO EXISTING TECHNIQUES

As described in Section II, any recursive filtering technique can be viewed as a mapping $\widehat{\Phi} : \mathcal{F}_{\mathrm{I}} \to \mathcal{F}_{\mathrm{O}}$. To discuss existing filtering techniques under the proposed framework, we list the families of distributions and the approaches used by representative techniques in Table I.[13]

The framework presented in Section II advocates the usage of filtering techniques $\widehat{\Phi}$ formed by the concatenation of the exact mapping Φ with a BC ψ, i.e., $\widehat{\Phi} = \psi \circ \Phi$. Note that BC is performed after the mapping Φ. This implies that BCF represents the complex posterior distribution by a tractable one only in the final stage of each recursive step (see Fig. 2(b)), thus minimizing the loss of information. Moreover, unlike most existing techniques,[14] the methodology of BCF allows the use of different families for inputs and outputs, thus adapting the accuracy versus complexity trade-off at each time step.[15]

The methodology of BCF can lead to different families of filtering techniques depending on the discrepancy function and family of distributions, as well as on the algorithm used to perform BC. Several existing filtering techniques can be cast as special cases of BCF: PFs use discrete distributions and random sampling; VM-based techniques use separable distributions and left sided KL minimization; and EP-based techniques use exponential families and right sided KL minimization. PFs correspond to BCF since they approximate the posterior distribution after the mapping Φ,[16] whereas VM- and EP-based techniques correspond to BCF in the case that they approximate the posterior distribution after the mapping Φ.

Note that the essential component for BCF is an algorithm to accurately represent complex distributions, referred to as BC. In this paper we also develop new algorithms for continuous and

[13]Probability distributions satisfying $f(\mathbf{x}_1, \mathbf{x}_2) = f(\mathbf{x}_1) f(\mathbf{x}_2)$.

[14]Certain PFs are an exception since they adapt the number of particles for different time steps [46], [47].

[15]These more general approaches are desirable in cases where $\widehat{f}_{k-1} \in \mathcal{F}_{\mathrm{I}}$, but $\Phi(\widehat{f}_{k-1})$ cannot be accurately approximated by a distribution in $\mathcal{F}_{\mathrm{I}}$ or, conversely, where $\Phi(\widehat{f}_{k-1})$ can be accurately approximated with a distribution in a simpler family than $\mathcal{F}_{\mathrm{I}}$.

[16]Note that $\Phi(\widehat{f}_{k-1})$ given by (16) is equivalent to the "empirical filtering density" given by (4) in [30].

discrete BC.[17] The former minimize right sided KL divergence using mixtures of exponential families generalizing EP, while the latter minimize KS distance using discrete distributions outperforming random sampling.

The discrepancy functions, families of distributions, and BC algorithms described above are by no means exhaustive, and the proposed methodology puts forward important research directions in two fronts. In the theoretical front, it is of interest to understand the dependence of filtering performance on discrepancy functions and families of distributions. In the algorithmic front, it is of interest to develop additional BC algorithms with guaranteed accuracy for general distributions.

V. Case Study: Navigation in Harsh Environments

In this section, we investigate the filtering problem for navigation in harsh environments as a case study to evaluate the performance of BCF. We consider a scenario in which an agent node obtains range measurements with respect to several anchor nodes as well as inertial measurements from an inertial measurement unit (IMU). In the following, we describe the dynamic and measurement models.

A. Dynamic Model

Let $\mathbf{p}_k$ and $\boldsymbol{\varphi}_k$ denote the position and orientation of the agent node at time t_k.[18] If $\mathbf{x}_k \in \mathbb{R}^n$ is the state vector formed by position, orientation, as well as several of their time derivatives at time t_k, the dynamic model can be written as

$$\mathbf{x}_{k+1} = \mathbf{F}_k \mathbf{x}_k + \mathbf{n}_{\mathrm{d},k} \tag{8}$$

where $\mathbf{F}_k$ is the state transition matrix and the error $\mathbf{n}_{\mathrm{d},k}$ is commonly modeled as a zero-mean Gaussian random vector (i.e., discrete Wiener process) [50].

B. Measurements Models

The set of measurements obtained by an agent at each time instant t_k form the vector $\mathbf{y}_k$. The relationship between the positional state vector and the measurements can be described by the likelihood $f(\mathbf{y}_k|\mathbf{x}_k)$. Here we focus on the case in which the agent obtains range and IMU measurements in harsh environments. In the following we describe realistic models for these measurements.

Range Measurements: A range measurement $y^{\mathrm{r}}_{k,i}$ at time t_k between the agent at position $\mathbf{p}_k$ and the anchor i at position $\mathbf{p}_i$ can be written as

$$y^{\mathrm{r}}_{k,i} = \|\mathbf{p}_k - \mathbf{p}_i\| + n^{\mathrm{r}}_{k,i} + b_{k,i} \tag{9}$$

where $n^{\mathrm{r}}_{k,i}$ is additive noise, and $b_{k,i}$ is the positive bias due to non-line-of-sight (NLOS) and multipath propagation.

Note that the relationship between the range measurements and the state vector is nonlinear. Moreover, the distribution $f(y^{\mathrm{r}}_{k,i}|\mathbf{x}_k)$ is generally non-Gaussian in harsh environments due to the existence of NLOS and multipath [51], [52].

IMU Measurements: An IMU device makes two kinds of measurements: the angular velocity $\mathbf{y}^{\omega}_k$ about the body frame and the force $\mathbf{y}^{f}_k$ [53]. Angular velocity measured by gyroscopes at time instant t_k is given by

$$\mathbf{y}^{\omega}_k = \boldsymbol{\omega}_k + \mathbf{n}^{\omega}_k \tag{10}$$

where $\boldsymbol{\omega}_k \in \mathbb{R}^3$ is the true angular velocity, and $\mathbf{n}^{\omega}_k \in \mathbb{R}^3$ is a noise vector.

The angular velocity $\boldsymbol{\omega}_k$ is related to the state vector by [49]

$$\boldsymbol{\omega}_k = \mathbf{C}_1^{\mathrm{T}}(\boldsymbol{\varphi}_k) \cdot \dot{\boldsymbol{\varphi}}_k \tag{11}$$

where $\mathbf{C}_1(\boldsymbol{\varphi}_k) \in \mathbb{R}^{3\times 3}$ is given by

$$\mathbf{C}_1(\boldsymbol{\varphi}_k) = \mathbf{I}_3 + \frac{1-\cos\|\boldsymbol{\varphi}_k\|}{\|\boldsymbol{\varphi}_k\|^2}[\boldsymbol{\varphi}_k]_\times + \frac{\|\boldsymbol{\varphi}_k\| - \sin\|\boldsymbol{\varphi}_k\|}{\|\boldsymbol{\varphi}_k\|^3}[\boldsymbol{\varphi}_k]^2_\times$$

in which $[\boldsymbol{\varphi}_k]_\times$ is the skew-symmetric form of the rotation vector $\boldsymbol{\varphi}_k = (\varphi_{k,x}, \varphi_{k,y}, \varphi_{k,z})$

$$[\boldsymbol{\varphi}_k]_\times = \begin{pmatrix} 0 & -\varphi_{k,z} & \varphi_{k,y} \\ \varphi_{k,z} & 0 & -\varphi_{k,x} \\ -\varphi_{k,y} & \varphi_{k,x} & 0 \end{pmatrix}.$$

Hence, the relationship between $\mathbf{y}^{\omega}_k$ and $\mathbf{x}_k$ is nonlinear.[19]

