Abstract—This article presents an up-to-date tutorial review of nonlinear Bayesian estimation. State estimation for nonlinear systems has been a challenge encountered in a wide range of engineering fields, attracting decades of research effort. To date, one of the most promising and popular approaches is to view and address the problem from a Bayesian probabilistic perspective, which enables estimation of the unknown state variables by tracking their probabilistic distribution or statistics (e.g., mean and covariance) conditioned on the system’s measurement data. This article offers a systematic introduction of the Bayesian state estimation framework and reviews various Kalman filtering (KF) techniques, progressively from the standard KF for linear systems to extended KF, unscented KF and ensemble KF for nonlinear systems. It also overviews other prominent or emerging Bayesian estimation methods including the Gaussian filtering, Gaussian-sum filtering, particle filtering and moving horizon estimation and extends the discussion of state estimation forward to more complicated problems such as simultaneous state and parameter/input estimation.

Index Terms—State estimation, nonlinear Bayesian estimation, Kalman filtering, stochastic estimation.

I. INTRODUCTION

As a core subject of control systems theory, state estimation for nonlinear dynamic systems has been undergoing active research and development for a few decades. Considerable attention is gained from a wide community of researchers, thanks to its significant applications in signal processing, navigation and guidance, and econometrics, just to name a few. When stochastic systems, i.e., systems subjected to the effects of noise, are considered, the Bayesian estimation approaches have evolved as a leading estimation tool enjoying wide popularity. Bayesian analysis traces back to the 1763 essay [1], published two years after the death of its author, Rev. Thomas Bayes. This seminal work was meant to tackle the following question: “Given the number of times in which an unknown event has happened and failed: Required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named”. Rev. Bayes developed a solution to examine the case of only continuous probability, single parameter and a uniform prior, which is an early form of the Bayes’ rule known to us nowadays. Despite its preciousness, this work remained obscure for many scientists and even mathematicians of that era. The change came when the French mathematician Pierre-Simon de Laplace rediscovered the result and presented the theorem in the complete and modern form. A historical account and comparison of Bayes’ and Laplace’s work can be found in [2]. From today’s perspective, the Bayes’ theorem is a probability-based answer to a philosophical question: How should one update an existing belief when given new evidence [3]? Quantifying the degree of belief by probability, the theorem modifies the original belief by producing the probability conditioned on new evidence from the initial probability. This idea was applied in the past century from one field to another whenever the belief update question arose, driving numerous intriguing explorations. Among them, a topic of relentless interest is Bayesian state estimation, which is concerned with determining the unknown state variables of a dynamic system using the Bayesian theory.

The capacity of the Bayesian analysis to provide a powerful framework for state estimation has been well recognized now. A representative method within the framework is the well-known Kalman filter (KF), which “revolutionized the field of estimation ... (and) opened up many new theoretical and practical possibilities” [4]. KF was initially developed by using the least squares in the 1960 paper [5] but reinterpreted from a Bayesian perspective in [6], only four years after its invention. Further envisioned in [6] was that “the Bayesian approach offers a unified and intuitive viewpoint particularly adaptable to handling modern-day control problems”. This investigation and vision ushered a new statistical treatment of nonlinear estimation problems, laying a foundation for prosperity of research on this subject.

In this article, we offer a systematic and bottom-to-up introduction to major Bayesian state estimators, with a particular emphasis on the KF family. We begin with outlining the essence of Bayesian thinking for state estimation problems, showing that its core is the model-based prediction and measurement-based update of the probabilistic belief of unknown state variables. A conceptual KF formulation can be made readily in the Bayesian setting, which tracks the mean and covariance of the states modeled as random vectors throughout the evolution of the system. Turning a conceptual KF into executable algorithms requires certain approximations to nonlinear systems; and depending on the approximation adopted, different KF methods are derived. We demonstrate three primary members of the KF family in this context: extended KF (EKF), unscented KF (UKF), and ensemble
KF (EnKF), all of which have achieved proven success both theoretically and practically. A review of other important Bayesian estimators and estimation problems is also presented briefly in order to introduce the reader to the state of the art of this vibrant research area.

II. A BAYESIAN VIEW OF STATE ESTIMATION

We consider the following nonlinear discrete-time system:

\[
\begin{align*}
x_{k+1} &= f(x_k) + w_k, \\
y_k &= h(x_k) + v_k,
\end{align*}
\]

where \( x_k \in \mathbb{R}^{n_x} \) is the unknown system state, and \( y_k \in \mathbb{R}^{n_y} \) the output, with both \( n_x \) and \( n_y \) being positive integers. The process noise \( w_k \) and the measurement noise \( v_k \) are mutually independent, zero-mean white Gaussian sequences with covariances \( Q_k \) and \( R_k \), respectively. The nonlinear mappings \( f : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_x} \) and \( h : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_y} \) represent the process dynamics and the measurement model, respectively. The system in (1) is assumed input-free for simplicity of presentation, but the following results can be easily extended to an input-driven system.

The state vector \( x_k \) comprises a set of variables that fully describe the status or behavior of the system. It evolves through time as a result of the system dynamics. The process of states over time hence represents the system’s behavior. Because it is unrealistic to measure the complete state in most practical applications, state estimation is needed to infer \( x_k \) from the output \( y_k \). More specifically, the significance of estimation comes from the crucial role it plays in the study of dynamic systems. First, one can monitor how a system behaves with state information and take corresponding actions when any adjustment is necessary. This is particularly important to ensure the detection and handling of internal faults and anomalies at the earliest phase. Second, high-performing state estimation is the basis for the design and implementation of many control strategies. The past decades have witnessed a rapid growth of control theories, and most of them, including optimal control, model predictive control, sliding mode control and adaptive control, premise the design on the availability of state information.

While state estimation can be tackled in a variety of ways, the stochastic estimation has drawn remarkable attention and been profoundly developed in terms of both theory and applications. Today, it is still receiving continued interest and intense research effort. From a stochastic perspective, the system in (1) can be viewed as a generator of random vectors \( x_k \) and \( y_k \). The reasoning is as follows. Owing to the initial uncertainty or lack of knowledge of the initial condition, \( x_0 \) can be considered as a random vector subject to variation due to chance. Then, \( f(x_0) \) represents a nonlinear transformation of \( x_0 \), and its combination with \( w_0 \) modeled as another random vector generates a new random vector \( x_1 \). Following this line, \( x_k \) for any \( k \) is a random vector, and the same idea applies to \( y_k \). In practice, one can obtain the sensor measurement of the output at each time \( k \), which can be considered as a sample drawn from the distribution of the random vector \( y_k \). For simplicity of notation, we also denote the output measurement as \( y_k \) and the measurement set at time \( k \) as \( \mathbb{Y}_k := \{y_1, y_2, \cdots, y_k\} \). The state estimation then is to build an estimate of \( x_k \) using \( \mathbb{Y}_k \) at each time \( k \). To this end, one’s interest then lies in how to capture \( p(x_k|\mathbb{Y}_k) \), i.e., the probability density function (pdf) of \( x_k \) conditioned on \( \mathbb{Y}_k \). This is because \( p(x_k|\mathbb{Y}_k) \) captures the information of \( x_k \) conveyed in \( \mathbb{Y}_k \) and can be leveraged to estimate \( x_k \).

A “prediction-update” procedure\(^1\) can be recursively executed to obtain \( p(x_k|\mathbb{Y}_k) \). Standing at time \( k-1 \), we can predict what \( p(x_k|\mathbb{Y}_{k-1}) \) is like using \( p(x_k|\mathbb{Y}_{k-1}) \). When the new measurement \( y_k \) conveying information about \( x_k \) arrives, we can update \( p(x_k|\mathbb{Y}_{k-1}) \) to \( p(x_k|\mathbb{Y}_k) \). Characterizing a probabilistic belief about \( x_k \) before and after the arrival of \( y_k \), \( p(x_k|\mathbb{Y}_{k-1}) \) and \( p(x_k|\mathbb{Y}_k) \) are referred to as the \textit{a priori} and \textit{a posteriori} pdf’s, respectively. Specifically, the prediction at time \( k-1 \), demonstrating the pass from \( p(x_{k-1}|\mathbb{Y}_{k-1}) \) to \( p(x_k|\mathbb{Y}_{k-1}) \), is given by

\[
p(x_k|\mathbb{Y}_{k-1}) = \int p(x_k|x_{k-1})p(x_{k-1}|\mathbb{Y}_{k-1})dx_{k-1}. \tag{2}
\]

