Distributionally Robust State Estimation for Nonlinear Systems

Shixiong Wang, *Member, IEEE*

***Abstract*—Uncertainties unavoidably exist in modeling for nonlinear systems: state equation, measurement equation, and/or noises statistics might be uncertain. Such model mismatches render the performance of nominally optimal state estimators being deteriorated or even unsatisfactory. Therefore, robust filters that are insensitive to modeling uncertainties have to be designed. The challenge is to quantitatively describe the uncertainties and then design accordingly efficient robust filters. Since uncertainties in nominal models make prior state distributions and likelihood distributions uncertain as well, this article proposes a distributionally robust particle filtering framework for nonlinear systems subject to modeling uncertainties. Specifically, we use worst-case prior state distributions (near the nominal prior state distributions) to generate prior state particles and/or determine their weights. Likewise, worst-case likelihood distributions (near the nominal likelihood distributions) are used to evaluate the worst-case likelihoods of prior state particles at given measurements. The "worst-case" scenario is quantified by entropy of distributions, and maximum entropy distributions are found in balls centered at nominal distributions with radii defined by statistical similarity measures such as moments-based similarity, Wasserstein distance, and Kullback-Leibler divergence. We prove that Gaussian approximation filters (e.g., unscented/cubature/ensemble Kalman filter) are distributionally robust in the sense that they use maximum entropy prior state distributions and maximum entropy likelihood distributions. Moreover, we show that the distributionally robust particle filtering framework provides a likelihood evaluation method for general nonlinear measurement equation with non-additive and non-multiplicative measurement noises. At last, we discuss measurement outlier treatment strategies in the distributionally robust particle filtering framework.**

***Index Terms*—Approximated Bayesian inference, Kullback-Leibler divergence, maximum entropy, particle filter, sequential Monte Carlo, Wasserstein distance.**

I. Introduction

RESEARCH on state estimation for nonlinear systems is lastingly active in several academic communities such as target tracking [1], power systems [2], reliability engineering [3], geodesy [4], sensor network [5], control and automation (e.g., robotics [6]), and astronautics [7]. Typical treatment frameworks include: 1) linearization methods, e.g., extended Kalman filters [8] and Takagi–Sugeno fuzzy approximation [9], 2) Gaussian approximation methods [10] including unscented Kalman filters [11], cubature Kalman filters [12], ensemble Kalman filters [13], etc., and 3) approximated Bayesian inference methods such as variational Bayesian inference [14], [15], [16], [17] and sequential Monte Carlo methods (a.k.a. particle filters) [18], [19], [20], [21]. All these methods try to find the posterior state distribution, at every time step, conditional on the past measurements. Some key points can be summarized as follows: 1) Particle filters are optimal methods in the sense that they can compute the exact posterior state distributions if given sufficiently many particles [22]; 2) Ensemble Kalman filters are approximation methods of particle filters, which adopt the closed-form Kalman iterations to reduce the computational complexity at the cost of scarifying the approximation accuracy for the posterior state distributions [13]; 3) Cubature Kalman filters and unscented Kalman filters are special cases of ensemble Kalman filters and they use only a very limited number of particles (which are called sigma points) to further reduce the computational complexity.

Linearization methods and Gaussian approximation methods are doubted for their incapability of capturing severe nonlinearities, while approximated Bayesian inference methods are criticized for their high computational burdens. However, continuous improvement in computation powers of modern microprocessors/computers is reducing such criticisms on approximated Bayesian inference methods and encouraging signal processing practitioners to implement these methods for higher estimation accuracy. On this basis, sequential Monte Carlo methods (i.e., particle filters) are of more interest because solving functional optimization problems in variational Bayesian inference is theoretically challenging and therefore additional assumptions, e.g. parameterized function representation and mean field approximation [23], are required.

Over the years, tremendous efforts have been made to perfect particle filters, especially in designing efficient sampling and resampling techniques [18], [24], [25], [26]. However, virtually all of the past literature assumes that the state equation and measurement equation are accurate. This assumption is suspect because uncertainties are unavoidable in modeling; i.e., nominal models designed by scientists/engineers are not guaranteed to be exactly the same as the true governing models. Such uncertainties may be incurred by oscillating but unknown values of elements in circuits (e.g., resistors/inductors influenced by thermal/electromagnetic noises), by uncontrollable factors

Manuscript received 15 November 2021; revised 9 July 2022; accepted 26 August 2022. Date of publication 31 August 2022; date of current version 15 September 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Monica F. Bugallo.

The author is with the Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore 119077 (e-mail: s.wang@u.nus.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TSP.2022.3203225, provided by the author.

Digital Object Identifier 10.1109/TSP.2022.3203225

in model identification (e.g., numerical errors in parameter estimation; mismatched model assumptions), etc. Therefore, uncertainty-aware particle-based state estimation solutions have to be studied.[1] There are two philosophies in statistics, optimization, and also engineering to cope with uncertainties. The one is to reduce such uncertainties by, e.g., jointly estimating the true values of the uncertain factors whenever it is possible [32], [33], [34], [35], [36], whereas the other is to tolerate the uncertainties by, e.g., designing robust solutions that are insensitive to them [37], [38], [39], [40], [41]. The former is referred to as **adaptive methods**, and the latter is termed **robust methods**. Specifically, in the state estimation literature (for linear systems), adaptive methods include unknown-input Kalman filters [34], adaptive Kalman filters [35], etc., while robust methods contain, e.g., robust Kalman filters [42], [43] and distributionally robust state estimators [39], [40]. Since not all uncertain factors can be correctly characterized, quantitatively modeled, and exactly estimated, in general, robust solutions are attractive. Distributionally robust optimization theory, an offspring of robust statistics and optimization theories, is a mainstream framework dealing with modeling uncertainties. It is currently popular in operations research [44], machine learning [38], [45], systems control [46], to name a few. When some statistical information of uncertain factors are known in prior, distributionally robust optimization methods are preferable over classical robust optimization methods which only take into account possible values that the uncertain factors can take. This is because distributional information can be utilized to counteract conservativeness, to some extent [47].

In this article, distributionally robust optimization theory is leveraged to robustify particle filters. This is because particle filters are Bayesian statistical methods, and therefore, natural to be discussed in "distributional" contexts.[2] When a nominal state equation is not guaranteed to be exactly the same as the true one, we argue that the associated nominal prior state distributions, which are represented by weighted particles that are generated from this nominal state equation, are different from the true prior state distributions as well. Therefore, we propose to find the worst-case prior state distributions near the nominal prior state distributions, and use these worst-case distributions as surrogates to generate new prior state particles and/or update their weights. On the other hand, when the measurement equation is inexact, the likelihoods of the prior state particles cannot be exactly evaluated either. Likewise, we suggest finding worst-case likelihood distributions for prior state particles to evaluate their likelihoods. Intuitively, we recall the Bayesian posterior estimation principle: $p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x}) \cdot p(\boldsymbol{x})$. Hence, if we trust more the prior distribution $p(\boldsymbol{x})$ and doubt the exactness of the likelihood distribution $p(\boldsymbol{y}|\boldsymbol{x})$, we should let $p(\boldsymbol{y}|\boldsymbol{x})$ be noninformative/uncertain and make main use of $p(\boldsymbol{x})$. Conversely, we should let $p(\boldsymbol{x})$ be noninformative/uncertain and mainly utilize $p(\boldsymbol{y}|\boldsymbol{x})$.

The "uncertainty" of distributions can be quantified by their entropy [48], [49]. A large entropy value implies that the distribution is not concentrated/informative,[3] and instead scattered/noninformative; i.e., we are less certain about happenings. Therefore, an eligible method to quantify the "worst-case" scenario is to use entropy; it is due to the principle of maximum entropy: "the probability distribution which best represents the current state of knowledge about a system is the one with largest entropy [53]". The principle of maximum entropy is also popular in robust Bayesian methods [50], [51], [54], especially in choosing robust prior distributions. To be specific, according to [51, p.229], flat-tailed priors and noninformative priors can robustify a Bayesian statistical method. In fact, maximum entropy distributions tend to be flat-tailed because maximizing the entropy of a variable (n.b., not fixed) distribution admits minimizing the Kullback–Leibler divergence of this distribution from a uniform distribution [cf. (39)], and uniform distributions are most flat-tailed. In this article, we utilize classical statistical metrics and divergences, such as the Wasserstein metric and the Kullback-Leibler divergence, to construct balls containing a family of distributions centered at a nominal prior state distribution or at a nominal likelihood distribution. Then, we find maximum entropy distributions in the balls to generate new prior state particles and/or update their weights, and to evaluate the worst-case likelihoods of these prior state particles. As a result, the worst-case posterior state particles are immediate to be obtained by particle filters.

The contributions of this article can be summarized as follows.

1) We propose a robustification scheme for particle filters. Specifically, in implementing a particle filter, we use worst-case (i.e., maximum-entropy) prior state distribution near the nominal prior state distribution to generate new prior state particles and/or update their weights, and use worst-case (i.e., maximum-entropy) likelihood distribution near the nominal likelihood distribution to evaluate the worst-case likelihoods of these prior state particles. For details, see Sections II and IV.
2) We derive maximum entropy distributions in balls centered at nominal distributions with radii defined by the Wasserstein distance and the Kullback-Leibler divergence. For details, see Section III-B and III-C, especially Theorems 3, 4, 5, and 6.
3) We show that this robustification scheme serves yet a new resampling strategy against particle degeneracy. In detail, maximum entropy distributions tend to have uniform probability for each support point, and therefore, in particle filter, particles tend to have equal weights. For details, see Section IV-A, especially (39) and (40).
4) We show that the proposed robustification scheme offers a universal likelihood evaluation method for prior state particles when measurement equation is driven by

[1] Although there exist robustified extended Kalman filters [27], [28], robustified cubature Kalman filters [29], [30], and robustified Gaussian filters [31] for uncertain nonlinear system models subject to, e.g., non-Gaussian measurement noises, they are based on sub-optimal Gaussian approximation filters which cannot sufficiently handle nonlinearities.