Similarly, the force measured by accelerometers at time instant t_k is given by

$$\mathbf{y}^{f}_k = \mathbf{f}_k + \mathbf{n}^{f}_k \tag{12}$$

where $\mathbf{f}_k \in \mathbb{R}^3$ is the force vector written in the body frame reference, and $\mathbf{n}^{f}_k \in \mathbb{R}^3$ is a noise vector. The force vector in the body frame $\mathbf{f}_k$ is related to the state vector by

$$\mathbf{f}_k = \mathbf{C}_2^{\mathrm{T}}(\boldsymbol{\varphi}_k)(\mathbf{a}_k - \mathbf{g}) \tag{13}$$

where $\mathbf{a}_k, \mathbf{g} \in \mathbb{R}^3$ are the acceleration and the gravity written in the fixed frame reference. Moreover, $\mathbf{C}_2(\boldsymbol{\varphi}_k) \in \mathbb{R}^{3\times 3}$ is given by the Rodrigues' rotation formula [48] as[20]

$$\mathbf{C}_2(\boldsymbol{\varphi}_k) = \mathbf{I}_3 + \frac{\sin\|\boldsymbol{\varphi}_k\|}{\|\boldsymbol{\varphi}_k\|}[\boldsymbol{\varphi}_k]_\times + \frac{1-\cos\|\boldsymbol{\varphi}_k\|}{\|\boldsymbol{\varphi}_k\|^2}[\boldsymbol{\varphi}_k]^2_\times.$$

Hence, the relationship between $\mathbf{y}^{f}_k$ and $\mathbf{x}_k$ is nonlinear.

More sophisticated models have been developed for range measurements and IMU inertial measurements [53]. Such models take also into account drifts and biases. In this paper we use the compact models described above for simplicity, while more advanced models can be analogously incorporated by adding the corresponding variables to the state vector.

VI. Performance Evaluation

In this section, we show the performance of the proposed continuous BCFs for the navigation example described in Section V

[17]Notice that BC can be performed simply by evaluating some characteristics of $\Phi(\widehat{f}_{k-1})$, e.g., statistical moments and percentiles, as shown in Section III.

[18]The orientation $\boldsymbol{\varphi}_k$ can be represented by a rotation vector, $\boldsymbol{\varphi}_k \in \mathbb{R}^3$ if $\mathbf{p}_k \in \mathbb{R}^3$ and $\boldsymbol{\varphi}_k \in \mathbb{R}$ if $\mathbf{p}_k \in \mathbb{R}^2$ [48], [49].

[19]Notice that the measurement model for angular velocity in 2D scenarios is linear.

[20]The matrix $\mathbf{C}_2(\boldsymbol{\varphi}_k)$ is the direction cosine matrix which transform the coordinates of vectors with respect to the body frame to those with respect to the fixed reference frame. Hence, it is a matrix formed from the rotation vector $\boldsymbol{\varphi}_k$ that represents the rotation like a linear transformation.

Algorithm 1: Continuous Belief Condensation Filter.

1: INITIALIZATION:

2: Set $\widehat{f}_0$ equal to the prior distribution of $\mathbf{x}_0$.

3: $\widehat{\Phi}$ IMPLEMENTATION:

4: **for** $k = 1, 2, \ldots$ **do**

5: (i) Choose a family of mixtures of exponential families $\mathcal{F}_{\Xi^{m_k}}$ and an initial distribution

$$g(\mathbf{x}_k; \boldsymbol{\xi}_k) = \sum_{i=1}^{m_k} \alpha_{k,i} g_i(\mathbf{x}_k; \boldsymbol{\theta}_{k,i}).$$

6: (ii) $\Phi(\widehat{f}_{k-1})$ CONDENSATION: Repeat until convergence,[21] for $i = 1, 2, \ldots, m_k$,

$$\alpha_{k,i} \leftarrow \alpha_{k,i} \cdot \mathbb{E}_{g_i(\mathbf{x}_k;\boldsymbol{\theta}_{k,i})} \left\{ \frac{\Phi(\widehat{f}_{k-1})}{g(\mathbf{x}_k; \boldsymbol{\xi}_k)} \right\}$$

$$\boldsymbol{\theta}_{k,i} \leftarrow \boldsymbol{\vartheta} \quad \text{with}$$

$$\mathbb{E}_{g_i(\mathbf{x}_k;\boldsymbol{\vartheta}))} \{\mathbf{t}_i(\mathbf{x}_k)\} = \frac{\mathbb{E}_{g_i(\mathbf{x}_k;\boldsymbol{\theta}_{k,i})} \left\{ \frac{\Phi(\widehat{f}_{k-1})}{g(\mathbf{x}_k;\boldsymbol{\xi}_k)} \mathbf{t}_i(\mathbf{x}_k) \right\}}{\mathbb{E}_{g_i(\mathbf{x}_k;\boldsymbol{\theta}_{k,i})} \left\{ \frac{\Phi(\widehat{f}_{k-1})}{g(\mathbf{x}_k;\boldsymbol{\xi}_k)} \right\}}$$

where

$$\Phi(\widehat{f}_{k-1}) \propto f(\mathbf{y}_k|\mathbf{x}_k) \times \int f(\mathbf{x}_k|\mathbf{x}_{k-1}) \widehat{f}_{k-1}(\mathbf{x}_{k-1}) d\mathbf{x}_{k-1} \quad (14)$$

7: (iii) Approximate $f(\mathbf{x}_k|\mathbf{y}_{1:k})$ as

$$\widehat{f}_k = g(\mathbf{x}_k; \boldsymbol{\xi}_k) = \sum_{i=1}^{m_k} \alpha_{k,i} g_i(\mathbf{x}_k; \boldsymbol{\theta}_{k,i})$$

8: **end for**

and that of discrete BCFs for a widely explored 1-D filtering problem [29], [36].

A. Continuous BCF

In the following, we illustrate the performance of the proposed continuous BCF in the navigation filtering problem through simulations with measurements emulating sensors' behavior in harsh propagation environments. We considered a scenario in which one agent moves in the horizontal plane and obtains both GPS and IMU measurements. In this example, the state vector is eight dimensional and is given by $\mathbf{x}_k = [\mathbf{p}_k, \mathbf{v}_k, \mathbf{a}_k, \varphi_k, \dot{\varphi}_k] \in \mathbb{R}^8$, where $\mathbf{p}_k \in \mathbb{R}^2$, $\mathbf{v}_k \in \mathbb{R}^2$, $\mathbf{a}_k \in \mathbb{R}^2$, φ_k, and $\dot{\varphi}_k$ are the position, velocity, acceleration, orientation, and derivate of orientation at time t_k, respectively. The dynamic model is given by (8) with matrix

$$\mathbf{F}_k = \begin{pmatrix} \mathbf{I}_2 & \Delta_k \mathbf{I}_2 & \frac{\Delta_k^2}{2} \mathbf{I}_2 & & \\ & \mathbf{I}_2 & \Delta_k \mathbf{I}_2 & & \\ & & \mathbf{I}_2 & & \\ & & & 1 & \Delta_k \\ & & & & 1 \end{pmatrix}$$

[21]The simulation results in this paper are obtained after 3 iterations.