Let us explain how to achieve (2). By the Chapman-Kolmogorov equation, it can be seen that

\[
p(x_k|\mathbb{Y}_{k-1}) = \int p(x_k,x_{k-1}|\mathbb{Y}_{k-1})dx_{k-1},
\]

which, according to the Bayes’ rule, can be written as

\[
p(x_k|\mathbb{Y}_{k-1}) = \int p(x_k|x_{k-1}, \mathbb{Y}_{k-1})p(x_{k-1}|\mathbb{Y}_{k-1})dx_{k-1}.
\]

It reduces to (2), because \( p(x_k|x_{k-1}, \mathbb{Y}_{k-1}) = p(x_k|x_{k-1}) \) as a result of the Markovian propagation of the state. Then on the arrival of \( y_k \), \( p(x_k|\mathbb{Y}_{k-1}) \) can be updated to yield \( p(x_k|\mathbb{Y}_k) \), which is governed by

\[
p(x_k|\mathbb{Y}_k) = \frac{p(y_k|x_k)p(x_k|\mathbb{Y}_{k-1})}{p(y_k|\mathbb{Y}_{k-1})}. \tag{3}
\]

The above equation is also owing to the use of the Bayes’ rule:

\[
p(x_k|\mathbb{Y}_k) = \frac{p(x_k,\mathbb{Y}_k)}{p(\mathbb{Y}_k)} = \frac{p(x_k,y_k,\mathbb{Y}_{k-1})}{p(y_k,\mathbb{Y}_{k-1})} = \frac{p(y_k|x_k,\mathbb{Y}_{k-1})p(x_k,\mathbb{Y}_{k-1})}{p(y_k,\mathbb{Y}_{k-1})} = \frac{p(y_k|x_k,\mathbb{Y}_{k-1})p(x_k|\mathbb{Y}_{k-1})}{p(y_k|\mathbb{Y}_{k-1})}.
\]

Note that we have \( p(y_k|x_k, \mathbb{Y}_{k-1}) = p(y_k|x_k) \) from the fact that \( y_k \) only depends on \( x_k \). Then, (3) is obtained. Together, (2)-(3) represent the fundamental principle of Bayesian state estimation for the system in (1), describing the sequential propagation of the \textit{a priori} and \textit{a posteriori} pdf’s. The former captures our belief over the unknown quantities in the presence of only the prior evidence, and the latter updates this belief using the Bayesian theory when new evidence becomes available. The two steps, prediction and update, are executed alternately through time, as illustrated in Fig. 1.

\(^1\)The two steps are equivalently referred to as ‘time-update’ and ‘measurement-update’, or ‘forecast’ and ‘analysis’, in different literature.
the difficulty in the design of state estimators. To simplify
equivalently, uncertainty) measure. Reducing the pdf tracking
of the random state vector
x
k
, the prediction turns against the development of executable state
or impossible to derive an exact or closed-form solution. This
will only need the statistics (mean and
variances) of
x
k
, in physical sense. In addition, it can easily translate into a
philosophical abstraction that anything unknown, in one’s
mind, is subject to variations due to chance, the randomness-
function, as the uncertainty. Second, the output
measurements of the system are samples drawn from a certain
uncertainties in a system, e.g., state, are viewed from the
Bayesian estimation. First, all the unknown quantities or
measurable quantities is regarded as a random variable. Second, the output
measurements in most
applications are inferred from the
system. Finally, the system model represents transformations that the unknown and
random state variables undergo over time. Originating from the
philosophical abstraction that anything unknown, in one’s
mind, is subject to variations due to chance, the randomness-
represented by the random variables.

Looking at the above Bayesian filtering principle, we can
summarize three elements that constitute the thinking of
Bayesian estimation. First, all the unknown quantities or
uncertainties in a system, e.g., state, are viewed from the
probabilistic perspective. In other words, any unknown variable is regarded as a random variable. Second, the output
measurements of the system are samples drawn from a certain
probability distribution dependent on the concerned variables. They provide data evidence for state estimation. Finally, the
system model represents transformations that the unknown and
random state variables undergo over time. Originating from the
philosophical abstraction that anything unknown, in one’s
mind, is subject to variations due to chance, the randomness-
represented by the random variables.

Fig. 1. The Bayesian filtering principle. The running of a dynamic system
propagates the state
x
k
through time and produces output measurement
y
k
at each time
k
. For the purpose of estimation, the Bayesian filtering principle
tracks the pdf of
x
k
given the measurement set
Y
k
= \{y_1, y_2, \cdots, y_k\}. It consists of two steps sequentially implemented: the prediction step from
p(x_{k-1}|Y_{k-1})
to
p(x_k|Y_k)
, and the update step from
p(x_k|Y_{k-1})
to
p(x_k|y_k)
upon the arrival of
y_k
.

III. FROM BAYESIAN FILTERING TO KALMAN FILTERING

In the above discussion, we have shown the probabilistic
nature of state estimation and presented the Bayesian filtering principle (2)-(3) as a solution framework. However, this does
not mean that one can simply use (2)-(3) to track the conditional pdf of a random vector passing through nonlinear
transformations, because the nonlinearity often makes it difficult
or impossible to derive an exact or closed-form solution. This
challenge turns against the development of executable state
estimation algorithms, since a dynamic system’s state propagation
and observation are based on the nonlinear functions of the random state vector
x
k
, i.e.,
f(x_k)
and
h(x_k)
. Yet for the sake of estimation, one only needs the statistics (mean and
variances) of
x
k
conditioned on the measurements in most
circumstances, rather than a full grasp of its conditional pdf.
A straightforward and justifiable way is to use the mean as the
estimate of
x
k
and the covariance as the confidence (or equivalently, uncertainty) measure. Reducing the pdf tracking to the mean and covariance tracking can significantly mitigate the
difficulty in the design of state estimators. To simplify
the problem further, certain Gaussianity approximations can be made because of the mathematical tractability and
statistical soundness of Gaussian distributions (for the reader’s
cvenience, several properties of the Gaussian distribution to
be used next are summarized in the Appendix.). Proceeding
in this direction, we can reach a formulation of the Kalman
filtering (KF) methodology, as shown below.

In order to predict
x
k
at time
k−1
, we consider the
minimum-variance unbiased estimation, which gives that the
best estimate of
x
k
given
Y_{k−1}
, denoted as
\hat{x}_{k|k−1}
, is
E(x_k|Y_{k−1})
[7, Theorem 3.1]. That is,
\hat{x}_{k|k−1} = E(x_k|Y_{k−1}) = \int x_k p(x_k|Y_{k−1}) dx_k. \hspace{1cm} (4)

Inserting (2) into the above equation, we have
\hat{x}_{k|k−1} = \int \left[\int x_k p(x_k|x_{k−1})dx_k\right] \cdot p(x_{k−1}|Y_{k−1})dx_{k−1}.
\hspace{1.5cm} (5)

By assuming that
w_k
is a white Gaussian noise independent of
x_k
, we have
x_k|x_{k−1} \sim \mathcal{N}(f(x_{k−1}), Q)
and then \int x_k p(x_k|x_{k−1}) dx_k = f(x_{k−1}) according to (A.1).
Hence, (5) becomes
\hat{x}_{k|k−1} = \int f(x_{k−1}) p(x_{k−1}|Y_{k−1}) dx_{k−1}
\hspace{1cm} = E[f(x_{k−1}|Y_{k−1})]. \hspace{1cm} (6)

For \hat{x}_{k|k−1} in (6), the associated prediction error covariance
is
P_{k|k−1} = E[(x_k - \hat{x}_{k|k−1})(x_k - \hat{x}_{k|k−1})^T] = \int (x_k - \hat{x}_{k|k−1})(x_k - \hat{x}_{k|k−1})^T p(x_k|Y_{k−1}) dx_k. \hspace{1cm} (7)

With the use of (2) and (A.1), we can obtain
P_{k|k−1} = \int x_k x_k^T p(x_k|Y_{k−1}) dx_k - \hat{x}_{k|k−1} \hat{x}_{k|k−1}^T
\hspace{1cm} = \int \left[\int x_k x_k^T p(x_k|x_{k−1})dx_k\right] p(x_{k−1}|Y_{k−1}) dx_{k−1}
\hspace{1cm} - \hat{x}_{k|k−1} \hat{x}_{k|k−1}^T
\hspace{1cm} = \int \left[f(x_{k−1}) + Q\right] f(x_{k−1}) p(x_{k−1}|Y_{k−1}) dx_{k−1}
\hspace{1cm} - \hat{x}_{k|k−1} \hat{x}_{k|k−1}^T
\hspace{1cm} = \int [f(x_{k−1}) - \hat{x}_{k|k−1}] [f(x_{k−1}) - \hat{x}_{k|k−1}]^T
\hspace{1cm} \cdot p(x_{k−1}|Y_{k−1}) dx_{k−1} + Q
\hspace{1cm} = \text{Cov}[f(x_{k−1})|Y_{k−1}] + Q. \hspace{1cm} (8)