[2] The term "distributional" means probability-distribution-related. One should differentiate it with another term "distributed" in engineering literature.

[3] Maximum-entropy distributions tend to be noninformative [50, Section 2.3], [51], [52].

non-additive and non-multiplicative noises. For details, see Section IV-B, especially Methods 4 and 5.

5) We illustrate that Gaussian approximation state estimators are distributionally robust. For details, see Section III-A, especially Corollary 1.
6) We provide a measurement outlier identification and treatment method for particle filters. For details, see Section IV-C.

Notations: Let $\mathbb{R}^d$ denote the d-dimensional Euclidean space, L^1 the absolutely integrable function space, and l^1 the absolutely summable vector space. We use $\mathbb{P}_{\boldsymbol{x}}$ to denote the distribution of the random vector $\boldsymbol{x}$ and $\mathbb{E}\boldsymbol{x}$ its expectation. Let the probability density function of $\boldsymbol{x}$ be $p(\boldsymbol{x})$. Let $\boldsymbol{Y}_k := \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k\}$ denote the measurement set up to and including the time k. Let $\delta_{\boldsymbol{x}_0}(\boldsymbol{x})$ be the Dirac delta function: $\delta_{\boldsymbol{x}_0}(\boldsymbol{x}) = \infty$ if $\boldsymbol{x} = \boldsymbol{x}_0$ and 0 otherwise; $\int \delta_{\boldsymbol{x}_0}(\boldsymbol{x}) d\boldsymbol{x} = 1$. We use $\boldsymbol{F}^T$ to denote the transpose of a matrix $\boldsymbol{F}$. For a given integer N, we let $[N] := \{1, 2, 3, \ldots, N\}$ denote a running index set.

II. Problem Formulation

Consider a nonlinear system model

$$\begin{cases} \boldsymbol{x}_k = \boldsymbol{f}_k(\boldsymbol{x}_{k-1}, \boldsymbol{w}_{k-1}) \\ \boldsymbol{y}_k = \boldsymbol{h}_k(\boldsymbol{x}_k, \boldsymbol{v}_k) \end{cases} \tag{1}$$

in which $\boldsymbol{x}_k \in \mathbb{R}^n$ is the state vector, $\boldsymbol{y}_k \in \mathbb{R}^m$ is the measurement vector, $\boldsymbol{w}_{k-1} \in \mathbb{R}^p$ is the process noise vector, $\boldsymbol{v}_k \in \mathbb{R}^q$ is the measurement noise vector, $\boldsymbol{f}_k(\cdot, \cdot)$ is the state evolution function, and $\boldsymbol{h}_k(\cdot, \cdot)$ is the state measurement function; $k = 1, 2, 3, \ldots$ denotes the discrete time index. We assume that $\boldsymbol{x}_k$, $\boldsymbol{y}_k$, $\boldsymbol{w}_k$, and $\boldsymbol{v}_k$ have finite second moments, and $\boldsymbol{f}_k(\cdot, \cdot)$ and $\boldsymbol{h}_k(\cdot, \cdot)$ have finite operator norms (i.e., bounded inputs give bounded outputs). The task is to estimate the hidden state vector $\boldsymbol{x}_k$ based on the measurement set $\boldsymbol{Y}_k$. In this article, we exclusively consider sequential Monte Carlo methods (i.e., particle filters) as explained in Introduction I.

The first issue is that the nominal nonlinear system model (1) might be uncertain [55, Chapter 1], [56, Chapter 9]; see Appendix A in the online supplementary materials for the concept of "model uncertainty". Specifically, for given k, at least one of the state evolution function $\boldsymbol{f}_k(\cdot, \cdot)$, state measurement function $\boldsymbol{h}_k(\cdot, \cdot)$, and types and/or parameters of distributions of $\boldsymbol{w}_k$ and $\boldsymbol{v}_k$ might be inexact. In designing robust state estimation solutions that are insensitive to these uncertainties, the challenge is quantifying and bounding such modeling uncertainties. Special cases when (1) takes linear forms have been discussed in the author's previous works [39], [40]. In this article, we exclusively investigate non-degenerate nonlinear cases. As measurements $\boldsymbol{y}_k$ sequentially arrive, we focus on a time-incremental state estimation problem, i.e., studying the problem at every k given the measurement set $\boldsymbol{Y}_k$ [19], [21]. Hence, it suffices to examine the following single-stage Bayesian inference problem:

$$\begin{cases} \boldsymbol{z} \sim \mathbb{P}_{\boldsymbol{z}} \\ \boldsymbol{x} = \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{w}) \\ \boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{v}) \end{cases} \tag{2}$$

where $\boldsymbol{z} := \boldsymbol{x}_{k-1} | \boldsymbol{Y}_{k-1}$ represents the conditional posterior state at $k-1$ given the past measurement set $\boldsymbol{Y}_{k-1}$, $\boldsymbol{x} := \boldsymbol{x}_k$ the state at k, $\boldsymbol{y} := \boldsymbol{y}_k$ the measurement at k, $\boldsymbol{w} := \boldsymbol{w}_{k-1}$ the process noise at $k-1$, and $\boldsymbol{v} := \boldsymbol{v}_k$ the measurement noise at k; nominal distributions $\mathbb{P}_{\boldsymbol{z}}$, $\mathbb{P}_{\boldsymbol{w}}$, and $\mathbb{P}_{\boldsymbol{v}}$ are known; nominal nonlinear functions $\boldsymbol{f}(\cdot, \cdot)$ and $\boldsymbol{h}(\cdot, \cdot)$ are known as well. The time index k is dropped to avoid notational clutter. In particle filters, all involved distributions $\mathbb{P}_{\boldsymbol{z}}$, $\mathbb{P}_{\boldsymbol{w}}$, $\mathbb{P}_{\boldsymbol{v}}$, $\mathbb{P}_{\boldsymbol{x}}$, and $\mathbb{P}_{\boldsymbol{y}}$ are represented/approximated by particles; they are discrete distributions. Specifically, for example, $p(\boldsymbol{z}) := \sum_{i=1}^{N_{\boldsymbol{z}}} u_{\boldsymbol{z}^i} \cdot \delta_{\boldsymbol{z}^i}(\boldsymbol{z})$ where $N_{\boldsymbol{z}}$ is the number of particles; particles $\boldsymbol{z}^i$ are sampled from $\mathbb{P}_{\boldsymbol{z}}$ and $u_{\boldsymbol{z}^i}$ are weights. Since uncertain state equation (resp. uncertain measurement equation) would let the true prior state distribution $\mathbb{P}_{\boldsymbol{x}}$ (resp. true likelihood distribution $\mathbb{P}_{\boldsymbol{y}|\boldsymbol{x}}$) deviate from the nominal prior state distribution $\bar{\mathbb{P}}_{\boldsymbol{x}}$ (resp. nominal likelihood distribution $\bar{\mathbb{P}}_{\boldsymbol{y}|\boldsymbol{x}}$), particle filters can be robustified by considering that prior state distributions (resp. likelihood distributions) are uncertain, and finding worst-case state priors (resp. likelihoods). To be specific, we propose to find the worst-case prior state distribution near the nominal prior state distribution $\bar{\mathbb{P}}_{\boldsymbol{x}}$ to generate worst-case prior state particles $\boldsymbol{x}^j$ and/or update their weights. Likewise, worst-case likelihood distributions near the nominal ones $\bar{\mathbb{P}}_{\boldsymbol{y}|\boldsymbol{x}^j}$ are leveraged to evaluate the worst-case likelihoods of prior state particles $\boldsymbol{x}^j$ at the measurement $\boldsymbol{y}$. The principle of maximum entropy supports us to explore and exploit the maximum entropy distribution when given limited distributional information. Since the limited (i.e., inexact) distributional information of prior state is conveyed in the nominal $\bar{\mathbb{P}}_{\boldsymbol{x}}$, the following optimization problem has to be solved:

$$\begin{aligned} \max_{p(\boldsymbol{x}) \in L^1} \quad & \int -p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} \\ s.t. \quad & \begin{cases} D\left(\mathbb{P}_{\boldsymbol{x}}, \bar{\mathbb{P}}_{\boldsymbol{x}}\right) & \le \theta \\ \int p(\boldsymbol{x}) d\boldsymbol{x} & = 1 \end{cases} \end{aligned} \tag{3}$$

where the objective is the entropy of $\mathbb{P}_{\boldsymbol{x}}$ whose density is $p(\boldsymbol{x})$, and $D(\mathbb{P}_{\boldsymbol{x}}, \bar{\mathbb{P}}_{\boldsymbol{x}})$ is a statistical similarity measure between $\mathbb{P}_{\boldsymbol{x}}$ and $\bar{\mathbb{P}}_{\boldsymbol{x}}$. When $\mathbb{P}_{\boldsymbol{x}}$ is also assumed to be discrete [i.e., $p(\boldsymbol{x}) := \sum_{j=1}^{M} u_{\boldsymbol{x}^j} \delta_{\boldsymbol{x}^j}(\boldsymbol{x})$], the following alternative problem needs to be solved:

$$\begin{aligned} \max_{p(\boldsymbol{x}) \in l^1} \quad & \sum_{j=1}^{M} -p\left(\boldsymbol{x}^j\right) \ln p\left(\boldsymbol{x}^j\right) \\ s.t. \quad & \begin{cases} D\left(\mathbb{P}_{\boldsymbol{x}}, \bar{\mathbb{P}}_{\boldsymbol{x}}\right) & \le \theta \\ \sum_j p\left(\boldsymbol{x}^j\right) & = 1, \end{cases} \end{aligned} \tag{4}$$

where $p(\boldsymbol{x}^j) := u_{\boldsymbol{x}^j}$ denotes the weight of $\boldsymbol{x}^j$. Note that the support sets of the uncertain prior $p(\boldsymbol{x}) := \sum_{j=1}^{M} u_{\boldsymbol{x}^j} \delta_{\boldsymbol{x}^j}(\boldsymbol{x})$ and the nominal prior $\bar{p}(\boldsymbol{x}) := \sum_{i=1}^{N} u_{\boldsymbol{x}^i} \delta_{\boldsymbol{x}^i}(\boldsymbol{x})$ may not be the same: $p(\boldsymbol{x})$ is supported on $\{\boldsymbol{x}^j\}$ for $j \in [M]$, while $\bar{p}(\boldsymbol{x})$ is supported on $\{\boldsymbol{x}^i\}$ for $i \in [N]$. We call $\{\boldsymbol{x}^i\}_{i \in [N]}$ the nominal prior state particles and $\{\boldsymbol{x}^j\}_{j \in [M]}$ the worst-case prior state particles. Suppose that $p^*(\boldsymbol{x})$ solves (3). The worst-case prior state particles $\boldsymbol{x}^j$ can be sampled from $p^*(\boldsymbol{x})$. If $p^*(\boldsymbol{x})$ solves (4), $\boldsymbol{x}^j$ are worst-case prior state particles whose weights are $p^*(\boldsymbol{x}^j)$, respectively. The philosophy handling inexact likelihood information conveyed in nominal likelihood distributions $\bar{p}(\boldsymbol{y}|\boldsymbol{x}^j) := \sum_{r=1}^{R} u_{\boldsymbol{y}^r|\boldsymbol{x}^j} \delta_{\boldsymbol{y}^r|\boldsymbol{x}^j}(\boldsymbol{y})$ keeps consistent. Specifically, for every prior state particle $\boldsymbol{x}^j$, we need to solve

$$\begin{array}{ll} \max\limits_{p_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y})\in L^1} & \displaystyle\int -p_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y}) \ln p_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y}) d\boldsymbol{y} \\ s.t. & \begin{cases} D\left(\mathbb{P}_{\boldsymbol{y}|\boldsymbol{x}^j}, \bar{\mathbb{P}}_{\boldsymbol{y}|\boldsymbol{x}^j}\right) & \le \theta \\ \int p_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y}) d\boldsymbol{y} & = 1 \end{cases} \end{array} \tag{5}$$

or its discrete version

$$\begin{array}{ll} \max\limits_{p_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y})\in l^1} & \displaystyle\sum_{t=1}^{T} -p_{\boldsymbol{y}|\boldsymbol{x}^j}\left(\boldsymbol{y}^t\right) \ln p_{\boldsymbol{y}|\boldsymbol{x}^j}\left(\boldsymbol{y}^t\right) \\ s.t. & \begin{cases} D\left(\mathbb{P}_{\boldsymbol{y}|\boldsymbol{x}^j}, \bar{\mathbb{P}}_{\boldsymbol{y}|\boldsymbol{x}^j}\right) & \le \theta \\ \sum_t p_{\boldsymbol{y}|\boldsymbol{x}^j}\left(\boldsymbol{y}^t\right) & = 1. \end{cases} \end{array} \tag{6}$$

Likewise, the support sets of the uncertain likelihood distributions $p_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y}) := \sum_{t=1}^{T} u_{\boldsymbol{y}^t|\boldsymbol{x}^j} \delta_{\boldsymbol{y}^t|\boldsymbol{x}^j}(\boldsymbol{y})$ and the nominal likelihood distributions $\bar{p}_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y}) := \sum_{r=1}^{R} u_{\boldsymbol{y}^r|\boldsymbol{x}^j} \delta_{\boldsymbol{y}^r|\boldsymbol{x}^j}(\boldsymbol{y})$ may not be the same: $p_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y})$ is supported on $\{\boldsymbol{y}^t\}$ for $t \in [T]$, while $\bar{p}_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y})$ is supported on $\{\boldsymbol{y}^r\}$ for $r \in [R]$. We call $\{\boldsymbol{y}^r\}_{r\in[R]}$ the nominal likelihood particles and $\{\boldsymbol{y}^t\}_{t\in[T]}$ the worst-case likelihood particles. Suppose that $p^*_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y})$ solves (5). The worst-case likelihood of the prior state particle $\boldsymbol{x}^j$ given the measurement $\boldsymbol{y}$ can be evaluated by $p^*_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y})$. Instead, if $p^*_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y})$ solves (6) and one of $\boldsymbol{y}^t$ is the same as the given measurement $\boldsymbol{y}$, the worst-case likelihood of the prior state particle $\boldsymbol{x}^j$ given the measurement $\boldsymbol{y}$ can be evaluated by $p^*_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y}^t)$. This is possible because we can let the collected $\boldsymbol{y}$ be a supporting point to solve (6); i.e., $\boldsymbol{y} \in \{\boldsymbol{y}^t\}_{t\in[T]}$. Note that the support set $\{\boldsymbol{y}^t\}_{t\in[T]}$ is specified by filter designers.

The second issue is to evaluate likelihoods of particles $\boldsymbol{x}^j$ given the measurement $\boldsymbol{y}$ when the measurement noise $\boldsymbol{v}$ is non-additive and non-multiplicative. If the measurement noise is additive [i.e., $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{v}$] or multiplicative [i.e., $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}) \cdot \boldsymbol{v}$], the evaluation method is straightforward. For this reason, virtually all of the existing state-estimation literature tacitly takes the premise of additive/multiplicative measurement noises, which, however, is not always tenable in practice. This article, therefore, also aims to study a likelihood evaluation method for a general nonlinear measurement equation with non-additive and non-multiplicative measurement noises. This can be done by (5) and (6), in a robust (i.e., worst-case) sense. To be specific, we can first use the measurement equation $\boldsymbol{y}^r = \boldsymbol{h}(\boldsymbol{x}^j, \boldsymbol{v}^r)$, where $\boldsymbol{v}^r$ is a particle sampled from $\mathbb{P}_{\boldsymbol{v}}$, to generate nominal likelihood particles $\{\boldsymbol{y}^r\}_{r\in[R]}$ for the prior state particle $\boldsymbol{x}^j$, and then use (5) and (6) to evaluate the worst-case likelihood of $\boldsymbol{x}^j$ at the measurement $\boldsymbol{y}$.

The third issue is to identify possible outliers in measurements and take actions to remove/attenuate them [40]. Motivated by the M-estimation theory [37], we claim that this can be done by evaluating the likelihoods of prior state particles at the given measurement: if the largest likelihood of the prior state particles is smaller than a threshold (e.g., 5%), we treat this measurement as an outlier because none of these prior state particles can possibly generate this measurement. Then, this measurement can be directly trashed and all prior state particles directly become posterior (cf. re-descending influence functions, e.g., Hampel's [37, Eq. (4.90)], in M-estimation). This measurement can also be replaced by the nearest likelihood particle generated by the prior state particle that has the largest likelihood (cf. monotonic influence functions, e.g., Huber's [37, Eq. (4.53)], in M-estimation).

To the core, this article needs to find solutions of (3), (4), (5), and (6). In the following sections, we first explicitly choose eligible forms of the statistical similarity measure $D(\cdot, \cdot)$. Then, we find maximum entropy distributions for generating worst-case prior state particles and evaluating their worst-case likelihoods. Third, we identify and handle measurement outliers. At last, the overall distributionally robust state estimation framework for nonlinear systems is outlined.

III. Find Maximum Entropy Distributions

Mathematically, (3) and (5) are the same problem, so are (4) and (6). The former is a maximum entropy problem for a continuous distribution family given a discrete reference distribution, while the latter is a maximum entropy problem for a discrete distribution family given a discrete reference distribution. Therefore, for notational simplicity, we investigate a unified form for (3) and (5):

$$\begin{array}{ll} \max\limits_{p(\boldsymbol{x})\in L^1} & \displaystyle\int -p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} \\ s.t. & \begin{cases} D[p(\boldsymbol{x}), q(\boldsymbol{x})] \le \theta \\ \int p(\boldsymbol{x}) d\boldsymbol{x} \quad = 1 \end{cases} \end{array} \tag{7}$$

where $q(\boldsymbol{x}) = \sum_{i=1}^{N} u_{\boldsymbol{x}^i} \delta_{\boldsymbol{x}^i}(\boldsymbol{x})$ is a N-point discrete reference distribution; $q(\boldsymbol{x})$ is the density function of $\mathbb{Q}_{\boldsymbol{x}}$. Likewise, supposing $p(\boldsymbol{x}) = \sum_{j=1}^{M} u_{\boldsymbol{x}^j} \delta_{\boldsymbol{x}^j}(\boldsymbol{x})$ is a M-point discrete distribution (which is the density function of $\mathbb{P}_{\boldsymbol{x}}$), the unified form for (4) and (6) is

$$\begin{array}{ll} \max\limits_{\boldsymbol{p}\in l^1} & \displaystyle\sum_{j=1}^{M} -p_j \ln p_j \\ s.t. & \begin{cases} D[\boldsymbol{p}, \boldsymbol{q}] & \le \theta \\ \sum_{j=1}^{M} p_j & = 1 \end{cases} \end{array} \tag{8}$$

where $p_j := u_{\boldsymbol{x}^j}$, $q_i := u_{\boldsymbol{x}^i}$, $\boldsymbol{p} := [p_1, p_2, \ldots, p_j, \ldots p_M]^T$, and $\boldsymbol{q} := [q_1, q_2, \ldots, q_i, \ldots q_N]^T$. In (8), M might be equal to N but this is not always the case. Besides, even when $M = N$, $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{Q}_{\boldsymbol{x}}$ can be supported on different discrete sets; the former is $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$ and the latter is $\{\boldsymbol{x}^i\}_{i=1,2,\ldots,N}$.