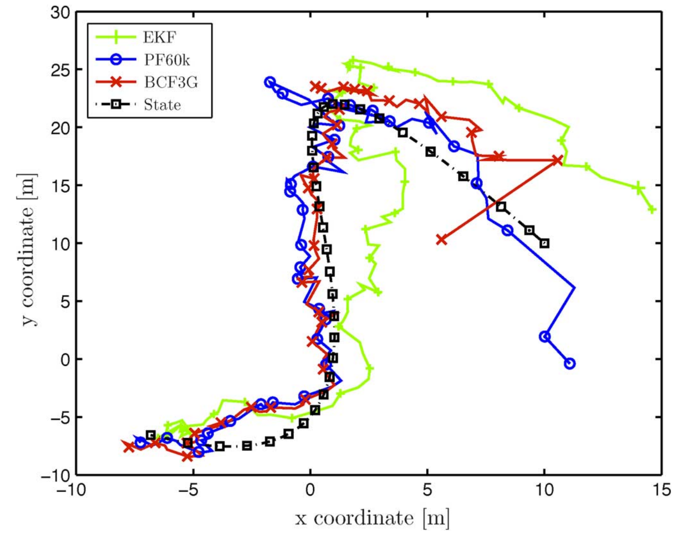


Fig. 3. Comparison of EKF, PF, and proposed continuous BCF in one instantiation.

where $\Delta_k = t_{k+1} - t_k$. The measurements model is given by (9)–(13), where the first two components of the rotation vector are zero for 2-D navigation in the horizontal plane.

We simulated range measurements from 4 GPS satellites in NLOS conditions. The additive noise associated with these measurements is modeled as a zero mean Gaussian RV with standard deviation of 2 m, while the positive bias introduced by the NLOS propagation is modeled as an exponential RV with mean 6 m, 8 m, 10 m, and 14 m for each satellite. The errors in the force and angular velocity measurements made by IMU are modeled as zero mean Gaussian RVs with standard deviations of 0.07 N and 0.02 rad/sec, respectively. The motion of the agent was simulated in 100 positions as shown in Fig. 3 with a mean and maximum velocity of 1.02 m/sec and 2.6 m/sec, respectively; mean and maximum acceleration of 0.122 $\mathrm{m/sec}^2$ and 0.267 $\mathrm{m/sec}^2$, respectively; and mean and maximum angular velocity of 0.04 rad/sec and 0.077 rad/sec, respectively.

The positional state of the agent is estimated using the proposed BCF with one Gaussian distribution (BCF1G) and mixtures of three Gaussian distributions (BCF3G), as well as commonly used techniques: EKF, UKF, SIR PF (Algorithm 4 in [33]), and APF (Section D.2 in [31]). Furthermore, we evaluated the Cramer-Rao lower bound (CRLB) as a theoretical benchmark [3], [54] and compared the root mean squared error (RMSE) of the BCF with the bound. We then quantified the performance of the proposed BCF both in terms of accuracy and complexity, comparing such performances to the above representative existing nonlinear/non-Gaussian filtering techniques.

Algorithm 1 summarizes the key steps needed for the implementation of continuous BC filters. We obtain the parameters of the Gaussian mixtures using the recursions (4), (6), and (7).[22] Note that in this example, the integrands in (4), (6), and (7) can be easily evaluated because a function proportional to $\Phi(\widehat{f}_{k-1})$ in (14) can be obtained in a closed form at each time step k since

[22]We use integration rules specific for Gaussian weights as in [27], [44].

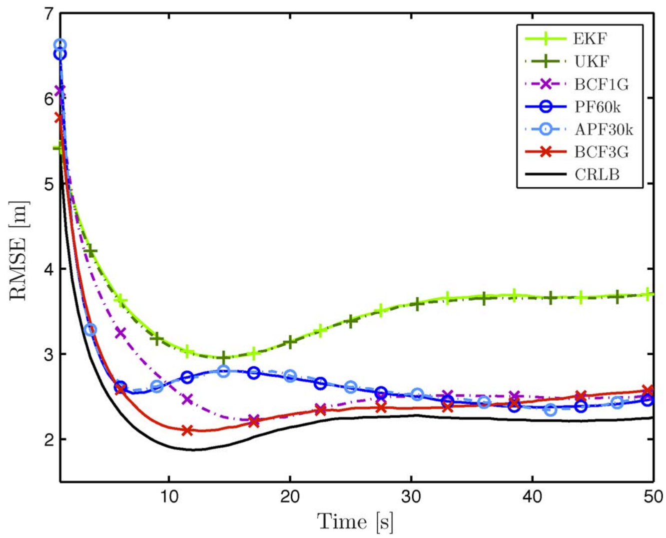


Fig. 4. Comparison of the CRLB and the RMSEs for EKF, PF, and proposed continuous BCFs.

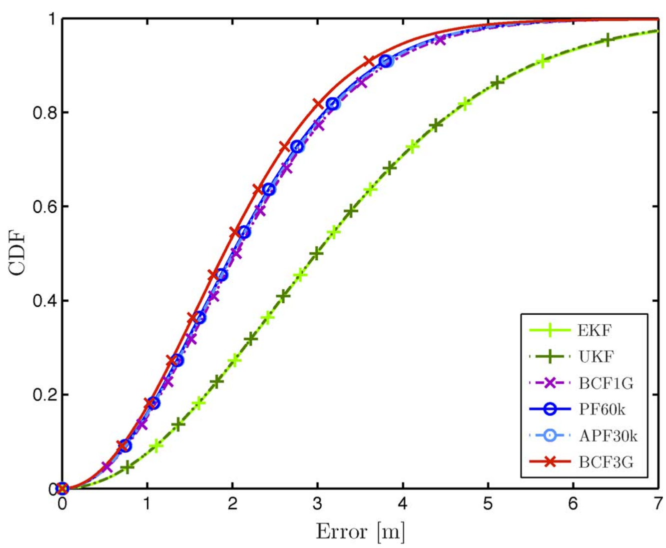


Fig. 5. Comparison of the CDFs for EKF, PF, and proposed continuous BCFs.

the dynamic model in (8) is linear and Gaussian, and the families of distributions for BC are mixtures of Gaussians.[23]

The positions inferred by different filters are shown in Fig. 3 for one specific instantiation. Figure 4 shows the RMSE obtained by the different filters for each time step based on Monte Carlo simulation with 16,000 instantiations, as well as the CRLB. We can observe that the proposed BCF can achieve remarkable RMSE performance in comparison with representative existing techniques and the CRLB. In addition, Fig. 5 shows the empirical CDFs of the positional errors and the overall RMSE are given in Table II.[24] In these simulation results, we observe that BCF1G outperforms other techniques that use Gaussian distributions[25] and obtains accuracies comparable to PFs that use a considerable

TABLE II
ACCURACY AND COMPLEXITY COMPARISON OF CURRENT FILTERING TECHNIQUES WITH CONTINUOUS BCF IN TERMS OF RMSE, PROCESSING TIME, AND REQUIRED MEMORY

Technique	RMSE [m]	Time [s]	Memory [Reals]
EKF	3.63	2.4×10^{-4}	44
UKF	3.62	1.65×10^{-3}	44
PF30k	2.61	11.46	2.4×10^5
PF60k	2.44	26.72	4.8×10^5
APF15k	2.61	17.99	1.2×10^5
APF30k	2.45	36.56	2.4×10^5
BCF1G	2.49	9.85×10^{-3}	44
BCF3G	2.44	1.27	134

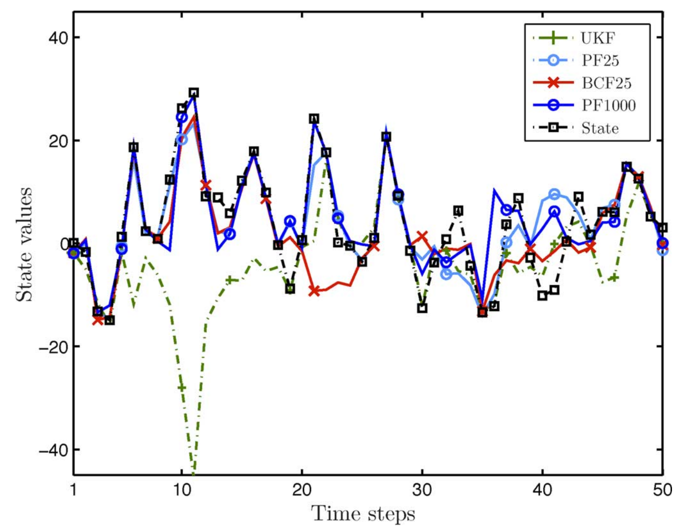


Fig. 6. Comparison of EKF, PF, and proposed discrete BCF in one instantiation.

number of particles (30k particles for PF30k and 15k particles for APF15k), whereas BCF3G obtains accuracies close to the CRLB and comparable with PFs that use a large number of particles (60k particles for PF60k and 30k particles for APF30k).