When
y_k
becomes available, we assume that
p(x_k, y_k|Y_{k−1})
can be approximated by a Gaussian distribution
\mathcal{N}\left(\begin{bmatrix} \hat{x}_{k|k−1} \\ \hat{y}_{k|k−1} \end{bmatrix}, \begin{bmatrix} P_{x|k−1} & P_{xy|k−1} \\ P_{xy|k−1}^T & P_{y|k−1} \end{bmatrix} \right), \hspace{1cm} (9)
where \hat{y}_{k|k−1} is the prediction of
y_k
given
Y_{k−1}
and expressed as
\hat{y}_{k|k−1} = E(y_k|Y_{k−1}) = \int y_k p(y_k|Y_{k−1}) dy_k. \hspace{1cm} (10)
For two jointly Gaussian random vectors, the conditional distribution of one given another is also Gaussian, which is summarized in (A.4) in Section IV. It then follows from (9) that a Gaussian approximation can be constructed for \( p(x_k | Y_k) \). Its mean and covariance can be expressed as

\[
\hat{x}_{k|k} = \hat{x}_{k|k-1} + P^{xy}_{k|k-1} (P^{y|y}_{k|k-1})^{-1} (y_k - \hat{y}_{k|k-1}),
\]

Kalman gain

\[
P^{x|y}_{k|k} = P^{x}_{k|k-1} - P^{xy}_{k|k-1} (P^{y|y}_{k|k-1})^{-1} (P^{xy}_{k|k-1})^T.
\]

Putting together (6)-(8), (13)-(14) and (15)-(17), we can establish a conceptual description of the KF technique, which is outlined in Fig. 2. Built in the Bayesian-Gaussian setting, it conducts state estimation through tracking the mean and covariance of a random state vector. It is noteworthy that one needs to develop explicit expressions to enable the use of the above KF equations. The key that bridges the gap is to find the mean and covariance of a random vector passing through nonlinear functions. For linear dynamic systems, the development is straightforward, because, in the considered context the involved pdf’s are strictly Gaussian and the linear transformation of the Gaussian state variables can be readily handled. The result is the standard KF to be shown in the next section. However, complications arise in the case of nonlinear systems. This issue has drawn significant interest from researchers. A wide range of ideas and methodologies have been developed, leading to a family of nonlinear KFs. The three most representative among them are EKF, UKF, and EnKF to be introduced following the review of the linear KF.

IV. STANDARD LINEAR KALMAN FILTER

In this section, KF for linear systems is reviewed briefly to pave the way for discussion of nonlinear KFs. Consider a linear time-invariant discrete-time system of the form

\[
\begin{align*}
\dot{x}_{k+1} &= Fx_k + w_k, \\
y_k &= Hx_k + v_k,
\end{align*}
\]

where: 1) \{\{w_k\}\} and \{\{v_k\}\} are zero-mean white Gaussian noise sequences with \( w_k \sim N(0, Q) \) and \( v_k \sim N(0, R) \). \( x_0 \) is Gaussian with \( x_0 \sim N(\tilde{x}_0, P_0) \), and 3) \( x_0 \), \{\{w_k\}\} and \{\{v_k\}\} are independent of each other. Note that, under these conditions, the Gaussian assumptions in Section III will exactly hold for the linear system in (18).

The standard KF for the linear dynamic system in (18) can be readily derived from the conceptual KF summarized in Fig. 2. Since the system is linear and under a Gaussian setting, \( p(x_{k-1} | Y_{k-1}) \) and \( p(x_k | Y_{k-1}) \) are strictly Gaussian according to the properties of Gaussian vectors. Specifically, \( x_{k-1} Y_{k-1} \sim N(\tilde{x}_{k-1|k-1}, P^{x}_{k-1|k-1}) \) and \( x_k Y_{k-1} \sim N(\tilde{x}_{k|k-1}, P^{x}_{k|k-1}) \). According to (6) and (8), the prediction is

\[
\begin{align*}
\hat{x}_{k|k-1} &= E(Fx_{k-1} Y_{k-1}) = F\hat{x}_{k-1|k-1}, \\
P^{x}_{k|k-1} &= Cov(Fx_{k-1} Y_{k-1}) + Q \\
&= FP^{x}_{k-1|k-1} F^T + Q.
\end{align*}
\]

The update can be accomplished along the similar lines. Based on (13)-(15), we have \( \hat{y}_{k|k-1} = H\hat{x}_{k|k-1}, P^{y}_{k|k-1} = P^{y}_{k|k-1} \).
A. Extended Kalman Filter

EKF is arguably the most celebrated nonlinear state estimation technique, with numerous applications across a variety of engineering areas and beyond [9]. It is based on the linearization of nonlinear functions around the most recent state estimate. When the state estimate $\hat{x}_{k-1|k-1}$ is generated, consider the first-order Taylor expansion of $f(x_{k-1})$ at this point:

$$f(x_{k-1}) \approx f(\hat{x}_{k-1|k-1}) + F_{k-1} \left( x_{k-1} - \hat{x}_{k-1|k-1} \right),$$  \hspace{1cm} (23)

$$F_{k-1} = \left. \frac{df}{dx} \right|_{\hat{x}_{k-1|k-1}}.$$  \hspace{1cm} (24)

For simplicity, let $p(x_{k-1}|\mathcal{Y}_{k-1})$ be approximated by a distribution with mean $\hat{x}_{k-1|k-1}$ and covariance $P_{x_{k-1|k-1}}$. Then inserting (23) to (6)-(8), we can readily obtain the one-step-forward prediction

$$\hat{x}_{k|k-1} = E \left[ f(x_{k-1}|\mathcal{Y}_{k-1}) \right] \approx f(\hat{x}_{k-1|k-1}),$$  \hspace{1cm} (25)

$$P_{x_{k|k-1}} = \text{Cov} \left[ x_{k} | \mathcal{Y}_{k-1} \right] + Q = F_{k-1}P_{x_{k-1|k-1}}F_{k-1}^\top + Q.$$  \hspace{1cm} (26)

Looking into (23), we find that the Taylor expansion approximates the nonlinear transformation of the random vector $x_{k}$ by an affine one. Proceeding on this approximation, we can easily estimate the mean and covariance of $f(x_{k-1})$ once provided the mean and covariance information of $x_{k-1}$ conditioned on $\mathcal{Y}_{k-1}$. This, after being integrated with the effect of the noise $w_{k}$ on the prediction error covariance, establishes a prediction of $x_{k}$, as specified in (25)-(26). After $\hat{x}_{k|k-1}$ is produced, we can investigate the linearization of $h(x_{k})$ around this new operating point in order to update the prediction. That is,

$$h(x_{k}) \approx h(\hat{x}_{k|k-1}) + H_{k} \left( x_{k} - \hat{x}_{k|k-1} \right),$$  \hspace{1cm} (27)

$$H_{k} = \left. \frac{dh}{dx} \right|_{\hat{x}_{k|k-1}}.$$  \hspace{1cm} (28)

Similarly, we assume that $p(x_{k} | \mathcal{Y}_{k-1})$ can be replaced by a distribution with mean $\hat{x}_{k|k-1}$ and covariance $P_{x_{k|k-1}}$. Using (27), the evaluation of (13)-(15) yields $\hat{y}_{k|k-1} \approx h(\hat{x}_{k|k-1})$, $P_{y_{k|k-1}} \approx H_{k}P_{x_{k|k-1}}H_{k}^\top + R$, and $P_{x_{k|k-1}} \approx P_{x_{k|k-1}}H_{k}^\top.$

Here, the approximate mean and covariance information of $h(x_{k})$ and $y_{k}$ are obtained through the linearization of $h(x_{k})$ around $\hat{x}_{k|k-1}$. With the aid of the Gaussianity assumption in (9), an updated estimate of $x_{k}$ is produced as follows:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + P_{x_{k|k-1}}H_{k}^\top \left( H_{k}P_{x_{k|k-1}}H_{k}^\top + R \right)^{-1} \cdot \left[ y_{k} - h(\hat{x}_{k|k-1}) \right],$$  \hspace{1cm} (29)

$$P_{x_{k|k}} = P_{x_{k|k-1}} - P_{x_{k|k-1}}H_{k}^\top \left( H_{k}P_{x_{k|k-1}}H_{k}^\top + R \right)^{-1} \cdot H_{k}P_{x_{k|k-1}}.$$  \hspace{1cm} (30)
Then, EKF consists of (25)-(26) for prediction and (29)-(30) for update. When comparing it with the standard KF in (19)-(22), we can find that they share significant resemblance in structure, except that EKF introduces the linearization procedure to accommodate the system nonlinearities.