Therefore, it suffices to consider only (7) and (8) in this section. In state-of-the-art distributionally robust optimization literature, the most commonly adopted statistical similarity measures are moments-based similarity [57], [58], Wasserstein distance [38], and ϕ-divergence [59]. We find the solutions to (7) and (8) based on these three statistical similarity measures, respectively. In the following sections, to avoid notational clutter, we no longer emphasize that a density function is in L^1 or a mass function is in l^1; they are implicitly admitted instead.

A. Solutions Using Moments-Based Similarity

Moments-Based statistical similarity claims that two random vectors are similar (in distribution) if they have similar moments up to the order of O (e.g., when $O = 2$, two random vectors have the same mean and covariance). This measure is also widely used in information theory [49, Chapter 11]. Note that the moments of the discrete reference distribution $\mathbb{Q}_{\boldsymbol{x}}$ can be estimated from its particles (using any eligible approaches, e.g., weighted sample mean and weighted sample covariance). Suppose the first two sample moments of $\mathbb{Q}_{\boldsymbol{x}}$ are given by $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}} := \sum_{i=1}^{N} u_{\boldsymbol{x}^i} \cdot \boldsymbol{x}^i$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} := \sum_{i=1}^{N} u_{\boldsymbol{x}^i} \cdot (\boldsymbol{x}^i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(\boldsymbol{x}^i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})^T$, respectively.

1) Solution to (7): The theorem below gives the continuous maximum entropy distribution when the first two moments are specified.

Theorem 1: If the first two moments of an absolutely continuous distribution $\mathbb{P}_{\boldsymbol{x}}$ are $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$, respectively, then the maximum entropy of $\mathbb{P}_{\boldsymbol{x}}$ is obtained by a Gaussian with mean $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$.

Proof: See [49, Theorem 9.6.5] or [60, Theorem 4.1.2]. Note that a Gaussian distribution is translation-invariant, and absolute continuity of $\mathbb{P}_{\boldsymbol{x}}$ implies the existence of its density almost everywhere. □

Theorem 1 can be extended to take into account higher order moments; see [49, Section 11.1]. We do not consider moments with orders equal to or higher than 3 because they are tensors for multivariate problems, and they are unnecessary for this article's contexts. Theorem 1 reveals the distributional robustness of the Gaussian approximation state estimators.

Corollary 1: The Gaussian approximation state estimators for nonlinear systems are distributionally robust in the sense that they use maximum entropy distributions for prior states and their likelihoods. □

Corollary 1 implies that when the nominal nonlinear system model is uncertain, Gaussian approximation filters, such as unscented Kalman filter (UKF), cubature Kalman filter (CKF), and Ensemble Kalman filter (EnKF), might outperform general particle filters. The benefit of such Gaussian approximation is that the induced filters (UKF, CKF, EnKF, etc.) are, strictly speaking, no longer computationally-intensive sequential Monte Carlo methods because they do not store prior state particles and explicitly evaluate their likelihoods. Instead, states and measurements are assumed to be marginally Gaussian and also jointly Gaussian, and therefore, closed-form solutions (i.e., canonical Kalman iterations) are applicable, which are computationally attractive.

In this sense, the philosophy of Gaussian approximation can also be applied in general particle filtering procedure. Specifically, we first sample (worst-case) prior state particles from the found maximum-entropy Gaussian prior distribution, and then evaluate their likelihoods using the found maximum-entropy Gaussian likelihood distributions. Finally, the posterior state particles can be generated. In fact, it is also possible to directly discover a discrete maximum-entropy Gaussian distribution supported on $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$ without sampling from a continuous Gaussian.

2) Solution to (8): In this subsection, we discuss the discrete maximum entropy distribution that is supported on the discrete set $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$, when the first two moments $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ are given.

Theorem 2: Among all discrete distributions supported on $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$ with first two moments $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$, the maximum entropy distribution is

$$p_j = \exp\left\{-1 - \gamma - \boldsymbol{\lambda}^T \boldsymbol{x}^j - \left(\boldsymbol{x}^j - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}\right)^T \boldsymbol{\Lambda}^T \left(\boldsymbol{x}^j - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}\right)\right\}, \tag{9}$$

$\forall j \in [M]$, where $\gamma \in \mathbb{R}^1$, $\boldsymbol{\lambda} \in \mathbb{R}^n$, and $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times n}$ are determined by the following three equalities

$$\begin{cases} \sum_{j=1}^{M} p_j = 1, \\ \sum_{j=1}^{M} \boldsymbol{x}^j \cdot p_j = \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}, \\ \sum_{j=1}^{M} \left(\boldsymbol{x}^j - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}\right)\left(\boldsymbol{x}^j - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}\right)^T \cdot p_j = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}. \end{cases} \tag{10}$$

Proof: Applying the Lagrange multiplier method to (8), the statements are immediate. □

Theorem 2 gives the worst-case weights of particles $\boldsymbol{x}^j$; i.e., $u_{\boldsymbol{x}^j} = p_j$. Therefore, particles $\boldsymbol{x}^j$ together with their weights $u_{\boldsymbol{x}^j}$ represent a worst-case prior state distribution [cf. (4)] or a worst-case likelihood distribution [cf. (6)]. The nonlinear root-finding problem (10) is, however, complicated even when only the first two moments are considered and only equalities are involved. If higher order moments and inequalities exist in (8), the complexity would be inconceivable (due to, e.g., tensors). However, the solution of (10) is just theoretically meaningful. In practice, when we take Gaussian assumption, it is pointless to store particles and evaluate their likelihoods; we prefer to apply canonical closed-form Kalman iterations.

B. Solutions Using Wasserstein Distance

The origin of the Wasserstein distance (i.e., Kantorovich-Rubinshtein metric) was inspired by the optimal transport theory [61]; see also [62]. It is currently of most interests in operations research [63] and machine learning [38], [64]. For any two distributions $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{Q}_{\boldsymbol{x}}$, the Wasserstein distance is defined as [61], [63]

$$\mathrm{W}(\mathbb{P}_{\boldsymbol{x}}, \mathbb{Q}_{\boldsymbol{x}}) := \inf_{\Pi_{\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}}} \int \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \Pi(d\boldsymbol{x}_{\mathbb{P}}, d\boldsymbol{x}_{\mathbb{Q}}) \tag{11}$$

where $\boldsymbol{x}_{\mathbb{P}}$ and $\boldsymbol{x}_{\mathbb{Q}}$ are random vectors associated with $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{Q}_{\boldsymbol{x}}$, respectively; $\Pi_{\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}}$ is any possible joint distribution of $(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}})$ whose marginals are $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{Q}_{\boldsymbol{x}}$; $\|\cdot\|$ denotes any possible vector norm. The benefit to use the Wasserstein distance is that it does not require the two involved distributions to have the same support. In other words, it is possible that either $\mathbb{P}_{\boldsymbol{x}}$ or $\mathbb{Q}_{\boldsymbol{x}}$ is continuous and the other one is discrete. Besides, the Wasserstein distance can also implicitly take higher-order-moment information of random vectors into consideration, unlike the Gaussian assumption that only focuses on the first two moments. In this article, as claimed, $\mathbb{Q}_{\boldsymbol{x}}$ is discrete and supported on $\{\boldsymbol{x}^i\}_{i=1,2,\ldots,N}$.

1) Solution to (7): Suppose $\mathbb{P}_{\boldsymbol{x}}$ and $\Pi_{\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}}$ are absolutely continuous, and the density of $\Pi_{\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}}$ is $\pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}})$; $\pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}) = I(\boldsymbol{x}_{\mathbb{Q}}|\boldsymbol{x}_{\mathbb{P}})p(\boldsymbol{x}_{\mathbb{P}})$ where $I(\boldsymbol{x}_{\mathbb{Q}}|\boldsymbol{x}_{\mathbb{P}})$ is the conditional density. We solve (7) using the Wasserstein distance. Hence, (7) can be written as

$$\begin{aligned} \max_{p(\boldsymbol{x})} \quad & \int -p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} \\ s.t. \quad & \begin{cases} \inf_{\pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}) d\boldsymbol{x}_{\mathbb{P}} d\boldsymbol{x}_{\mathbb{Q}} & \le \theta \\ \int p(\boldsymbol{x}) d\boldsymbol{x} & = 1. \end{cases} \end{aligned} \tag{12}$$

Note that $p(\boldsymbol{x}_{\mathbb{P}}) = p(\boldsymbol{x})$ and $p(\boldsymbol{x}_{\mathbb{Q}}) = q(\boldsymbol{x})$.