Finally, the processing time and memory used by each filtering technique are shown in Table II. One can observe that the complexity, both in terms of processing time and memory, of the proposed BCFs is several orders of magnitude smaller than that of PFs with a similar level of accuracy.[26]

B. Discrete BCF

We next evaluate the performance of proposed discrete BCF for a widely explored 1-D filtering problem [29], [36]. Specifically, the dynamic and measurement models for this example are given by

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

$$y_k = \frac{x_k^2}{20} + w_k$$

[23]In this example, mixtures of Gaussian distributions are suitable due to the conjugacy property of Gaussians with respect to the dynamic model. In other filtering problems, different exponential families may be more suitable depending on the system models.

[24]These results are obtained from the errors after 25 seconds of filtering.

[25]Notice that although BCF1G, EKF, and UKF all use Gaussian distributions, BCF1G does not approximate the KF equations but performs $\psi \circ \Phi$.

[26]To achieve the same level of accuracy as continuous BCF, PFs need to use a large number of particles and hence require a higher computational complexity even though the computation of each particle is simple.

Algorithm 2: Discrete Belief Condensation Filter.

1: INITIALIZATION:

2: Set $\widehat{F}_0$ equal to a prior distribution of $\mathbf{y}_0$.

3: $\widehat{\Phi}$ IMPLEMENTATION:

4: **for** $k = 1, 2, \ldots$ **do**

5: (i) Choose a number of discrete points m_k

6: (ii) $\Phi(\widehat{F}_{k-1})$ CONDENSATION: For $i = 1, 2, \ldots, m_k$, obtain $\theta_{k,i}$ as the value satisfying

$$\int_{-\infty}^{\theta_{k,i}} \Phi(\widehat{f}_{k-1}) dx_k = \frac{2i-1}{2m_k} \tag{15}$$

where

$$\Phi(\widehat{f}_{k-1}) \propto f(\mathbf{y}_k|x_k) \times \sum_{i=1}^{m_{k-1}} f(x_k|x_{k-1} = \theta_{k-1,i}) \tag{16}$$

7: (iii) Approximate $F(\mathbf{x}_k|\mathbf{y}_{1:k})$ as

$$\widehat{F}_k = \sum_{i=1}^{m_k} \frac{1}{m_k} \mathbb{1}_{\{z: z \geq \theta_{k,i}\}}(x)$$

8: **end for**

where v_k follows a Gaussian distribution with mean 0 and variance $\sigma_v^2 = 10$, and w_k follows a Gaussian distribution with mean 0 and variance $\sigma_w^2 = 1$. Notice that this filtering example is highly nonlinear and the measurements are not affected by changes in the sign of the state.[27]

We estimated the sequence of states $\{x_k\}_{k=1}^{50}$ using the proposed discrete BCF, as well as UKF and SIR PF. We utilized different number of particles or support points, specifically $m = 25, 100, 500$ and 1000. We then quantified the performance of the proposed BCF both in terms of accuracy and complexity, and compared such performance with those of UKF and SIR PF.

Algorithm 2 summarizes the key steps needed for the implementation of discrete BC filters, where the only non-straightforward step is the evaluation of the quantile function in (15). This equation can be solved numerically under accuracy and complexity tradeoff, where the following results were obtained using Monte Carlo integration.

The states inferred by different filters are shown in Fig. 6 for one specific instantiation, where it can be observed that UKF loses track when the state change its sign due to the ambiguity in the measurement model and the unimodality of Gaussian distributions. Figure 7 shows the RMSE obtained by the different filters for each time step based on Monte Carlo simulation with 30,000 instantiations. We can observe that the proposed BCF can achieve remarkable RMSE performance with less than 100 support points. In addition, Fig. 8 shows the empirical CDFs of the errors and the overall RMSE are given in Table III. From these results, we observe that the accuracies of the proposed discrete BCFs are comparable to PFs despite the fact that the discrete BCFs require a much smaller number of support points than that of PFs.

Finally, the processing time and memory used by each filtering technique are shown in Table III. Similarly to the continuous case, from this table one can observe that the proposed discrete BCFs require a smaller complexity than the one needed by SIR PFs with a similar level of accuracy.

[27]In this example the CRLB is not a meaningful performance metric due to the ambiguity induced from the measurement model.

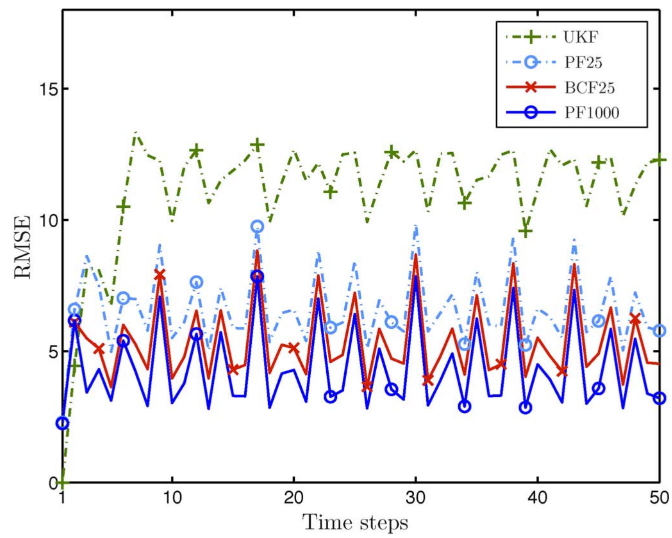


Fig. 7. Comparison of the RMSEs for UKF, PF, and proposed discrete BCF.

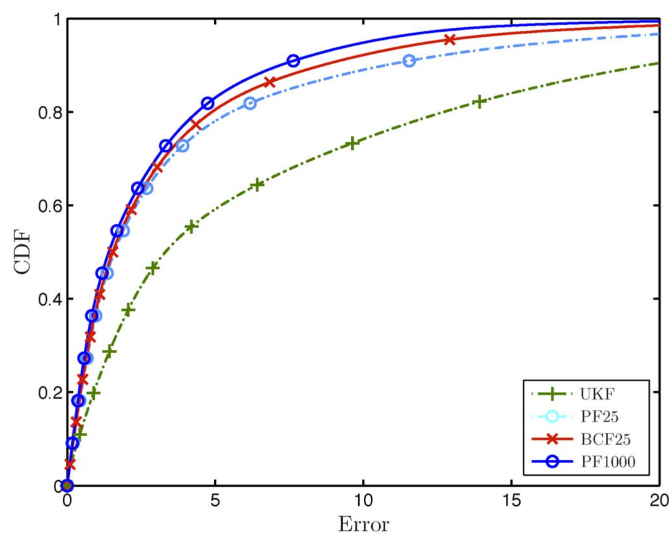


Fig. 8. Comparison of the CDFs for UKF, PF, and proposed discrete BCF.