Since the 1960s, EKF has gained wide use in the areas of aerospace, robotics, biomedical, mechanical, chemical, electrical and civil engineering, with great success in the real world witnessed. This is often ascribed to its relative easiness of design and execution. Another important reason is its good convergence from a theoretical viewpoint. In spite of the linearization-induced errors, EKF has provable asymptotic convergence under some conditions that can be satisfied by many practical systems [10–14]. However, it also suffers from some shortcomings. The foremost one is the inadequacy of its first-order accuracy for highly nonlinear systems. In addition, the need for explicit derivative matrices not only renders EKF futile for discontinuous or other non-differentiable systems, but also pulls it away from convenient use in view of programming and debugging, especially when nonlinear functions of a complex structure are faced. This factor, together with the computational complexity at \(O(n^3)\), limits the application of EKF to only low-dimensional systems.

Some modified EKFs have been introduced for improved accuracy or efficiency. In this regard, a natural extension is the second-order EKF with more accurate estimation [15–17]. Another important variant, named iterated EKF (IEKF), iteratively refines the state estimate around the current point at each time instant [18–20]. Though requiring an increased computational cost, it can achieve higher estimation accuracy even in the presence of severe nonlinearities in systems.

### B. Unscented Kalman Filter

As the performance of EKF degrades for systems with strong nonlinearities, researchers have been seeking better ways to conduct nonlinear state estimation. In the 1990s, UKF was invented [21, 22]. Since then, it has been gaining significant popularity among researchers and practitioners. This technique is based on the so-called “unscented transform (UT)”, which exploits the utility of deterministic sampling to track the mean and covariance information of a random variable passing through a nonlinear transformation [23–25]. The basic idea is to approximately represent a random variable by a set of sample points (sigma points) chosen deterministically to completely capture the mean and covariance. Then, projecting the sigma points through the nonlinear function concerned, one obtains a new set of sigma points and then use them to form the mean and covariance of the transformed variable for estimation.

To explain how UT tracks the statistics of a nonlinearly transformed random variable, we consider a random variable \(x \in \mathbb{R}^n\) and a nonlinear function \(z = g(x)\). It is assumed that the mean and covariance of \(x\) are \(\bar{x}\) and \(P_x\), respectively. The UT proceeds as follows [23, 24]. First, a set of sigma points \(\{x^i, i = 0, 1, \cdots, 2n\}\) are chosen deterministically:

\[
x^0 = \bar{x},
\]

(31)

\[
x^i = \bar{x} + \sqrt{n + \lambda} \left( \sqrt{P_x} \right)_i, \quad i = 1, 2, \cdots, n,
\]

(32)

\[
x^{i+n} = \bar{x} - \sqrt{n + \lambda} \left( \sqrt{P_x} \right)_i, \quad i = 1, 2, \cdots, n,
\]

(33)

where \([i]\) represents the \(i\)-th column of the matrix and the matrix square root is defined by \(\sqrt{P}\sqrt{P}^\top = P\) achievable through the Cholesky decomposition. The sigma points spread across the space around \(\bar{x}\). The width of spread is dependent on the covariance \(P\) and the scaling parameter \(\lambda\), where \(\lambda = \alpha^2(n + \kappa) - n\). Typically, \(\alpha\) is a small positive value (e.g., \(10^{-3}\)), and \(\kappa\) is usually set to 0 or 3−n [21]. Then the sigma points are propagated through the nonlinear function \(g(\cdot)\) to generate the sigma points for the transformed variable \(z\), i.e.,

\[
z^i = g(x^i), \quad i = 0, 1, \cdots, 2n.
\]

The mean and covariance of \(z\) are estimated as

\[
\bar{z} = \mathbb{E}[g(x)] \approx \sum_{i=0}^{2n} W^{(m)}_i z^i,
\]

(34)

\[
P_z = \mathbb{E} \left[ (g(x) - \bar{z}) (g(x) - \bar{z})^\top \right] 
\approx \sum_{i=0}^{2n} W^{(c)}_i \left( z^i - \bar{z} \right) (z^i - \bar{z})^\top,
\]

(35)

where the weights are

\[
W^{(m)}_0 = \frac{\lambda}{n + \lambda},
\]

(36)

\[
W^{(c)}_0 = \frac{\lambda}{n + \lambda}, \quad W^{(c)}_i = \frac{1}{2(n + \lambda)}, \quad i = 1, 2, \cdots, 2n.
\]

(37)

The parameter \(\beta\) in (37) can be used to include prior information on the distribution of \(x\). When \(x\) is Gaussian, \(\beta = 2\) is optimal. The UT procedure is schematically shown in Fig. 4.

To develop UKF, it is necessary to apply UT at both prediction and update steps, which involve nonlinear state transformations based on \(f\) and \(h\), respectively. For prediction, suppose that the mean and covariance of \(x_{k-1}, \hat{x}_{k-1|k-1}\) and \(P_{x|k-1|k-1}\), are given. To begin with, the sigma points for \(x_{k-1}\) are generated:

\[
x^{(0)}_k = \hat{x}_{k-1|k-1},
\]

(39)
Then, they are propagated forward through the nonlinear function $f(\cdot)$, that is,
\begin{equation}
\hat{x}_{k|k-1}^{(i)} = f(\hat{x}_{k-1|k-1}^{(i)}), \quad i = 0, 1, \cdots, 2n_x.
\end{equation}
These new sigma points are considered capable of capturing the mean and covariance of $f(x_{k-1})$. Using them, the prediction of $x_k$ can be achieved as follows:
\begin{equation}
\hat{x}_{k|k-1} = E[f(x_{k-1}) | Y_{k-1}] = \sum_{i=0}^{2n_x} W_i^{(m)} \hat{x}_{k|k-1}^{(i)},
\end{equation}
\begin{equation}
P_{x|k-1}^{k} = \text{Cov}[f(x_{k-1}) | Y_{k-1}] + Q
= \sum_{i=0}^{2n_x} W_i^{(c)} \left( \hat{x}_{k|k-1}^{(i)} - \hat{x}_{k|k-1} \right) \left( \hat{x}_{k|k-1}^{(i)} - \hat{x}_{k|k-1} \right) ^\top + Q.
\end{equation}
By analogy, the sigma points for $x_k$ need to be generated first in order to perform the update, which are
\begin{equation}
\hat{x}_{k|k}^{+} = \hat{x}_{k|k-1} + \sqrt{n_x + \lambda} \sqrt{P_{x|k-1}^{k}}, \quad i = 1, 2, \cdots, n_x,
\end{equation}
\begin{equation}
\hat{x}_{k|k}^{+} = \hat{x}_{k|k-1} - \sqrt{n_x + \lambda} \sqrt{P_{x|k-1}^{k}}, \quad i = 1, 2, \cdots, n_x.
\end{equation}
Propagating them through $h(\cdot)$, we can obtain the sigma points for $h(x_k)$, given by
\begin{equation}
\hat{y}_{k|k}^{(i)} = h(\hat{x}_{k|k}^{(i)}), \quad i = 0, 1, \cdots, 2n_x.
\end{equation}
The predicted mean and covariance of $y_k$ and the predicted cross-covariance between $x_k$ and $y_k$ are as follows:
\begin{equation}
\hat{y}_{k|k-1} = E[y_k | Y_{k-1}] = \sum_{i=0}^{2n_x} W_i^{(m)} \hat{y}_{k|k-1}^{(i)},
\end{equation}
\begin{equation}
P_{y|k-1}^{k} = \text{Cov}[h(x_k) | Y_{k-1}] + R
= \sum_{i=0}^{2n_x} W_i^{(c)} \left( \hat{y}_{k|k-1}^{(i)} - \hat{y}_{k|k-1} \right) \left( \hat{y}_{k|k-1}^{(i)} - \hat{y}_{k|k-1} \right) ^\top + R,
\end{equation}
\begin{equation}
P_{y|k-1}^{y} = \text{Cov}[y_k, h(x_k) | Y_{k-1}]
= \sum_{i=0}^{2n_x} W_i^{(c)} \left( \hat{x}_{k|k}^{+} - \hat{x}_{k|k} \right) \left( \hat{y}_{k|k}^{(i)} - \hat{y}_{k|k-1} \right) ^\top.
\end{equation}
With the above quantities becoming available, the Gaussian update in (16)-(17) can be leveraged to enable the projection from the predicted estimate $\hat{x}_{k|k-1}$ to the updated estimate $\hat{x}_{k|k}$.