We first study the constraint $\inf_{\pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}) d\boldsymbol{x}_{\mathbb{P}} d\boldsymbol{x}_{\mathbb{Q}} \le \theta$. The infimum optimization

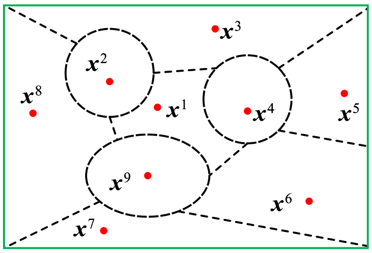


Fig. 1. The whole rectangular region C is divided into 9 sub-regions C_1, C_2,..., and C_9 whose centres (red dots) are $\boldsymbol{x}^1$, $\boldsymbol{x}^2$,..., and $\boldsymbol{x}^9$, respectively. Boundaries of sub-regions are marked by dashed lines.

problem on the left hand side of this constraint is functional and infinite-dimensional. Therefore, we aim to transform it into a vector-valued and finite-dimensional equivalent.

Lemma 1: The infinite-dimensional optimization problem $\inf_{\pi(\boldsymbol{x}_{\mathbb{P}},\boldsymbol{x}_{\mathbb{Q}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}) d\boldsymbol{x}_{\mathbb{P}} d\boldsymbol{x}_{\mathbb{Q}}$ is equivalent to a finite-dimensional optimization problem

$$\max_{\boldsymbol{\lambda}} \int p(\boldsymbol{x}) \min_{i\in[N]} \left\{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\right\} d\boldsymbol{x} + \sum_{i=1}^{N} q_i \lambda_i, \tag{13}$$

where $\boldsymbol{\lambda} := [\lambda_1, \lambda_2, \ldots, \lambda_N]^T$ and $\forall i \in [N], \lambda_i \in \mathbb{R}^1$.

Proof: See Appendix B in the online supplementary materials. □

We identify that (13) is a continuous-region partitioning problem for optimal transport [65]; intuitions can be found in Appendix B. Specifically, (13) is equivalent to

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & \int p(\boldsymbol{x})\sigma(\boldsymbol{x})d\boldsymbol{x} + \sum_{i=1}^{N} q_i\lambda_i \\ s.t. \quad & \begin{cases} \sigma(\boldsymbol{x}) = \min_{i\in[N]}\{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\}, & \\ \qquad\quad \le \|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i, & \forall i \in [N], \\ \sigma(\boldsymbol{x}) \ge 0, & \end{cases} \end{aligned} \tag{14}$$

which has the same form with [65, Eq. (5)]. Note that in [65, Eq. (4)], an auxiliary variable t was used, which introduced λ_i to [65, Eq. (5)]. (In [65], if t were cancelled, λ_i would disappear.) Note also that in [65], a generic measure dA was used. In the contexts of this article, it is instantiated to $dA := p(\boldsymbol{x})d\boldsymbol{x}$. Therefore, for any given $p(\boldsymbol{x})$ and $q(\boldsymbol{x}) = \sum_i q_i\delta_{\boldsymbol{x}^i}(\boldsymbol{x})$, an optimal partition exist [65]. For illustration, see Fig. 1, in which we suppose that $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are distributed over the whole rectangular region. However, $q(\boldsymbol{x})$ is discrete ($N = 9$), and supported on nine red dots. The optimal solution states that the optimal transport plan is to move all density of $p(\boldsymbol{x})$ in C_i to its centre $\boldsymbol{x}^i$. In other words, any density outside of C_i will strictly not be accepted at $\boldsymbol{x}^i$. Intuitively, this renders $\int I(\boldsymbol{x}^i|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = q_i, \forall i \in [N]$, and $I(\boldsymbol{x}^i|\boldsymbol{x})$ is in fact an indicator: $I(\boldsymbol{x}^i|\boldsymbol{x}) = 1$ if $\boldsymbol{x} \in C_i$ and $I(\boldsymbol{x}^i|\boldsymbol{x}) = 0$ otherwise (cf. Appendix B). Therefore, $\int_{\mathbb{R}^n} p(\boldsymbol{x})d\boldsymbol{x} = \sum_{i=1}^{N}\int_{C_i} p(\boldsymbol{x})d\boldsymbol{x} = \sum_{i=1}^{N} q_i = 1$.

Lemma 1 transforms (12) to

$$\begin{aligned} \max_{p(\boldsymbol{x})} \quad & \int -p(\boldsymbol{x})\ln p(\boldsymbol{x})d\boldsymbol{x} \\ s.t. \quad & \begin{cases} \displaystyle\max_{\boldsymbol{\lambda}} \int p(\boldsymbol{x}) \min_{i\in[N]} \left\{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\right\} d\boldsymbol{x} + \sum_{i=1}^{N} q_i\lambda_i \le \theta \\ \displaystyle\int p(\boldsymbol{x})d\boldsymbol{x} \qquad\qquad\qquad\qquad = 1. \end{cases} \end{aligned} \tag{15}$$

The solution to problem (15) is given in the theorem below.

Theorem 3: The maximum entropy distribution solving (15) is

$$p(\boldsymbol{x}) = \exp\left\{-v_0 \min_{i\in[N]} \left\{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\right\} - v_1 - 1\right\} \tag{16}$$

where $v_0 \in \mathbb{R}^1$, $v_1 \in \mathbb{R}^1$, and $\lambda_i \in \mathbb{R}^1, \forall i$ solve the following convex and smooth problem (n.b., almost-everywhere smooth in terms of λ_i; non-smooth only on zero-measure boundaries):

$$\begin{aligned} \min_{v_0, v_1, \boldsymbol{\lambda}} \quad & v_0 \cdot \left(\theta - \textstyle\sum_{i=1}^{N} \lambda_i q_i\right) + v_1 + \\ & \int \exp\left\{-v_0 \min_{i\in[N]} \left\{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\right\} - v_1 - 1\right\} d\boldsymbol{x} \\ s.t. \quad & v_0 \ge 0, \end{aligned} \tag{17}$$

where $\boldsymbol{\lambda} := [\lambda_1, \lambda_2, \ldots, \lambda_N]^T$.

Proof: See Appendix C in the online supplementary materials. □

Suppose that v_0^*, v_1^*, and $\boldsymbol{\lambda}^*$ solve (17). We claim that $p(\boldsymbol{x})$ in (16) admits

$$p(\boldsymbol{x}) = \exp\left\{-v_0^* \cdot \left\{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i^*\right\} - v_1^* - 1\right\}, \forall \boldsymbol{x} \in C_i, \tag{18}$$

where the sub-region/sub-space C_i is defined by

$$C_i := \left\{\boldsymbol{x} \in \mathbb{R}^n \big| \|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i^* \le \|\boldsymbol{x} - \boldsymbol{x}^j\| - \lambda_j^*\right\}, \forall j \ne i.$$

Note that $\{C_i\}_{i=1,2,\ldots,N}$ are collectively exhaustive and mutually exclusive; $C_i \bigcap C_j = \emptyset, \forall i \ne j$ and $\mathbb{R}^n = \bigcup_{i=1}^{N} C_i$.

Since (17) is convex[4] and smooth,[5] it can be solved using any first-order method (e.g., projected gradient descent). Let the objective of (17) be $f_{W-C}(v_0, v_1, \boldsymbol{\lambda})$; the subscripts "W" is for "Wasserstein" and "C" for "Continuous". By letting $g(\boldsymbol{x}, \boldsymbol{\lambda}) := \min_{i\in[N]}\{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\}$, the gradients of $f_{W-C}(v_0, v_1, \boldsymbol{\lambda})$ with respect to v_0, v_1, and λ_i are, respectively,

$$\begin{aligned} & \frac{\partial f_{W-C}}{\partial v_0} \\ & \quad = \theta - \sum_{i=1}^{N} \lambda_i q_i - \int_{\mathbb{R}^n} g(\boldsymbol{x}, \boldsymbol{\lambda}) \exp\{-v_0 g(\boldsymbol{x}, \boldsymbol{\lambda}) - v_1 - 1\} d\boldsymbol{x}, \end{aligned} \tag{19}$$

$$\frac{\partial f_{W-C}}{\partial v_1} = 1 - \textstyle\int_{\mathbb{R}^n} \exp\{-v_0 g(\boldsymbol{x}, \boldsymbol{\lambda}) - v_1 - 1\} d\boldsymbol{x}, \tag{20}$$

and

$$\frac{\partial f_{W-C}}{\partial \lambda_i} = -v_0 q_i + v_0 \int_{C_i} \exp\{-v_0 g(\boldsymbol{x}, \boldsymbol{\lambda}) - v_1 - 1\} d\boldsymbol{x},$$

[4] In the Lagrange duality sense, dual problems of any primal problems are always concave (resp. convex), no matter whether the primal problems are convex (resp. concave) or not [66, Chapter 5]. One can verify this point on (17) themselves by the definition of convexity. Note that for every two bounded functions f_1 and f_2 that have the same support, $\min(f_1 + f_2) \ge \min f_1 + \min f_2$.