TABLE III
ACCURACY AND COMPLEXITY COMPARISON OF CURRENT FILTERING TECHNIQUES WITH DISCRETE BCF IN TERMS OF RMSE, PROCESSING TIME, AND REQUIRED MEMORY

Technique	RMSE	Time [s]	Memory [Reals]
UKF	11.34	2.97×10^{-2}	2
PF25	6.74	3.60×10^{-2}	25
BCF25	5.50	4.97×10^{-2}	25
PF100	5.06	6.97×10^{-2}	100
BCF100	4.78	9.84×10^{-2}	100
PF500	4.63	1.55×10^{-1}	500
BCF500	4.57	2.19×10^{-1}	500
PF1000	4.58	2.18×10^{-1}	1000
BCF1000	4.55	3.10×10^{-1}	1000

VII. Conclusion

In this paper, we proposed a new methodology for nonlinear/non-Gaussian filtering called BCF. We first established a general framework for filtering, based on which we formulated an optimality criterion leading to BCF. Moreover, we developed efficient continuous and discrete BC algorithms to condense the complex distributions arising in the filtering process. We compared the accuracy and complexity of the proposed BCFs with representative existing techniques for the filtering problem in navigation as well as for a highly nonlinear problem widely explored in the literature. Our results suggest the advantages of the BCF methodology for broad filtering problems and show that the proposed techniques can obtain accuracies comparable to those of PFs, but with complexities much smaller than PFs. This paper provides a new methodology for the design and analysis of nonlinear/non-Gaussian filters that can be employed in a wide range of modern applications.

Appendix A
Proof of Theorem 1

Proof: For a given $\breve{\boldsymbol{\xi}} = (\breve{\alpha}_1, \breve{\boldsymbol{\theta}}_1, \breve{\alpha}_2, \breve{\boldsymbol{\theta}}_2, \ldots, \breve{\alpha}_m, \breve{\boldsymbol{\theta}}_m) \in \Xi^m$, we show that a parameter $\boldsymbol{\xi} \in \Xi^m$ with $D_{\mathrm{KL}}(f, g_{\boldsymbol{\xi}}) \leq D_{\mathrm{KL}}(f, g_{\breve{\boldsymbol{\xi}}})$ can be obtained as the solution of convex optimization problems by decomposing $D_{\mathrm{KL}}(f, g_{\boldsymbol{\xi}})$ and using Gibb's inequality [55].

We first obtain the decomposition of $D_{\mathrm{KL}}(f, g_{\boldsymbol{\xi}})$. It is clear that for $i = 1, 2, \ldots, m$

$$\log g(\mathbf{x}; \boldsymbol{\xi}) = \log\left(\alpha_i g_i(\mathbf{x}; \boldsymbol{\theta}_i)\right) - \log\left(\frac{\alpha_i g_i(\mathbf{x}; \boldsymbol{\theta}_i)}{g(\mathbf{x}; \boldsymbol{\xi})}\right).$$

Together with the fact that $\sum_{i=1}^m \frac{\breve{\alpha}_i g_i(\mathbf{x}; \breve{\boldsymbol{\theta}}_i)}{g(\mathbf{x}; \breve{\boldsymbol{\xi}})} = 1, \forall \mathbf{x} \in \mathbb{R}^n$ and $\breve{\boldsymbol{\xi}} \in \Xi^m$, we have that

$$\begin{aligned}\log g(\mathbf{x}; \boldsymbol{\xi}) &= \sum_{i=1}^m \log g(\mathbf{x}; \boldsymbol{\xi}) \frac{\breve{\alpha}_i g_i(\mathbf{x}; \breve{\boldsymbol{\theta}}_i)}{g(\mathbf{x}; \breve{\boldsymbol{\xi}})} \\ &= -Q(\mathbf{x}, \boldsymbol{\xi}, \breve{\boldsymbol{\xi}}) - H(\mathbf{x}, \boldsymbol{\xi}, \breve{\boldsymbol{\xi}})\end{aligned}$$

where

$$\begin{aligned}Q(\mathbf{x}, \boldsymbol{\xi}, \breve{\boldsymbol{\xi}}) &= -\sum_{i=1}^m \log(\alpha_i g_i(\mathbf{x}; \boldsymbol{\theta}_i)) \frac{\breve{\alpha}_i g_i(\mathbf{x}; \breve{\boldsymbol{\theta}}_i)}{g(\mathbf{x}; \breve{\boldsymbol{\xi}})} \\ H(\mathbf{x}, \boldsymbol{\xi}, \breve{\boldsymbol{\xi}}) &= \sum_{i=1}^m \log\left(\frac{\alpha_i g_i(\mathbf{x}; \boldsymbol{\theta}_i)}{g(\mathbf{x}; \boldsymbol{\xi})}\right) \frac{\breve{\alpha}_i g_i(\mathbf{x}; \breve{\boldsymbol{\theta}}_i)}{g(\mathbf{x}; \breve{\boldsymbol{\xi}})}.\end{aligned}$$

Taking the expectation over $f(\mathbf{x})$ in the above equation, we obtain

$$\begin{aligned}D_{\mathrm{KL}}\left(f(\mathbf{x}), g(\mathbf{x}; \boldsymbol{\xi})\right) + h\left(f(\mathbf{x})\right) = \mathbb{E}_{f(\mathbf{x})}\left\{Q(\mathbf{x}, \boldsymbol{\xi}, \breve{\boldsymbol{\xi}})\right\} \\ + \mathbb{E}_{f(\mathbf{x})}\left\{H(\mathbf{x}, \boldsymbol{\xi}, \breve{\boldsymbol{\xi}})\right\}\end{aligned}$$

where $h(f(\mathbf{x}))$ is the differential entropy of $f(\mathbf{x})$, which is finite by assumption. Therefore, in order to find a parameter $\boldsymbol{\xi} \in \Xi^m$ such that $D_{\mathrm{KL}}(f, g_{\boldsymbol{\xi}}) \leq D_{\mathrm{KL}}(f, g_{\breve{\boldsymbol{\xi}}})$, it is sufficient that

$$\mathbb{E}_{f(\mathbf{x})}\left\{Q(\mathbf{x}, \boldsymbol{\xi}, \breve{\boldsymbol{\xi}})\right\} \leq \mathbb{E}_{f(\mathbf{x})}\left\{Q(\mathbf{x}, \breve{\boldsymbol{\xi}}, \breve{\boldsymbol{\xi}})\right\}$$

since $H(\mathbf{x}, \boldsymbol{\xi}, \breve{\boldsymbol{\xi}}) \leq H(\mathbf{x}, \breve{\boldsymbol{\xi}}, \breve{\boldsymbol{\xi}}), \forall \boldsymbol{\xi} \in \Xi^m$ by Gibbs' inequality [55].

In the remaining of the proof, we show that for any $\breve{\boldsymbol{\xi}}$, the value $\boldsymbol{\xi}$ that minimizes $\mathbb{E}_{f(\mathbf{x})}\{Q(\mathbf{x}, \boldsymbol{\xi}, \breve{\boldsymbol{\xi}})\}$ can be obtained by convex optimization, leading to the recursion given in the theorem.