Summarizing the above equations leads to UKF sketched in Fig. 5. Compared with EKF, UKF incurs a computational cost of the same order $O(n_x^3)$ but offers second-order accuracy [23], implying an overall smaller estimation error. In addition, its operations are derivative-free, exempt from the burdensome calculation of the Jacobian matrices in EKF. This will not only bring convenience to practical implementation but also imply its applicability to discontinuous undifferentiable nonlinear transformations. Yet, it is noteworthy that, with a complexity of $O(n_x^3)$ and operations of $2n_x + 1$ sigma points, UKF faces substantial computational expenses when the system is high-dimensional with a large $n_x$, thus unsuitable for this kind of estimation problems.

Owing to its merits, UKF has seen a growing momentum of research since its advent. A large body of work is devoted to the development of modified versions. In this respect, square-root UKF (SR-UKF) directly propagates a square root matrix, which enjoys better numerical stability than squaring the propagated covariance matrices [26]; iterative refinement of the state estimate can also be adopted to enhance UKF as in IEKF, leading to iterated UKF (IUKF). The performance of UKF can be improved by selecting the sigma points in different ways. While the standard UKF employs symmetrically distributed $2n_x + 1$ sigma points, asymmetric point sets or sets with a larger number of points may bring
better accuracy [27–30]. Another interesting question is the
determination of the optimal scaling parameter \( \kappa \), which is
investigated in [31]. UKF can be generalized to the so-called
*sigma-point Kalman filtering* (SPKF), which refers to the class
of filters that uses deterministic sampling points to determine
the mean and covariance of a random vector through nonlinear
transformation [32, 33]. Other SPKF techniques include the
central-difference Kalman filter (CDFK) and Gauss-Hermite
filter (GHKF), which perform sigma-point-based filtering and
also can be interpreted from the perspective of Gaussian
quadrature-based filtering [34] (GHKF will receive further
discussion in Section VII).

C. Ensemble Kalman Filter

Since its early development in [35–37], ensemble Kalman
filter (EnKF) has established a strong presence in the field of
state estimation for large-scale nonlinear systems. Its design
is built on an integration of KF with the Monte Carlo method,
which is a prominent statistical method concerning simulation-
based approximation of probability distributions using samples
directly drawn from certain distributions. Basically, EnKF
maintains an ensemble representing the conditional distribu-
tion of a random state vector given the measurement set.
The state estimate is generated from the sample mean and
covariance of the ensemble. In view of the sample-based
approximation of probability distributions, EnKF shares sim-
ilarity with UKF; however, the latter employs deterministic
sampling while EnKF adopts non-deterministic sampling.

Suppose that there is an ensemble of samples, \( \tilde{x}_{k|k-1} \) for
\( i = 1, 2, \ldots, N_s \), drawn from \( p(x_{k-1} | \mathbb{Y}_{k-1}) \) to approximately
represent this pdf. Next, let an ensemble of samples, \( \{w_{k|k-1}^{(i)}\} \)
for \( i = 1, 2, \ldots, N_s \), be drawn independently and identically
from the Gaussian distribution \( N(0, Q) \) in order to account
for the process noise \( w_{k-1} \) at time \( k-1 \). Then, \( \tilde{x}_{k|k-1}^{(i)} \)
can hence be projected to generate a *priori* ensemble \( \{\tilde{x}_{k|k-1}^{(i)}\} \)
that represents \( p(x_{k|k-1} | \mathbb{Y}_{k-1}) \) as follows:

\[
\tilde{x}_{k|k-1}^{(i)} = f(\tilde{x}_{k-1|k-1}^{(i)}) + w_{k|k-1}^{(i)}, \quad i = 1, 2, \ldots, N_s \tag{53}
\]

The sample mean and covariance of this ensemble can be calculated as:

\[
\tilde{x}_{k|k-1} = E(x_{k|k-1} | \mathbb{Y}_{k-1}) \approx \frac{1}{N_s} \sum_{i=1}^{N_s} \tilde{x}_{k|k-1}^{(i)}, \tag{54}
\]

\[
P_{x_{k|k-1}} = Cov(x_{k|k-1} | \mathbb{Y}_{k-1}) \approx \frac{1}{N_s-1} \sum_{i=1}^{N_s} \left( \tilde{x}_{k|k-1}^{(i)} - \tilde{x}_{k|k-1} \right) \left( \tilde{x}_{k|k-1}^{(i)} - \tilde{x}_{k|k-1} \right)^\top, \tag{55}
\]

which form the prediction formulae.

The update step begins with the construction of the ensemble
for \( p(y_k | \mathbb{Y}_{k-1}) \) by means of

\[
\hat{y}_{k|k-1}^{(i)} = h(\tilde{x}_{k|k-1}^{(i)}) + v_k^{(i)}, \quad i = 1, 2, \ldots, N_s \tag{56}
\]

where \( v_k^{(i)} \) is generated as per the Gaussian distribution \( \mathcal{N}(0, R) \)
to delineate the measurement noise \( v_k \). The sample mean of
this ensemble is

\[
\hat{y}_{k|k-1} = \frac{1}{N_s} \sum_{i=1}^{N_s} \hat{y}_{k|k-1}^{(i)}, \tag{57}
\]

with the associated sample covariance

\[
P_{y_{k|k-1}} = \frac{1}{N_s-1} \sum_{i=1}^{N_s} \left( \hat{y}_{k|k-1}^{(i)} - \hat{y}_{k|k-1} \right) \left( \hat{y}_{k|k-1}^{(i)} - \hat{y}_{k|k-1} \right)^\top. \tag{58}
\]

The cross-covariance between \( x_k \) and \( y_k \) given \( \mathbb{Y}_{k-1} \) is

\[
P_{x_{k|k-1}} = \frac{1}{N_s-1} \sum_{i=1}^{N_s} \left( \hat{x}_{k|k-1}^{(i)} - \hat{x}_{k|k-1} \right) \left( \hat{y}_{k|k-1}^{(i)} - \hat{y}_{k|k-1} \right)^\top. \tag{59}
\]

Once it arrives, the latest measurement \( y_k \) can be applied to
each member of a *priori* ensemble in the way defined by (16), i.e.,

\[
\tilde{x}_{k|k}^{(i)} = \tilde{x}_{k|k-1}^{(i)} + P_{x_{k|k}} ^{y_{k|k-1}} \left( P_{y_{k|k-1}}^{-1} \right) \left( y_k - \hat{y}_{k|k-1}^{(i)} \right), \quad i = 1, 2, \ldots, N_s. \tag{60}
\]

This *a posteriori* ensemble \( \{\tilde{x}_{k|k}^{(i)}\} \) can be regarded as an
approximate representation of \( p(x_{k|k} | \mathbb{Y}_{k}) \). Then, the updated
estimation of the mean and covariance of \( x_k \) can be achieved by

\[
\tilde{x}_{k|k} = \frac{1}{N_s} \sum_{i=1}^{N_s} \tilde{x}_{k|k}^{(i)}, \tag{61}
\]

\[
P_{x_{k|k}} = \frac{1}{N_s-1} \sum_{i=1}^{N_s} \left( \tilde{x}_{k|k}^{(i)} - \tilde{x}_{k|k} \right) \left( \tilde{x}_{k|k}^{(i)} - \tilde{x}_{k|k} \right)^\top. \tag{62}
\]

The above ensemble-based prediction and update will be
repeated recursively, forming EnKF. Note that the computation
of estimation error covariance in (55) and (62) can be skipped
if a user has no interest in learning about the estimation
accuracy. This can further cut down EnKF’s storage and
computational cost.

EnKF is illustrated schematically in Fig. 6. It features
direct operation on the ensembles as a Monte Carlo-based
extension of KF. Essentially, it represents the pdf of a state
vector by using an ensemble of samples, propagates the
ensemble members and makes estimation by computing the
mean and covariance of the ensemble at each time instant. Its
complexity is \( O(n_x^3 + n_y^2 N_s + n_y n_x N_s^2 + n_x N_s^2) \) \( n_x \gg n_y \)
\( n_x \gg N_s \) for high-dimensional systems) [38], which contrasts
with \( O(n_x^2) \) of EKF and UKF. This, along with the derivative-
free computation and freedom from covariance matrix propaga-
tion, makes EnKF computationally efficient and appealing
to be the method of choice for high-dimensional nonlinear
systems. An additional contributing factor in this respect is
that the ensemble-based computational structure places it in an
advantageous position for parallel implementation [39]. It has
been reported that convergence of the EnKF can be fast even with a reasonably small ensemble size [40, 41]. In particular, its convergence to KF in the limit for large ensemble size and Gaussian state probability distributions is proven in [41].