[5] Non-smoothness over zero-measure subsets does not matter. Whenever necessary, one can use sub-gradients instead.

$$= -v_0 q_i + v_0 \int_{C_i} \exp\{-v_0(\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i) - v_1 - 1\} d\boldsymbol{x}. \tag{21}$$

When the optimality reaches (i.e., all gradients vanish), (19) implies that the Wasserstein distance is strictly equal to the prescribed budget θ, (20) indicates that $p(\boldsymbol{x})$ in (16) is indeed a density that is integrated to unit, and (21) means that a partition for optimal transport exists (i.e., $\int_{C_i} p(\boldsymbol{x}) d\boldsymbol{x} = q_i$). The projection step is straightforward in the gradient descent procedure: whenever $v_0 < 0$, let $v_0 = 0$.

In the projected gradient descent procedure, all involved integrals can be approximated by numerical methods, e.g., global adaptive quadrature [67] or Monte Carlo integration [68], [69], whichever is easier to be implemented for specific problems.

2) Solution to (8): Suppose $\mathbb{P}_{\boldsymbol{x}}$ is also discrete and supported on $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$. We solve (8) using the Wasserstein distance. Hence, (8) can be written as

$$\begin{aligned} \max_{\boldsymbol{p}} \quad & \sum_{j=1}^{M} -p_j \ln p_j \\ s.t. \quad & \begin{cases} \inf_{\pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}) d\boldsymbol{x}_{\mathbb{P}} d\boldsymbol{x}_{\mathbb{Q}} \leq \theta \\ \sum_{j=1}^{M} p_j = 1, \end{cases} \end{aligned} \tag{22}$$

where $\boldsymbol{p} := [p_1, p_2, \ldots, p_j, \ldots, p_M]^T$.

We first study the constraint $\inf_{\pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}) d\boldsymbol{x}_{\mathbb{P}} d\boldsymbol{x}_{\mathbb{Q}} \leq \theta$. In fact, the infimum optimization problem on the left hand side of this constraint (i.e., the Wasserstein distance) can be reformulated.

Lemma 2: If both $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{Q}_{\boldsymbol{x}}$ are discrete, and supported on $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$ and $\{\boldsymbol{x}^i\}_{i=1,2,\ldots,N}$, respectively, the Wasserstein distance $\inf_{\pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}) d\boldsymbol{x}_{\mathbb{P}} d\boldsymbol{x}_{\mathbb{Q}}$ is equivalent to a linear program

$$\begin{aligned} \min_{P_{ij}} \; & \sum_{i=1}^{N} \sum_{j=1}^{M} \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \cdot P_{ij} \\ s.t. \; & \sum_{j=1}^{M} P_{ij} = q_i, \quad \forall i \in [N], \\ & \sum_{i=1}^{N} P_{ij} = p_j, \quad \forall j \in [M], \\ & P_{ij} \geq 0, \forall i \in [N], \quad \forall j \in [M]. \end{aligned} \tag{23}$$

In (23), P_{ij} denotes a joint discrete distribution supported on $\{(\boldsymbol{x}^i, \boldsymbol{x}^j)\}_{i \in [N], j \in [M]}$.

Proof: See Appendix D in the online supplementary materials. □

Intuitively, (23) can be seen as a optimal transport problem as well (cf. Lemma 1 and Fig. 1): the resources are discretely distributed on some given points $\{\boldsymbol{x}^j\}$, whereas facilities are fixed at $\{\boldsymbol{x}^i\}$.

Lemma 2 transforms (22) to

$$\begin{aligned} \max_{\boldsymbol{p}} \quad & \sum_{j=1}^{M} -p_j \ln p_j \\ s.t. \quad & \begin{cases} \min_{P_{ij}} \sum_{i=1}^{N} \sum_{j=1}^{M} \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \cdot P_{ij} \leq \theta \\ \sum_{j=1}^{M} P_{ij} = q_i, \quad \forall i \in [N], \\ \sum_{i=1}^{N} P_{ij} = p_j, \quad \forall j \in [M], \\ P_{ij} \geq 0, \forall i \in [N], \quad \forall j \in [M]. \end{cases} \end{aligned} \tag{24}$$

The constraint $\sum_{j=1}^{M} p_j = 1$ is dropped because it is redundant to (24).

Since the left hand side of the first constraint is a minimization problem, we can directly drop the minimization. Thus, (24) is equivalent to

$$\begin{aligned} \max_{P_{ij}} \quad & -\sum_{i=1}^{N} \sum_{j=1}^{M} P_{ij} \ln \sum_{i=1}^{N} P_{ij} \\ s.t. \quad & \begin{cases} \sum_{i=1}^{N} \sum_{j=1}^{M} \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \cdot P_{ij} \leq \theta \\ \sum_{j=1}^{M} P_{ij} = q_i, \forall i \in [N], \\ P_{ij} \geq 0, \forall i \in [N], \forall j \in [M]. \end{cases} \end{aligned} \tag{25}$$

The solution to problem (25) is given in theorem below.

Theorem 4: If there exists a discrete distribution $\{P^0_{ij}\}_{\forall i, \forall j}$ that strictly satisfies the inequality $\sum_{i=1}^{N} \sum_{j=1}^{M} \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \cdot P^0_{ij} < \theta$ and simultaneously satisfies the equality $\sum_{j=1}^{M} P^0_{ij} = q_i$, the maximum entropy distribution solving (25) also solves

$$\min_{v_0, \boldsymbol{\lambda}} \max_{P_{ij}} v_0 \theta + \sum_{i=1}^{N} \lambda_i q_i + \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{P^2_{ij}}{\sum_{i=1}^{N} P_{ij}}$$

$$s.t.$$

$$\begin{cases} -\ln\left(\sum_{i=1}^{N} P_{ij}\right) - \frac{P_{ij}}{\sum_{i=1}^{N} P_{ij}} - v_0 \|\boldsymbol{x}^i - \boldsymbol{x}^j\| - \lambda_i = 0, \\ \qquad \forall i \in [N], \forall j \in [M], \\ P_{ij} \geq 0, \quad \forall i \in [N], \forall j \in [M], \\ v_0 \geq 0, \end{cases} \tag{26}$$

where $\boldsymbol{\lambda} := [\lambda_1, \lambda_2, \ldots, \lambda_N]^T$.

Proof: See Appendix E in the online supplementary materials. □

The problem (26) is intuitively uneasy to be solved because P_{ij} has no closed-form expression. Therefore, we try to relax the original maximum entropy problem (25). Since the entropy of a joint distribution is no larger than the sum of the entropy of marginals [49, Theorem 2.6.6]; i.e.,

$$-\sum_{i=1}^{N} \sum_{j=1}^{M} P_{ij} \ln P_{ij} \leq -\sum_{j=1}^{M} p_j \ln p_j - \sum_{i=1}^{N} q_i \ln q_i$$

and $-\sum_{i=1}^{N} q_i \ln q_i$ is a constant, we can use the entropy of the join distribution as a surrogate for optimization. Whenever the entropy of the join distribution is maximized, the entropy of $p(\boldsymbol{x})$ is improved as well. [Of course, under this approximation, the entropy of $p(\boldsymbol{x})$ induced from the optimal P_{ij} is not guaranteed to be maximal as in (24).] As a result, (25) can be relaxed as follows.

$$\begin{aligned} \max_{P_{ij}} & \quad -\sum_{i=1}^{N}\sum_{j=1}^{M} P_{ij}\ln P_{ij} \\ s.t. & \quad \begin{cases} \sum_{i=1}^{N}\sum_{j=1}^{M}\|\boldsymbol{x}^i-\boldsymbol{x}^j\|\cdot P_{ij}\le\theta \\ \sum_{j=1}^{M} P_{ij}=q_i, \forall i\in[N]. \end{cases} \end{aligned} \tag{27}$$

The solution to (27) is given in the theorem below.

Theorem 5: If there exists a discrete distribution $\{P_{ij}^0\}_{\forall i,\forall j}$ that strictly satisfies the inequality $\sum_{i=1}^{N}\sum_{j=1}^{M}\|\boldsymbol{x}^i-\boldsymbol{x}^j\|\cdot P_{ij}^0<\theta$ and simultaneously satisfies the equality $\sum_{j=1}^{M}P_{ij}^0=q_i$, then the maximum entropy distribution solving (27) is

$$P_{ij}=\exp\left\{-v_0\|\boldsymbol{x}^i-\boldsymbol{x}^j\|-\lambda_i-1\right\}, \forall i\in[N], \forall j\in[M], \tag{28}$$

where $v_0\in\mathbb{R}^1$ and $\lambda_i\in\mathbb{R}^1, \forall i$ solve the following convex and smooth problem:

$$\begin{aligned} \min_{v_0,\boldsymbol{\lambda}} & \quad v_0\cdot\theta+\sum_{i=1}^{N}\lambda_i q_i \\ & \quad +\sum_{i=1}^{N}\sum_{j=1}^{M}\exp\left\{-v_0\|\boldsymbol{x}^i-\boldsymbol{x}^j\|-\lambda_i-1\right\} \\ s.t. & \quad v_0\ge 0, \end{aligned} \tag{29}$$

where $\boldsymbol{\lambda}:=[\lambda_1,\lambda_2,\ldots,\lambda_N]^T$. In addition, the marginal distribution is given as $p_j=\sum_{i=1}^{N}P_{ij}, \forall j\in[M]$.