Note that

$$\begin{aligned}&\mathbb{E}_{f(\mathbf{x})}\left\{Q(\mathbf{x}, \boldsymbol{\xi}, \breve{\boldsymbol{\xi}})\right\} \\ &= -\mathbb{E}_{f(\mathbf{x})}\left\{\sum_{i=1}^m \log\left(\alpha_i g_i(\mathbf{x}; \boldsymbol{\theta}_i)\right)\left(\frac{\breve{\alpha}_i g_i(\mathbf{x}; \breve{\boldsymbol{\theta}}_i)}{g(\mathbf{x}; \breve{\boldsymbol{\xi}})}\right)\right\} \\ &= -\mathbb{E}_{f(\mathbf{x})}\left\{\sum_{i=1}^m \log \alpha_i \left(\frac{\breve{\alpha}_i g_i(\mathbf{x}; \breve{\boldsymbol{\theta}}_i)}{g(\mathbf{x}; \breve{\boldsymbol{\xi}})}\right)\right\} \\ &\quad - \mathbb{E}_{f(\mathbf{x})}\left\{\sum_{i=1}^m \log\left(g_i(\mathbf{x}; \boldsymbol{\theta}_i)\right)\left(\frac{\breve{\alpha}_i g_i(\mathbf{x}; \breve{\boldsymbol{\theta}}_i)}{g(\mathbf{x}; \breve{\boldsymbol{\xi}})}\right)\right\}. \quad (17)\end{aligned}$$

It can be observed that (17) is the sum of convex functions in $\boldsymbol{\xi}$ because $\log(\cdot)$ is a concave function, and each $g_i(\mathbf{x}; \boldsymbol{\theta}_i)$ is logconcave in $\boldsymbol{\theta}_i$ since they are members of exponential families and $\boldsymbol{\theta}_i$ are natural parameters [42]. Moreover, the sets $\{(\alpha_1, \alpha_2, \ldots, \alpha_m) \in \mathbb{R}^m : \sum_{i=1}^m \alpha_i = 1 \text{ and } \alpha_i \geq 0, \forall i\}$ and Θ_i are convex [42].

Since each parameter α_i and $\boldsymbol{\theta}_i$ for $i = 1, 2, \ldots, m$ appears in a different term, the parameter $\boldsymbol{\xi}$ minimizing (17) can be obtained by separately minimizing over $(\alpha_1, \alpha_2, \ldots, \alpha_m)$, and $\boldsymbol{\theta}_i$ for $i = 1, 2, \ldots, m$. The Lagrangian of the optimization problem corresponding to the first term in the right-hand side of (17) is

$$-\mathbb{E}_{f(\mathbf{x})}\left\{\sum_{i=1}^m \log \alpha_i \frac{\breve{\alpha}_i g_i(\mathbf{x}; \breve{\boldsymbol{\theta}}_i)}{g(\mathbf{x}; \breve{\boldsymbol{\xi}})}\right\} - \sum_{i=1}^m \lambda_i \alpha_i + \mu\left(\sum_{i=1}^m \alpha_i - 1\right).$$

Recall that the primal is a convex problem with an affine constraint, and

$$\begin{cases}\lambda_i = 0, & i = 1, 2, \ldots, m \\ \alpha_i = \breve{\alpha}_i \cdot \mathbb{E}_{f(\mathbf{x})}\left\{\frac{g_i(\mathbf{x}; \breve{\boldsymbol{\theta}}_i)}{g(\mathbf{x}; \breve{\boldsymbol{\xi}})}\right\}, & i = 1, 2, \ldots, m \\ \mu = 1 & \end{cases}$$

satisfies the Karush-Kuhn-Tucker conditions. Then, we obtain (4) as the solution to the convex problem [56], where one can check that

$$\sum_{i=1}^m \alpha_i = \mathbb{E}_{f(\mathbf{x})}\left\{\sum_{i=1}^m \frac{\breve{\alpha}_i g_i(\mathbf{x}; \breve{\boldsymbol{\theta}}_i)}{g(\mathbf{x}; \breve{\boldsymbol{\xi}})}\right\} = 1$$

since $g(\mathbf{x}; \breve{\boldsymbol{\xi}}) = \sum_{i=1}^m \breve{\alpha}_i g_i(\mathbf{x}; \breve{\boldsymbol{\theta}}_i)$.

For the optimization problems corresponding to the second term, it can be proved that

$$\mathbb{E}_{f(\mathbf{x})}\left\{\sum_{i=1}^m \log g_i(\mathbf{x}; \boldsymbol{\theta}_i) \frac{\breve{\alpha}_i g_i(\mathbf{x}; \breve{\boldsymbol{\theta}}_i)}{g(\mathbf{x}; \breve{\boldsymbol{\xi}})}\right\} \quad (18)$$

is differentiable with respect to $\boldsymbol{\theta}_i$ and its derivative is

$$\breve{\alpha}_i \cdot \mathbb{E}_{g_i(\mathbf{x}; \breve{\boldsymbol{\theta}}_i)}\left\{\frac{f(\mathbf{x})}{g(\mathbf{x}; \breve{\boldsymbol{\xi}})} \mathbf{t}_i(\mathbf{x})\right\} - \breve{\alpha}_i \cdot \frac{\partial A_i}{\partial \boldsymbol{\theta}_i} \mathbb{E}_{g_i(\mathbf{x}; \breve{\boldsymbol{\theta}}_i)}\left\{\frac{f(\mathbf{x})}{g(\mathbf{x}; \breve{\boldsymbol{\xi}})}\right\}.$$

Then, we can obtain (5) by using the fact that $\frac{\partial A_i}{\partial \boldsymbol{\theta}_i} = \mathbb{E}_{g_i(\mathbf{x}; \boldsymbol{\theta}_i)}\{\mathbf{t}_i(\mathbf{x})\}$ [42].

We have shown that $\{D_{\rm KL}(f, g_{\boldsymbol{\xi}^{[l]}})\}_{l\in\mathbb{Z}_+}$, with $\{g_{\boldsymbol{\xi}^{[l]}}\}_{l\in\mathbb{Z}_+}$ given by (4) and (5), is monotonically decreasing and hence converging. Finally, the result for the case $m=1$ can be directly obtained observing that in this case $D_{\rm KL}(f, g_{\boldsymbol{\theta}})$ is convex in $\boldsymbol{\theta}$ since $H(\mathbf{x}, \boldsymbol{\xi}, \breve{\boldsymbol{\xi}})$ vanishes. □

Appendix B
Proof of Proposition 3

Proof: Let $G_{\boldsymbol{\xi}}$ be a distribution in $\mathcal{F}_{\Xi^m}$ with $\boldsymbol{\xi} = (\alpha_1, \boldsymbol{\theta}_1, \alpha_2, \boldsymbol{\theta}_2, \ldots, \alpha_m, \boldsymbol{\theta}_m)$. Without loss of generality we assume that $F(\boldsymbol{\theta}_i) \leq F(\boldsymbol{\theta}_j)$ for $i \leq j$, since the components of $\boldsymbol{\xi}$ can be rearranged to satisfy this condition.