VI. APPLICATION TO SPEED SENSORLESS INDUCTION MOTORS

This section presents a case study of applying EKF, UKF and EnKF to state estimation for speed sensorless induction motors. Induction motors are used as an enabling component for numerous industrial systems, e.g., manufacturing machines, belt conveyors, cranes, lifts, compressors, trolleys, electric vehicles, pumps, and fans [42]. In an induction motor, electromagnetic induction from the magnetic field of the stator winding is used to generate the electric current that drives the rotor to produce torque. This dynamic process must be delicately controlled to ensure accurate and responsive operations. Hence, control design for this application was researched extensively during the past decades, e.g., [42–44]. Recent years have seen a growing interest in speed sensorless induction motors, which have no sensors to measure the rotor speed to reduce costs and increase reliability. However, the absence of the rotor speed makes control design more challenging. To address this challenge, state estimation is exploited to recover the speed and other unknown variables. It is also noted that an induction motor as a multivariable and highly nonlinear system makes a valuable benchmark for evaluating different state estimation approaches [44, 45].

The induction motor model in a stationary $d$-$q$ reference frame can be written as

$$\dot{i}_{ds} = -\gamma i_{ds} + \alpha \beta \psi_{dr} + \alpha \beta \psi_{qr} - u_{ds}/\sigma,$$

$$\dot{i}_{qs} = -\gamma i_{qs} - \beta \psi_{dr} + \alpha \beta \psi_{qr} + u_{qs}/\sigma,$$

where $(\psi_{dr}, \psi_{qr})$ is the rotor fluxes, $(i_{ds}, i_{qs})$ is the stator currents, and $(u_{ds}, u_{qs})$ is the stator voltages, all defined in a stationary $d$-$q$ frame. In addition, $\omega$ is the rotor speed to be estimated, $J$ is the rotor inertia, $T_L$ is the load torque, and $y$ is the output vector composed of the stator currents. The rest symbols are parameters, where $\sigma = L_s(1 - L_m^2/L_s L_r)$, $\alpha = R_s/L_r$, $\beta = L_m/\sigma L_r$, $\gamma = R_s/\sigma + \alpha \beta L_m$, $\mu = 3L_m^2/2L_r$; $(R_s, L_s)$ and $(R_r, L_r)$ are the resistance-inductance pairs of the stator and rotor, respectively; $L_m$ is the mutual inductance.

As shown above, the state vector $x$ comprises $i_{ds}$, $i_{qs}$, $\psi_{dr}$, $\psi_{qr}$, and $\omega$. The parameter setting follows [46]. Note that, because of the focus on state estimation, an open-loop control scheme is considered with $u_{ds}(t) = 380 \sin(100\pi t)$ and $u_{qs}(t) = -380 \sin(200\pi t)$. The state estimation problem is then to estimate the entire state vector through time using the measurement data of $i_{ds}$, $i_{qs}$, $u_{ds}$ and $u_{qs}$.

In the simulation, the model is initialized with $x_0 = [0 \ 0 \ 0 \ 0 \ 0]^T$. The initial state guess for all the filters is set to be $\tilde{x}_{0|0} = [1 \ 1 \ 1 \ 1]^T$ and $P_{0|0}^x = 10^{-6}I$. For EnKF, its estimation accuracy depends on the ensemble size. Thus, different sizes are implemented to examine this effect, with $N_s = 40, 100, 200$ and $400$. To make a fair evaluation, EnKF with each $N_s$ is run for 100 times as a means to reduce the influence of randomly generated noise. The estimation error for each run is defined as $\sum_k \|x_k - \tilde{x}_{k|k}\|_2$; the errors from the 100 runs are averaged to give the final estimation error for comparison.

Fig. 7 shows the estimation errors for $\omega$ along with $\pm 3\sigma$ bounds in a simulation run of EKF, UKF and EnKF with ensemble size of 100 (here, $\sigma$ stands for the standard deviation associated with the estimate of $\omega$, and $\pm 3\sigma$ bounds correspond to the 99% confidence region). It is seen that, in all three cases, the error is large at the initial stage but gradually decreases to a much lower level, indicating that the filters successfully adapt their running according to their own equations. In addition, UKF demonstrates the best estimation of $\omega$ overall. The average estimation errors over 100 runs are summarized in Table I. It also shows that UKF offers the most accurate estimation when all state variables are considered. In addition, the estimation accuracy of EnKF improves when the ensemble size increases.

We draw the following remarks about nonlinear state estimation from our extensive simulations with this specific problem and experience with other problems in our past research.

- The initial estimate can significantly impact the estimation accuracy. For the problem considered here, it is found that EKF and EnKF are more sensitive to an initial guess. It is noteworthy that an initial guess, if differing much from the truth, can lead to divergence of filters. Hence, one is encouraged to obtain a guess as close as possible to...
the truth through using prior knowledge or trying different guesses.

- A filter’s performance can be problem-dependent. A filter can provide estimation at a decent accuracy when applied to a problem but may fail when used to handle another. Thus, the practitioners are suggested to try different filters whenever allowed to find out the one that performs the best for his/her specific problem.

- Successful application of a filter usually requires to tune the covariance matrices and in some cases, parameters involved in a filter (e.g., $\lambda$, $\alpha$ and $\beta$ in UKF), because of their important influence on estimation [47]. The trial-and-error method is common in practice. Meanwhile, there also exist some studies of systematic tuning methods, e.g., [48, 49]. Readers may refer to them for further information.

- In choosing the best filter, engineers need to take into account all the factors relevant to the problem they are addressing, including but not limited to estimation accuracy, computational efficiency, system’s structural complexity, and problem size. To facilitate such a search, Table II summarizes the main differences and application areas of EKF, UKF and EnKF.

VII. OTHER FILTERING APPROACHES AND ESTIMATION PROBLEMS

Nonlinear stochastic estimation remains a major research challenge for the control research community. Continual research effort has been in progress toward the development of advanced methods and theories in addition to the KFs reviewed above. This section gives an overview of other major filtering approaches.

**Gaussian filters (GFs).** GFs are a class of Bayesian filters enabled by a series of Gaussian distribution approximations. They bear much resemblance with KFs in view of their prediction-update structure and thus, in a broad sense, belong to the KF family. As seen earlier, the KF-based estimation relies on the evaluation of a set of integrals indeed—for example, the prediction of $x_k$ is attained in (6) by computing the conditional mean of $f(x_{k-1})$ on $Y_{k-1}$. The equation is repeated here for convenience of reading:

$$\hat{x}_{k|k-1} = E[f(x_{k-1})|Y_{k-1}] = \int f(x_{k-1})p(x_{k-1}|Y_{k-1})dx_{k-1}.$$ 

GFs approximate $p(x_{k-1}|Y_{k-1})$ with a Gaussian distribution having mean $\hat{x}_{k-1|k-1}$ and covariance $P_{k-1|k-1}^{x}$. Namely,
Continuing with this assumption, one can use the Gaussian quadrature integration rules to evaluate the integral. A quadrature is a means of approximating a definite integral of a function by a weighted sum of values obtained by evaluating the function at a set of deterministic points in the domain of integration. An example of a one-dimensional Gaussian quadrature is the Gauss-Hermite quadrature, which plainly states that, for a given function \( g(x) \),

\[
\int_{-\infty}^{\infty} g(x) \cdot N(x; 0, 1) \, dx \approx \sum_{i=1}^{m} w_i g(x_i),
\]

where \( m \) is the number of points used, \( x_i \) for \( i = 1, 2, \ldots, m \) the roots of the Hermite polynomial \( H_m(x) \), and \( w_i \) the associated weights

\[
w_i = \frac{2^{m-1} m! \sqrt{\pi}}{m^2 \left[H_m(-1(x_i))\right]^2}.
\]

Exact equality holds for polynomials of order up to \( 2m - 1 \). Applying the multivariate version of this quadrature, one can obtain a filter in a KF form, which is named Gauss-Hermite KF (GHKF) [34, 50]. GHKF reduces to UKF in certain cases [34]. Besides, the cubature rules for numerical integration can also be used in favor of a KF realization, which yields a cubature Kalman filter (CKF) [51, 52]. It is noteworthy that CKF is a special case of UKF given \( \alpha = 1, \beta = 0 \) and \( \kappa = 0 \) [53].