Proof: Similar to the proof of Theorem 4. □

Since (29) is convex and smooth, it can be solved using any first-order method (e.g., projected gradient descent). Let the objective of (29) be $f_{W-D}(v_0,\boldsymbol{\lambda})$; the subscripts "W" is for "Wasserstein" and "D" for "Discrete". The gradients of $f_{W-D}(v_0,\boldsymbol{\lambda})$ with respect to v_0 and λ_i are, respectively,

$$\begin{aligned} &\frac{\partial f_{W-D}}{\partial v_0}= \\ &\theta-\sum_{i=1}^{N}\sum_{j=1}^{M}\|\boldsymbol{x}^i-\boldsymbol{x}^j\|\exp\left\{-v_0\|\boldsymbol{x}^i-\boldsymbol{x}^j\|-\lambda_i-1\right\}, \end{aligned} \tag{30}$$

and

$$\frac{\partial f_{W-D}}{\partial\lambda_i}=q_i-\sum_{j=1}^{M}\exp\left\{-v_0\|\boldsymbol{x}^i-\boldsymbol{x}^j\|-\lambda_i-1\right\}. \tag{31}$$

Likewise, when all gradients vanish, the minimum transport cost coincides with the prescribed Wasserstein budget θ, and an (discrete-version) optimal transport exists (i.e., $q_i=\sum_{j=1}^{M}P_{ij}$). The projection step is straightforward in the gradient descent procedure: whenever $v_0<0$, let $v_0=0$.

C. Solutions Using ϕ-Divergence

Suppose $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{Q}_{\boldsymbol{x}}$ have the same support $\mathcal{S}$. If $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{Q}_{\boldsymbol{x}}$ are absolutely continuous with respect to the Lebesgue measure and $\mathbb{P}_{\boldsymbol{x}}$ is absolutely continuous with respect to $\mathbb{Q}_{\boldsymbol{x}}$, then the ϕ-divergence of $\mathbb{P}_{\boldsymbol{x}}$ from $\mathbb{Q}_{\boldsymbol{x}}$ is defined as

$$\int_S\phi\left(\frac{d\mathbb{P}_{\boldsymbol{x}}}{d\mathbb{Q}_{\boldsymbol{x}}}\right)d\mathbb{Q}_{\boldsymbol{x}}=\int_S\phi\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)q(\boldsymbol{x})d\boldsymbol{x}, \tag{32}$$

where $\phi(t), t\ge 0$ is a convex function such that $\phi(1):=0$ and $0\phi(0/0):=0$; $d\mathbb{P}_{\boldsymbol{x}}/d\mathbb{Q}_{\boldsymbol{x}}$ is the Radon-Nikodym derivative. Alternatively, if $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{Q}_{\boldsymbol{x}}$ are discrete on the same support, the ϕ-divergence of $\boldsymbol{p}$ from $\boldsymbol{q}$ is defined as

$$\sum_{i=1}^{N}q_i\phi\left(\frac{p_i}{q_i}\right). \tag{33}$$

The ϕ-divergence is a generalization of the Kullback-Leibler divergence. Letting $\phi(t):=t\ln t$ or $\phi(t):=t\ln t-t+1$, the ϕ-divergence degenerates to the Kullback-Leibler divergence. Other possible choice of $\phi(t)$ can be found in, e.g., [59, Table 2]. For the demonstration purpose only, results in this section are only based on the Kullback-Leibler divergence. This is because the Kullback-Leibler divergence is the most popular one which also has clear physical meaning in information theory [48], [70]. Interested readers may try other $\phi(\cdot)$ themselves.

Since the reference distribution $\mathbb{Q}_{\boldsymbol{x}}$ in this article is limited to be discrete, it is pointless to consider the continuity of $\mathbb{P}_{\boldsymbol{x}}$. Otherwise, $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{Q}_{\boldsymbol{x}}$ would have discrepant supports so that the ϕ-divergence is undefined. Thus, we only study the solution to (8) when $\mathbb{P}_{\boldsymbol{x}}$ is discrete and neglect the continuous case (7).

1) Solution to (8): We solve (8) using the Kullback-Leibler divergence. Hence, (8) can be written as

$$\begin{aligned} \max_{\boldsymbol{p}} & \quad \sum_{i=1}^{N}-p_i\ln p_i \\ s.t. & \quad \begin{cases} \sum_{i=1}^{N}p_i\ln\left(\frac{p_i}{q_i}\right) & \le\theta \\ \sum_{i=1}^{N}p_i & =1, \end{cases} \end{aligned} \tag{34}$$

where $\boldsymbol{p}:=[p_1,p_2,\ldots,p_j,\ldots,p_N]^T$ (n.b., $M=N$). The solution to (34) is outlined in the theorem below.

Theorem 6: The distribution solving (34) is given by

$$p_i=\exp\left\{\frac{-\lambda_0\ln(q_i)+\lambda_1}{-(\lambda_0+1)}-1\right\}, \forall i\in[N], \tag{35}$$

where $\lambda_0\in\mathbb{R}^1, \lambda_1\in\mathbb{R}^1$ solve the following the convex and smooth problem:

$$\begin{aligned} \min_{\lambda_0,\lambda_1} & \quad \lambda_0\theta+\lambda_1+(\lambda_0+1)\sum_{i=1}^{N}p_i \\ s.t. & \quad \lambda_0\ge 0. \end{aligned} \tag{36}$$

Proof: See Appendix F in the online supplementary materials. □

Since (36) is convex and smooth, it can be solved using any first-order method (e.g., projected gradient descent). Let the objective of (36) be $f_{KL-D}(\lambda_0,\lambda_1)$; the subscripts "KL" is for "Kullback-Leibler" and "D" for "Discrete". The gradients of $f_{KL-D}(\lambda_0,\lambda_1)$ with respect to λ_0 and λ_1 are, respectively,

$$\frac{\partial f_{KL-D}(\lambda_0,\lambda_1)}{\partial\lambda_0}=\theta+\sum_{i=1}^{N}\left[1+\frac{\ln(q_i)+\lambda_1}{\lambda_0+1}\right]p_i, \tag{37}$$

and

$$\frac{\partial f_{KL-D}(\lambda_0,\lambda_1)}{\partial\lambda_1}=1-\sum_{i=1}^{N}p_i. \tag{38}$$

Likewise, when the optimality reaches, the Kullback-Leibler divergence between $\boldsymbol{p}$ and $\boldsymbol{q}$ conincides with the prescribed budget θ, and the sum of $\boldsymbol{p}$ is unit. The projection step is straightforward in the gradient descent procedure: whenever $\lambda_0<0$, let $\lambda_0=0$.

D. Comparisons for the Three Statistical Similarity Measures

As we can see, the moments-based similarity and Wasserstein distance do not require that the two distributions to have the same support. Therefore, a discrete distribution and a continuous distribution can be discussed in a same maximum entropy problem, so can be two discrete distributions with different supports. In addition, the advantage of the Wasserstein distance and the ϕ-divergence is that they can implicitly take into account high-order moments of random variables even for multivariate problems. However, using the Wasserstein distance and the ϕ-divergence implies that computationally intensive numerical problems have to be solved (cf. Theorem 3, Theorem 5, and Theorem 6). Instead, using the moments-based similarity gives the Gaussian approximation state estimation framework which means that closed-form solutions exist (i.e., canonical Kalman iterations).

E. Projected Gradient Descent Algorithm for Maximum Entropy Problems

Since all maximum entropy problems subject to the Wasserstein distance and the ϕ-divergence can be solved by the projected gradient descent algorithm, we depict it in Algorithm 1. Without loss of generality, we use the problem under the Kullback-Leibler divergence [i.e. (34)] as an example; see Theorem 6.

Algorithm 1: Projected Gradient Descent Method for Maximum Entropy Problem Under the Kullback-Leibler Divergence.

```
Definition: $S$ as maximum allowed iteration steps and $s$ the
  current iteration step; $\alpha$ as step size; $\epsilon$ as numerical
  precision threshold; $\mathrm{abs}(\cdot)$ returns absolute value.
Remark: Since (36) is convex, in principle, any initial values
  for $\lambda_0 \geq 0$ and $\lambda_1$ are acceptable. If early stopping is
  applied (i.e., $S$ is not sufficiently large for time-saving
  purpose), a normalization procedure is necessary to
  guarantee $1 = \sum_i p_i$.
Input: $S, \alpha, \epsilon, \lambda_0, \lambda_1$
$s \leftarrow 0$
while true do
  // Gradient Descent
  $\lambda_0 \leftarrow \lambda_0 - \alpha \cdot \frac{\partial f_{KL-D}}{\partial \lambda_0}$     // See (37)
  $\lambda_1 \leftarrow \lambda_1 - \alpha \cdot \frac{\partial f_{KL-D}}{\partial \lambda_1}$     // See (38)
  // Projection
  if $\lambda_0 < 0$ then
    $\lambda_0 \leftarrow 0$
  end if
  // Next Iteration
  $s \leftarrow s + 1$;
  if $s > S$ or $\mathrm{abs}(\frac{\partial f_{KL-D}}{\partial \lambda_1}) < \epsilon$ then
    if $1 \neq \sum_i p_i$ then     // Early Stopping Applied
      $p_i \leftarrow p_i / \sum_j p_j$     // Normalization
    end if
    Exit Algorithm
  end if
end while
Output: $p_i$ in (35)
```

IV. Distributionally Robust State Estimation Framework for Nonlinear Systems

This section outlines the overall distributionally robust particle-based state estimation method.

A. Generate Worst-Case Prior State Particles

We use the solutions to (3) and (4) to generate worst-case prior state particles. Solutions under the moments-based similarity measure are just used to argue for the distributional robustness of the Gaussian approximation framework; see Corollary 1. Therefore, we do not cover them in this subsection. Suppose the worst-case prior state particles are $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$.