Now we let

$$\delta_i = \begin{cases} F(\boldsymbol{\theta}_1), & i=1 \\ \frac{(F(\boldsymbol{\theta}_i)-F(\boldsymbol{\theta}_{i-1}))}{2}, & i=2,3,\ldots,m \\ 1-F(\boldsymbol{\theta}_m), & i=m+1 \end{cases}.$$

Note that at least one δ_i is larger than or equal to $1/2m$ since $\delta_1 + 2\sum_{i=2}^{m}\delta_i + \delta_{m+1} = 1$. Let $i_0 \in \{1, 2, \ldots, m+1\}$ with $\delta_{i_0} \geq 1/2m$, and A be the set

$$A = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq \boldsymbol{\theta}_i \quad \forall i \leq i_0, \mathbf{x} \ngeq \boldsymbol{\theta}_j \quad \forall j > i_0\}$$

we have that

$$D_{\rm KS}(F, G_{\boldsymbol{\xi}}) \geq \max_{\mathbf{x}\in A} |F(\mathbf{x}) - G_{\boldsymbol{\xi}}(\mathbf{x})|.$$

Calling $M_1 = \min_{\mathbf{x}\in A} F(\mathbf{x})$ and $M_2 = \max_{\mathbf{x}\in A} F(\mathbf{x})$, since F is monotonically increasing and continuous, we have that

(1) if $i_0 = 1$, since $G_{\boldsymbol{\xi}}(\mathbf{x}) = 0, \forall \mathbf{x} \in A$,

$$D_{\rm KS}(F, G_{\boldsymbol{\xi}}) \geq M_2 = F(\boldsymbol{\theta}_1) \geq \frac{1}{2m}$$

(2) if $i_0 = 2, \ldots m$, since $G_{\boldsymbol{\xi}}(\mathbf{x}_1) = G_{\boldsymbol{\xi}}(\mathbf{x}_2), \forall \mathbf{x}_1, \mathbf{x}_2 \in A$,

$$D_{\rm KS}(F, G_{\boldsymbol{\xi}}) \geq \frac{M_2 - M_1}{2} = \frac{F(\boldsymbol{\theta}_{i_0+1}) - F(\boldsymbol{\theta}_{i_0})}{2} \geq \frac{1}{2m}$$

(3) if $i_0 = m+1$, since $G_{\boldsymbol{\xi}}(\mathbf{x}) = 1, \forall \mathbf{x} \in A$,

$$D_{\rm KS}(F, G_{\boldsymbol{\xi}}) \geq 1 - M_1 = 1 - F(\boldsymbol{\theta}_m) \geq \frac{1}{2m}.$$

Therefore, in all cases $D_{\rm KS}(F, G_{\boldsymbol{\xi}}) \geq \frac{1}{2m}$. □

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments. They would also like to thank B. Appleby for drawing their attention to the modern applications of nonlinear and non-Gaussian filtering.

References

[1] M. Z. Win, A. Conti, S. Mazuelas, Y. Shen, W. M. Gifford, D. Dardari, and M. Chiani, "Network localization and navigation via cooperation," *IEEE Commun. Mag.*, vol. 49, no. 5, pp. 56–62, May 2011.

[2] B. Ristic, M. S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Norwood, MA, USA: Artech House, 2004.

[3] Y. Shen, S. Mazuelas, and M. Z. Win, "Network navigation: Theory and interpretation," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 9, pp. 1823–1834, Oct. 2012.

[4] R. A. Iltis, "Density function approximation using reduced sufficient statistics for joint estimation of linear and nonlinear parameters," *IEEE Trans. Signal Process.*, vol. 47, no. 8, pp. 2089–2099, Aug. 1999.

[5] K. J. Kim, M.-O. Pun, and R. A. Iltis, "Joint carrier frequency offset and channel estimation for uplink MIMO-OFDMA systems using parallel Schmidt Rao-Blackwellized particle filters," *IEEE Trans. Commun.*, vol. 58, no. 9, pp. 2697–2708, Sep. 2010.

[6] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[7] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[8] C.-J. Kim and C. R. Nelson, *State-space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications*. Cambridge, MA, USA: MIT Press, 1999.

[9] H. F. Lopes and R. S. Tsay, "Particle filters and Bayesian inference in financial econometrics," *J. Forecast.*, vol. 30, pp. 168–209, Jul. 2011.

[10] Y.-F. Huang, S. Werner, J. Huang, N. Kashyap, and V. Gupta, "State estimation in electric power grids," *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 33–43, Sep. 2012.

[11] E. Ghahremani and I. Kamwa, "Dynamic state estimation in power system by applying the extended Kalman filter with unknown inputs to phasor measurements," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 2556–2566, Nov. 2011.

[12] O. Cappe, E. Moulines, and T. Ryden, *Inference in Hidden Markov Models*, ser. Springer Series in Statistics. New York, NY, USA: Springer, 2007.

[13] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004.

[14] F. Daum, "Nonlinear filters: Beyond the Kalman filter," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 20, no. 8, pp. 57–69, Aug. 2005.

[15] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forsell, J. Jansson, R. Karlsson, and P. J. Nordlund, "Particle filters for positioning, navigation and tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 425–437, Feb. 2002.

[16] H. Wymeersch, J. Lien, and M. Z. Win, "Cooperative localization in wireless networks," *Proc. IEEE*, vol. 97, no. 2, pp. 427–450, Feb. 2009.

[17] K. L. Anderson and R. A. Iltis, "A distributed bearings-only tracking algorithm using reduced sufficient statistics," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 1, pp. 339–349, Jan. 1996.

[18] R. Verdone, D. Dardari, G. Mazzini, and A. Conti, *Wireless Sensor and Actuator Networks: Technologies, Analysis and Design*. Amsterdam, The Netherlands: Elsevier, 2008.

[19] U. A. Khan, S. Kar, and J. M. F. Moura, "Distributed sensor localization in random environments using minimal number of anchor nodes," *IEEE Trans. Signal Process.*, vol. 57, no. 5, pp. 2000–2016, May 2009.

[20] Y. Shen and M. Z. Win, "Fundamental limits of wideband localization—Part I: A general framework," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4956–4980, Oct. 2010.

[21] Y. Shen, H. Wymeersch, and M. Z. Win, "Fundamental limits of wideband localization—Part II: Cooperative networks," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4981–5000, Oct. 2010.

[22] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME J. Basic Eng.*, vol. 82, pp. 34–45, 1960.

[23] F. Daum, "Exact finite-dimensional nonlinear filters," *IEEE Trans. Autom. Control*, vol. 31, no. 7, pp. 616–622, Jul. 1986.

[24] R. S. Bucy and K. D. Senne, "Digital synthesis of non-linear filters," *Automatica*, vol. 7, no. 3, pp. 287–298, May 1971.

[25] H. W. Sorenson, "On the development of practical nonlinear filters," *Inf. Sci.*, vol. 7, pp. 253–270, 1974.

[26] V. Smidl and A. Quinn, "Variational bayesian filtering," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5020–5030, Oct. 2008.

[27] I. Arasaratnam, S. Haykin, and T. R. Hurd, "Cubature Kalman filtering for continuous-discrete systems: Theory and simulations," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 4977–4993, Oct. 2010.

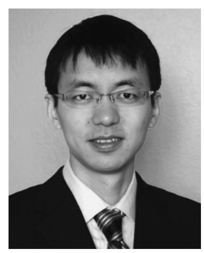
[28] I. Arasaratnam, S. Haykin, and R. J. Elliot, "Discrete-time nonlinear filtering algorithms using Gauss-Hermite quadrature," *Proc. IEEE*, vol. 95, no. 5, pp. 953–977, May 2007.