**Gaussian-sum filters (GSFs).** Though used widely in the development of GFs and KFs, Gaussian approximations are often inadequate and performance-limiting for many systems. To deal with a non-Gaussian pdf, GSFs represent it by a weighted sum of Gaussian basis functions [7]. For instance, the a posteriori pdf of \( x_k \) is approximated by

\[
p(x_k|\mathbb{Y}_K) = \sum_{i=1}^{m} W_{k|i}^i N\left(x_k; \hat{x}_{k|k}^i, P_{k|k}^i\right),
\]

where \( W_{k|i}^i, \hat{x}_{k|k}^i, \) and \( P_{k|k}^i \) are the weight, mean and covariance of the \( i \)-th Gaussian basis function (kernel), respectively. This can be justified by the Universal Approximation Theorem, which states that a continuous function can be approximated by a group of Gaussian functions with arbitrary accuracy under some conditions [54]. A GSF then recursively updates \( \hat{x}_{k|k}^i, P_{k|k}^i, \) and \( W_{k|i}^i \). In the basic form, \( \hat{x}_{k|k}^i, P_{k|k}^i \) for \( i = 1, 2, \ldots, m \) are updated individually through EKF, and \( W_{k|i}^i \) updated according to the output-prediction accuracy of \( \hat{x}_{k|k}^i \). The assumption for the EKF-based update is that the system’s nonlinear dynamics can be well represented by aggregating linearizations around a sufficient number of different points (means). In recent years, more sophisticated GSFs have been developed by merging the Gaussian-sum pdf approximation with other filtering techniques such as UKF, EnKF, GFs and particle filtering [34, 55–58] or optimization techniques [59].

**Particle filters (PFs).** The PF approach was first proposed in the 1990s [60] and came to prominence soon after that owing to its capacity for high-accuracy nonlinear non-Gaussian estimation. Today they have grown into a broad class of filters. As random-sampling-enabled numerical realizations of the Bayesian filtering principle, they are also known as the sequential Monte Carlo methods in the literature. Here, we introduce the essential idea with minimum statistical theory to offer the reader a flavor of this approach. Suppose that \( N_s \) samples, \( \hat{x}_k^{(i)} \) for \( i = 1, 2, \ldots, N_s \), are drawn from \( p(x_k|\mathbb{Y}_k) \) at time \( k \). The \( i \)-th sample is associated with a weight \( W_k^{(i)} \), and \( \sum_{i=1}^{N_s} W_k^{(i)} = 1 \). Then, \( p(x_k|\mathbb{Y}_k) \) can be empirically described as

\[
p(x_k|\mathbb{Y}_k) \approx \sum_{i=1}^{N_s} W_k^{(i)} \delta\left(x_k - \hat{x}_k^{(i)}\right).
\]

This indicates that the estimate of \( x_k \) is

\[
\hat{x}_k = \int x_k p(x_k|\mathbb{Y}_k) \, dx = \sum_{i=1}^{N_s} W_k^{(i)} \hat{x}_k^{(i)}.
\]

The samples can be propagated one-step forward to generate a sampling-based description of \( x_k \), i.e.,

\[
\hat{x}_k = f(\hat{x}_{k-1}^{(i)}) + w_k^{(i)}, \quad i = 1, 2, \ldots, N_s,
\]

where \( w_k^{(i)} \) for \( i = 1, 2, \ldots, N_s \) are samples drawn from the distribution of \( w_{k-1} \). After the propagation, each new sample should take a different weight in order to be commensurate with its probabilistic importance with respect to the others. To account for this, one can evaluate \( p(y_k|\hat{x}_k) \), which quantifies the likelihood of \( y_k \) given the \( i \)-th sample \( \hat{x}_k^{(i)} \). Then, the weight can be updated and normalized on \([0, 1]\) by

\[
W_k^{(i)} = \frac{W_k^{(i)} p(y_k|\hat{x}_k^{(i)})}{\sum_{i=1}^{N_s} W_k^{(i)}}.
\]

Then, an empirical sample-based distribution is built for \( p(x_k|\mathbb{Y}_k) \) as in (63), and the estimate of \( x_k \) can be computed as

\[
\hat{x}_k = \sum_{i=1}^{N_s} W_k^{(i)} \hat{x}_k^{(i)}.
\]
where \( \Phi(\cdot) \) should be incorporated. This produces the formulation of MHE here should be subject to the system model in (1). Meanwhile, the information up to the beginning of the horizon. The minimization of the above cost function becomes

\[
\min_{x_{k-1}, \ldots, x_{k}} \Phi(x_{k-1}) + \sum_{l=k-N}^{k-1} w_l^	op Q^{-1} w_l + \sum_{l=k-N}^{k} v_l^	op R^{-1} v_l,
\]

subject to:
\[
x_{l+1} = f(x_l) + w_l, \\
y_l = h(x_l) + v_l, \\
x \in \mathcal{X}, w \in \mathcal{W}, v \in \mathcal{V},
\]

where \( \mathcal{X}, \mathcal{W} \) and \( \mathcal{V} \) are, respectively, the sets of all feasible values for \( x, w \) and \( v \) and imposed as the constraints. It is seen that MHE tackles the state estimation through constrained optimization executed over time in a receding-horizon manner, as shown in Fig. 9. For an unconstrained linear system, MHE reduces to the standard KF. It is worth noting that the arrival cost \( \Phi(x_{k-1}) \) is crucial for the performance or even success of MHE approach. In practice, an exact expression is often unavailable, thus requiring an approximation [71, 72].

With the deployment of constrained optimization, MHE is computationally expensive and usually more suited for slow dynamic processes; however, the advancement of real-time optimization has brought some promises to its faster implementation [73, 74].

**Simultaneous state and parameter estimation (SSPE).** In state estimation problems, a system model is considered fully known a priori. This may not be true in various real-world situations, where part or even all of the model parameters are unknown or subject to time-varying changes. Lack of knowledge of the parameters can disable an effort for state estimation in such a scenario. Hence, SSPE is motivated to enable state estimation self-adapting to the unknown parameters. Despite the complications, a straightforward and popular way for SSPE is through state augmentation. To deal with the parameters, the state vector is augmented to include them, and on account of this, the state-space model will be transformed accordingly to one with increased dimensions. Then, a state estimation technique can be applied directly to the new model to estimate the augmented state vector, which is a joint estimation of the state variables and parameters. In the literature, EKF, UKF, EnKF and PFs have been modified using this idea for a broad range of applications [20, 75–78]. Another primary solution is the so-called dual Kalman filtering. By “dual”, it means that the state estimation and parameter estimation are performed in parallel and alternately. As such, the state estimate is used to estimate the parameters, and the parameter estimate is used to update the state estimation. Proceeding with this idea, EKF, UKF and EnKF can be dualized [79–82]. It should be pointed out that caution should be taken when an SSPE approach is developed. Almost any SSPE problem is nonlinear by nature with coupling between state variables and parameters. The joint state observability and parameter identifiability may be unavailable, or the estimation may get stuck in local minima. Consequently, the convergence can be vulnerable or unguaranteed, diminishing the chance of successful estimation. Thus application-specific SSPE analysis and development are recommended.

**Simultaneous state and input estimation (SSIE).** Some practical applications entail not only unknown states but also unknown inputs. An example is the operation monitoring for an industrial system subject to unknown disturbance, where the

A possible limitation is their computational complexity, which comes at \( O(N_s n_x^2) \) with \( N_s \gg n_x \). Yet, a strong anticipation is that the rapid growth of computing power tends to overcome this limitation, enabling wider application of PFs. A plethora of research has also been undertaken toward computationally efficient PFs [65]. A representative means is the Rao-Blackwellization that applies the standard KF to the linear part of a system and a PF to the nonlinear part and reduces the number of samples to operate on [16]. The performance of PFs often relies on the quality of samples used. To this end, KFs can be used in combination to provide high-probability particles for PFs, leading to a series of combined KF-PF techniques [66–68]. A recent advance is the implicit sampling method to generate samples capable of an improved approximation of the pdf [69, 70].