[29] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *Proc. Inst. Electr. Eng.—F*, vol. 140, no. 2, pp. 107–113, 1993.
[30] M. K. Pitt and N. Shephard, "Filtering via simulation: Auxiliary particle filters," *J. Amer. Statist. Assoc.*, vol. 94, no. 446, pp. 590–599, Jun. 1999.
[31] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for bayesian filtering," *Stat. Comput.*, vol. 10, no. 3, pp. 197–208, 2000.
[32] S. Saha, P. K. Mandal, Y. Boers, and H. Driessen, "Gaussian proposal density using moment matching in SMC," *Stat. Comput.*, vol. 19, no. 2, pp. 203–208, Jun. 2009.
[33] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.
[34] J. H. Kotecha and P. M. Djuric, "Gaussian sum particle filtering," *IEEE Trans. Signal Process.*, vol. 51, no. 10, pp. 2602–1612, Oct. 2003.
[35] B. D. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1979.
[36] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*. New York, NY, USA: Springer, 2001.
[37] D. L. Alspach and H. W. Sorenson, "Nonlinear Bayesian estimation using Gaussian sum approximations," *IEEE Trans. Autom. Control*, vol. 17, no. 4, pp. 439–448, Aug. 1972.
[38] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. 17th Annual Conf. Uncertainty in Artif. Intell.*, Seattle, WA, USA, Aug. 2001, pp. 362–369.
[39] R. E. Kalman, P. L. Falb, and M. A. Arbib, *Topics in Mathematical System Theory*. New York, NY, USA: McGraw-Hill, 1969.
[40] S. K. Mitter, "Filtering and stochastic control: A historical perspective," *IEEE Control Syst.*, vol. 16, no. 3, pp. 67–76, Jun. 1996.
[41] T. Kailath, "A view of three decades of linear filtering theory," *IEEE Trans. Inf. Theory*, vol. 20, no. 2, pp. 146–181, Mar. 1974.
[42] P. J. Bickel and K. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 2001, vol. 1.
[43] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.*, vol. 26, no. 2, pp. 195–239, Apr. 1984.
[44] A. H. Stroud and D. Secrest, "Approximate integration formulas for certain spherically symmetric regions," *Math. Comput.*, vol. 17, no. 82, pp. 105–135, Apr. 1963.
[45] J. A. Wellner, "Empirical processess in action: A review," *Int. Stat. Rev.*, vol. 60, no. 3, pp. 247–269, Dec. 1992.
[46] D. Fox, "Adapting the sample size in particle filters through KLDsampling," *Int. J. Robot. Res.*, vol. 22, no. 12, pp. 985–1003, Dec. 2003.
[47] O. Lanz, "An information theoretic rule for sample size adaptation in particle filtering," in *Proc. Int. Conf. Image Anal. Process.*, Modena, Italy, Sep. 2007, pp. 317–322.
[48] D. Koks, *Explorations in Mathematical Physics*. New York, NY, USA: Springer, 2006.
[49] J. Bortz, "A new mathematical formulation for strapdown inertial navigation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 7, no. 1, pp. 61–66, Jan. 1971.
[50] X. R. Li and V. P. Jilkov, "Survey of maneuvering target tracking. Part I: Dynamic models," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1333–1364, Oct. 2003.
[51] D. Dardari, A. Conti, U. J. Ferner, A. Giorgetti, and M. Z. Win, "Ranging with ultrawide bandwidth signals in multipath environments," *Proc. IEEE*, vol. 97, no. 2, pp. 404–426, Feb. 2009.
[52] A. Conti, M. Guerra, D. Dardari, N. Decarli, and M. Z. Win, "Network experimentation for cooperative localization," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 2, pp. 467–475, Feb. 2012.
[53] D. H. Titterton and J. L. Weston, *Strapdown Inertial Navigation Technology*. Washington DC, USA: Amer. Inst. of Aeronautics and Astronautics/IEE, 2004.
[54] P. Tichavský, C. H. Muravchik, and A. Nehorai, "Posterior Cramér-Rao bounds for discrete-time nonlinear filtering," *IEEE Trans. Signal Process.*, vol. 46, no. 5, pp. 1386–1396, May 1998.
[55] D. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[56] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

Santiago Mazuelas (M'10) received the Ph.D. in Mathematics and Ph.D in Telecommunications Engineering from the University of Valladolid, Spain, in 2009 and 2011, respectively.

Since 2009, he has been a postdoctoral fellow in the Wireless Communication and Network Sciences Laboratory at MIT. He previously worked from 2006 to 2009 as a Researcher and Project Manager in a Spanish technological center, as well as a junior lecturer in the University of Valladolid. His research interests are the application of mathematical and statistical theories to signal processing and navigation networks.

Dr. Mazuelas served as member of the TPC for the IEEE Globecom in 2010–2013, ICC in 2013, VTC in 2012, and ICUWB in 2011 and 2013.He received the Best Student Prize in Telecommunications Technical Engineering from University of Valladolid in 2006, the Best Doctorate Thesis Award from University of Valladolid in 2011, and the Young Scientist Prize from the Union Radio-Scientifique Internationale (URSI) Symposium in 2007. His papers received the IEEE Communications Society Fred W. Ellersick Prize in 2012, and Best Paper Awards from the IEEE ICUWB in 2011 and the IEEE Globecom in 2011.



Yuan Shen (S'05) received the B.S. degree (with highest honor) from Tsinghua University, China, in 2005, and S.M. degree from the Massachusetts Institute of Technology (MIT), Cambridge, MA, in 2008, both in electrical engineering.

Since 2005, he has been with Wireless Communications and Network Science Laboratory at MIT, where he is now a Ph.D. candidate. He was with the Hewlett-Packard Labs in winter 2009 and the Corporate R&D of Qualcomm Inc. in summer 2008. His current research focuses on network localization and navigation, network inference techniques, wireless secrecy techniques, and localization network optimization.

Mr. Shen served as a member of the TPC for the IEEE Globecom (2010–2013), ICC (2010–2013), WCNC (2009–2013), ICUWB (2011–2013), and ICCC (2012). He is a recipient of the Marconi Society Paul Baran Young Scholar Award (2010), the MIT EECS Ernst A. Guillemin Best S.M. Thesis Award (2008), the Qualcomm Roberto Padovani Scholarship (2008), and the MIT Walter A. Rosenblith Presidential Fellowship (2005). His papers received the IEEE Communications Society Fred W. Ellersick Prize (2012) and three Best Paper Awards from the IEEE Globecom (2011), the IEEE ICUWB (2011), and the IEEE WCNC (2007).

Moe Z. Win (S'85–M'87–SM'97–F'04) received both the Ph.D. in Electrical Engineering and the M.S. in Applied Mathematics as a Presidential Fellow at the University of Southern California (USC) in 1998. He received the M.S. in Electrical Engineering from USC in 1989, and the B.S. (*magna cum laude*) in Electrical Engineering from Texas A&M University in 1987.

He is a Professor at the Massachusetts Institute of Technology (MIT). Prior to joining MIT, he was with AT&T Research Laboratories for five years and with the Jet Propulsion Laboratory for seven years. His research encompasses fundamental theories, algorithm design, and experimentation for a broad range of real-world problems. His current research topics include network localization and navigation, network interference exploitation, intrinsic wireless secrecy, adaptive diversity techniques, and ultra-wide bandwidth systems.

Professor Win is an elected Fellow of the AAAS, the IEEE, and the IET, and was an IEEE Distinguished Lecturer. He was honored with two IEEE Technical Field Awards: the IEEE Kiyo Tomiyasu Award (2011) and the IEEE Eric E. Sumner Award (2006, jointly with R. A. Scholtz). Other recognitions include the Laurea Honoris Causa from the University of Ferrara (2008), the Technical Recognition Award of the IEEE ComSoc Radio Communications Committee (2008), and the U.S. Presidential Early Career Award for Scientists and Engineers (2004).

Dr. Win is an elected Member-at-Large on the IEEE Communications Society Board of Governors (2011–2013). He was the Chair (2004–2006) and Secretary (2002–2004) for the Radio Communications Committee of the IEEE Communications Society. Over the last decade, he has organized and chaired numerous international conferences. He served as Editor (2006–2012) for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and as Area Editor (2003–2006) and Editor (1998–2006) for the IEEE TRANSACTIONS ON COMMUNICATIONS.