**Moving-horizon estimators (MHEs).** MHEs are an emerging estimation tool based on constrained optimization. In general, they aim to find the state estimate through minimizing a cost function subject to certain constraints. The cost function is formulated on the basis of the system’s behavior in a moving horizon. To demonstrate the idea, we consider the Maximum a Posteriori estimation (MAP) for the system in (1) during the horizon \([k-N, k]\) as shown in (64). Assuming \( w_k \sim \mathcal{N}(0, Q) \) and \( v_k \sim \mathcal{N}(0, R) \) and using the logarithmic transformation, the above cost function becomes

\[
\min_{x_{k-N}, \ldots, x_{k}} \Phi(x_{k-N}) + \sum_{l=k-N}^{k-1} w_l^	op Q^{-1} w_l + \sum_{l=k-N}^{k} v_l^	op R^{-1} v_l,
\]

where \( \Phi(x_{k-N}) \) is the arrival cost summarizing the past information up to the beginning of the horizon. The minimization here should be subject to the system model in (1). Meanwhile, some physically motivated constraints for the system behavior should be incorporated. This produces the formulation of MHE given as

\[
\{\hat{x}_{k-N}, \ldots, \hat{x}_k\} = \arg\min_{x_{k-N}, \ldots, x_{k}} \Phi(x_{k-N})
\]
\[ \{\hat{x}_{k-N}, \ldots, \hat{x}_k\} = \arg \max_{x_{k-N}, \ldots, x_k} p(x_{k-N}, \ldots, x_k | y_{k-N}, \ldots, y_k) \]

\[ = \arg \max_{x_{k-N}, \ldots, x_k} p(x_{k-N}) \prod_{l=k-N}^k p(y_l | x_l) \prod_{l=k-N}^{k-1} p(x_{l+1} | x_l) \]

\[ = \arg \max_{x_{k-N}, \ldots, x_k} p(x_{k-N}) \prod_{l=k-N}^k p(v_l) \prod_{l=k-N}^{k-1} p(w_l). \] (64)

\[ \text{Output measurements} \]

\[ \text{Arrival cost} \quad \Phi(x_{k-N}) \]

\[ k-1 \quad k \quad N + 1 \]

\[ \hat{x}_{k-N}, \ldots, \hat{x}_k \]

\[ y_k \quad y_{k-1} \quad \ldots \quad y_{k-N} \]

Fig. 9. MHE performed over a moving observation horizon that spans \( N + 1 \) time instants. For estimation at time \( k \), the arrival cost \( \Phi(x_{k-N}) \) is determined first, which summarizes the information of the system behavior up to the beginning of the horizon. Then, the output measurements within the horizon, \( y_{k-N}, \ldots, y_k \), are used, along with the arrival cost, to conduct estimation of \( \hat{x}_{k-N}, \ldots, \hat{x}_k \) through constrained optimization.

Operational status is the state and the disturbance the input. In maneuvering target tracking, the tracker often wants to estimate the state of the target, e.g., position and velocity, and the input, e.g., the acceleration. Another example is the wildfire data assimilation extensively investigated in the literature. The spread of wildfire is often driven by local meteorological conditions such as the wind. This gives rise to the need for a joint estimation of both the fire perimeters (state) and the wind speed (input) toward accurate monitoring of the fire growth.

The significance of SSIE has motivated a large body of work. A lead was taken in [83] with the development of aKF-based approach to estimate the state and external white process noise of a linear discrete-time system [83]. Most recent research builds on the existing state estimation techniques. Among them, we highlight KF [84, 85], MHE [86], \( H_\infty \)-filtering [87], sliding mode observers [88, 89], and minimum-variance unbiased estimation [90–94]. SSIE for nonlinear systems involves more complexity, with fewer results reported. In [95, 96], SSIE is investigated for a special class of nonlinear systems that consist of a nominally linear part and a nonlinear part. However, the Bayesian statistical thinking has been generalized to address this topic, exemplifying its power in the development of nonlinear SSIE approaches. In [97, 98], a Bayesian approach along with numerical optimization is taken to achieve SSIE for nonlinear systems of a general form. This Bayesian approach is further extended in [99, 100] to build an ensemble-based SSIE method, as a counterpart of EnKF, for high-dimensional nonlinear systems. It is noteworthy that SSIE and SSPE would overlap if we consider the parameters as a special kind of inputs to the system. In this case, the SSIE approaches may find their use in solving SSPE problems.

VIII. Conclusion

This article offered a state-of-the-art review of nonlinear state estimation approaches. As a fundamental problem encountered across a few research areas, nonlinear stochastic estimation has stimulated a sustaining interest during the past decades. In the pursuit of solutions, the Bayesian analysis has proven to be a time-tested and powerful methodology for addressing various types of problems. In this article, we first introduced the Bayesian thinking for nonlinear state estimation, showing the nature of state estimation from the perceptive of Bayesian update. Based on the notion of Bayesian state estimation, a general form of the celebrated KF is derived. Then, we illustrated the development of the standard KF for linear systems and EKF, UKF and EnKF for nonlinear systems. A case study of state estimation for speed sensorless induction motors was provided to present a comparison of the EKF, UKF and EnKF approaches. We further extended our view to a broader horizon including GF, GSF, PF and MHE approaches, which are also deeply rooted in the Bayesian state estimation and thus can be studied from a unified Bayesian perspective to a large extent.

Despite remarkable progress made thus far, it is anticipated that nonlinear Bayesian estimation continues to see intensive research in the coming decades. This trend will be partially driven by the need to use state estimation as a mathematical tool to enable various emerging systems in contemporary industry and society, stretching from autonomous transportation to sustainable energy and smart X (grid, city, planet, geosciences, etc.). Here, we envision several directions that may shape the future research in this area. The first one lies in accurately characterizing the result of a nonlinear transformation applied to a probability distribution. Many of the present methods such as EKF, UKF and EnKF were more or less motivated to address this fundamental challenge. However, there still exists no solution generally acknowledged as being satisfactory, leaving room for further exploration. Second, much research is needed to deal with uncertainty. Uncertainty is intrinsic to many practical systems because of unmodeled dynamics, external disturbances, inherent variability of a dynamic process, and sensor noise, representing a major threat to successful estimation. Although the literature contains many results on state estimation with robustness to uncertainty, the research has not reached a level of maturity because of the difficulty involved. A third research direction is optimal sensing structure design. Sensing structure
or sensor deployment is critical for data informativeness and thus can significantly affect the effectiveness of estimation. An important question thus is how to achieve co-design of a sensing structure and Bayesian estimation approach to maximize estimation accuracy. Fourth, Bayesian estimation in a cyber-physical setting is an imperative. Standing at the convergence of computing, communication and control, cyber-physical systems (CPSs) are foundationally important and underpinning today’s smart X initiatives. They also present new challenges for estimation, which include communication constraints or failures, computing limitations, and cyber data attacks. The current research is particularly rare on nonlinear Bayesian estimation for CPSs. Finally, many emerging industrial and social applications are data-intensive, thus asking Bayesian estimation for CPSs. Finally, many emerging industrial and social applications are data-intensive, thus asking Bayesian estimation for CPSs. Finally, many emerging industrial and social applications are data-intensive, thus asking Bayesian estimation for CPSs.

APPENDIX

This appendix offers a summary of the properties of the Gaussian distribution. Suppose $z \in \mathbb{R}^n$ is a Gaussian random vector with $z \sim \mathcal{N}(\bar{z}, P_z)$. The pdf of $z$ is expressed as

$$p(z) = \frac{1}{(\sqrt{2\pi})^n |P_z|} \exp\left(-\frac{(z - \bar{z})^T P_z^{-1} (z - \bar{z})}{2}\right).$$

Some useful properties of the Gaussian vectors are as follows [101].

1) $\int z p(z) dz = \bar{z}$, $\int (z - \bar{z})(z - \bar{z})^T p(z) dz = P_z$;

$$\int z z^T p(z) dz = P_z + \bar{z} \bar{z}^T. \quad (A.1)$$

2) The affine transformation of $z$, $Az + b$, is Gaussian, i.e.,

$$A z + b \sim \mathcal{N} \left( A \bar{z} + b, A P_z A^T \right). \quad (A.2)$$

3) The sum of two independent Gaussian random vectors is Gaussian; i.e., if $z_1 \sim \mathcal{N}(\bar{z}_1, P_{z_1})$ and $z_2 \sim \mathcal{N}(\bar{z}_2, P_{z_2})$ and if $z_1$ and $z_2$ are independent, then

$$A z_1 + B z_2 \sim \mathcal{N} \left(A \bar{z}_1 + B \bar{z}_2, AP_{z_1} A^T + BP_{z_2} B^T \right). \quad (A.3)$$

4) For two random vectors jointly Gaussian, the conditional distribution of one given the other is Gaussian. Specifically, if $z_1$ and $z_2$ are jointly Gaussian with

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} \bar{z}_1 \\ \bar{z}_2 \end{bmatrix}, \begin{bmatrix} P_{z_1} & P_{z_1, z_2} \\ P_{z_1, z_2}^T & P_{z_2} \end{bmatrix} \right),$$

then

$$z_1 | z_2 \sim \mathcal{N} \left( \bar{z}_1 + P_{z_1, z_2} P_{z_2}^{-1} (z_2 - \bar{z}_2), P_{z_1} - P_{z_1, z_2} P_{z_2}^{-1} P_{z_1, z_2}^T \right). \quad (A.4)$$