First, we suppose $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$ are preset and only their weights are expected to be updated. For example, we can let $M := N$ and $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$ be a copy of $\{\boldsymbol{x}^i\}_{i=1,2,\ldots,N}$. For another example, $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$ can be uniformly sampled from a subset of $\mathbb{R}^n$ and this subset is usually the smallest hyperrectangle or hyperellipsoid containing $\{\boldsymbol{x}^i\}_{i=1,2,\ldots,N}$. We have the following method for generating worst-case prior state particles.

Method 1: Given worst-case prior state particles $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$ and nominal prior state particles $\{\boldsymbol{x}^i\}_{i=1,2,\ldots,N}$,

1) If the two sets $\{\boldsymbol{x}^j\}$ and $\{\boldsymbol{x}^i\}$ are identical, the worst-case weights $u_{\boldsymbol{x}^j}$ of particles $\boldsymbol{x}^j$ can be determined by Theorem 5 or Theorem 6.
2) If the two sets are different, the worst-case weights $u_{\boldsymbol{x}^j}$ of particles $\boldsymbol{x}^j$ can be determined by Theorem 5.
3) No matter whether the two sets are identical or not, the worst-case weights $u_{\boldsymbol{x}^j}$ of particles $\boldsymbol{x}^j$ can also be determined by Theorem 3 by letting $u_{\boldsymbol{x}^j} \propto p(\boldsymbol{x}^j)$, where $p(\boldsymbol{x})$ is defined in (18). Note that in this case, a normalization procedure is necessary; $u_{\boldsymbol{x}^j} \leftarrow u_{\boldsymbol{x}^j} / \sum_j u_{\boldsymbol{x}^j}$. □

Second, we suppose $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$ are not preset. Hence, we can directly sample M particles from $p(\boldsymbol{x})$ in (18). Since $p(\boldsymbol{x})$ is defined in a partitioned region/space, the first step is to choose a sub-region, and the second step is to draw a worst-case prior state particle from this sub-region. We have the following method.

Method 2: First, draw an integer $j \in [N]$ according to the discrete reference distribution $\mathbb{Q}_{\boldsymbol{x}}$ (i.e., choose a sub-region C_j whose probability being chosen is q_j). Second, draw a sample $\boldsymbol{x}^j$ from C_j using $p(\boldsymbol{x})$ defined in (18). Repeat the two steps above M times to obtain M worst-case prior state particles. In this case, all particles $\boldsymbol{x}^j$ have the same weight $1/M$. □

At last, we highlight that the proposed approaches for worst-case prior state particle generation based on entropy-maximization strategy can counteract particle degeneracy. In fact, maximizing the entropy of a variable distribution implies minimizing the Kullback-Leibler divergence of this distribution from a uniform distribution. This can be seen from

$$-\sum_{j=1}^{M} p_j \ln p_j = \ln M - \sum_{j=1}^{M} p_j \ln \frac{p_j}{1/M}. \tag{39}$$

Therefore, the worst-case prior state particles have more balanced weights than the corresponding nominal prior state particles (n.b., uniformly distributed weights are most balanced). On the other hand,

$$-\sum_{j=1}^{M} p_j \ln p_j \geq \sum_{j=1}^{M} p_j(1-p_j) = 1 - \sum_{j=1}^{M} p_j^2. \tag{40}$$

It means that any methods reducing the variance of weights (i.e., improving the effective sample size) implicitly elevate the entropy of weights of prior state particles; cf. [18, Eq. (51)] or [26, Eq. (5)].

B. Evaluate Worst-Case Likelihoods

When the nominal measurement noise is additive, i.e., $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{v}$, the nominal likelihood distribution is $p_{\boldsymbol{v}}[\boldsymbol{y} - \boldsymbol{h}(\boldsymbol{x})]$. As a result, the worst-case likelihood distribution can be chosen near $p_{\boldsymbol{v}}[\boldsymbol{y} - \boldsymbol{h}(\boldsymbol{x})]$, and worst-case likelihood of a prior state particle (or a worst-case prior state particle; depending on whether the state equation is uncertain or not) given $\boldsymbol{y}$ can be evaluated accordingly. Likewise, when the nominal measurement noise is multiplicative, the nominal likelihood distribution is $p_{\boldsymbol{v}}[\boldsymbol{h}^{-1}(\boldsymbol{x}) \cdot \boldsymbol{y}]$ if $\boldsymbol{h}(\boldsymbol{x})$ is invertible. To be specific, we take a Gaussian case as an example to explain the worst-case likelihood evaluation method under additive and multiplicative measurement noises.

Method 3: If the nominal likelihood distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$ is $p_{\boldsymbol{v}}[\boldsymbol{y} - \boldsymbol{h}(\boldsymbol{x}); \boldsymbol{\mu}, \boldsymbol{\Sigma}]$ or $p_{\boldsymbol{v}}[\boldsymbol{h}^{-1}(\boldsymbol{x}) \cdot \boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}]$, and $p_{\boldsymbol{v}}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, then the worst-case likelihood distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$ is $p_{\boldsymbol{v}}(\cdot; \boldsymbol{\mu}, \theta\boldsymbol{\Sigma})$ where $\theta \geq 1$. □

By multiplying $\boldsymbol{\Sigma}$ by a scalar $\theta \geq 1$, a worst-case maximum-entropy likelihood distribution can be obtained because the entropy of the m-dimensional Gaussian distribution $p_{\boldsymbol{v}}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is $\frac{m}{2} + \frac{m}{2}\ln(2\pi) + \frac{1}{2}\ln(|\boldsymbol{\Sigma}|)$. Hence, improving the covariance implies raising the entropy. Method 3 can be straightforwardly extended to other noise distributions such as the Student's t distribution. We do not cover details here.

However, when the nominal measurement noise is non-additive and non-multiplicative, such closed-form evaluation methods are unavailable. Therefore, numerical methods are indispensable. The first step is to generate nominal likelihood particles $\{\boldsymbol{y}^r|\boldsymbol{x}^j\}_{r=1,2,\ldots,R}$ for each worst-case prior state particle $\boldsymbol{x}^j$ (n.b., when the state equation is exact, worst-case $\{\boldsymbol{x}^j\}_{j\in[M]}$ and nominal $\{\boldsymbol{x}^i\}_{i\in[N]}$ are the same). This can be done by the nominal measurement equation $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}^j, \boldsymbol{v})$. Specifically, we need to generate R samples from $\boldsymbol{v}$, say $\boldsymbol{v}^r$, and obtain $\{\boldsymbol{y}^r|\boldsymbol{x}^j\}$ by $\boldsymbol{y}^r|\boldsymbol{x}^j := \boldsymbol{h}(\boldsymbol{x}^j, \boldsymbol{v}^r), \forall r \in [R]$. Since $\boldsymbol{v}$ is high-dimensional, we use the importance sampling method [71, Section 11.1.4]: 1) uniformly draw R samples in the support of $\boldsymbol{v}$, and 2) use $p(\boldsymbol{v})$ to determine their weights; $u_{\boldsymbol{v}^r} \propto p(\boldsymbol{v}^r)$ (n.b., a normalization procedure is hence necessary). Based on nominal likelihood particles $\{\boldsymbol{y}^r|\boldsymbol{x}^j\}_{r=1,2,\ldots,R}$ whose weights are $u_{\boldsymbol{y}^r|\boldsymbol{x}^j} = u_{\boldsymbol{v}^r}$, the worst-case likelihood of $\boldsymbol{x}^j$ is ready to be evaluated. Suppose the support set of the worst-case likelihood distribution is $\{\boldsymbol{y}^t|\boldsymbol{x}^j\}_{t=1,2,\ldots,T}$. As the case that generates worst-case prior state particles in Section IV-A, $\{\boldsymbol{y}^t|\boldsymbol{x}^j\}_{t=1,2,\ldots,T}$ can be just a copy of $\{\boldsymbol{y}^r|\boldsymbol{x}^j\}_{r=1,2,\ldots,R}$ (thus $T := R$) or uniformly sampled from a subset of $\mathbb{R}^m$. The subset can be the smallest hyperrectangle or hyperellipsoid containing $\{\boldsymbol{y}^r|\boldsymbol{x}^j\}_{r=1,2,\ldots,R}$.

We have two methods to evaluate the worst-case likelihood of $\boldsymbol{x}^j$ given the measurement $\boldsymbol{y}$.

Method 4: Suppose $p^*_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y})$ solves (5) using the Wasserstein distance. The worst-case likelihood of $\boldsymbol{x}^j$ given the measurement $\boldsymbol{y}$ is $p^*_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y})$. □

Method 5: Augment $\boldsymbol{y}$ into the support sets of worst-case likelihood distributions for $\boldsymbol{x}^j, j \in [M]$; i.e., let $\{\boldsymbol{y}^t|\boldsymbol{x}^j\}_{t=1,2,\ldots,T+1} := \{\boldsymbol{y}\} \bigcup \{\boldsymbol{y}^t|\boldsymbol{x}^j\}_{t=1,2,\ldots,T}$. Suppose $p^*_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y})$ solves (6) using the Wasserstein distance (n.b., the Kullback-Leibler divergence is not applicable because after augmentation, the two support sets are hardly identical). The worst-case likelihood of $\boldsymbol{x}^j$ given the measurement $\boldsymbol{y}$ is $p^*_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y})$. □

Compared to Method 4 and Method 5, Method 3 is likely to be of more interest in engineering for two reasons: 1) many measurement equation are driven by additive measurement noises, and 2) the involved likelihood distribution has a closed-form expression which allows fast computation.

C. Outlier Treatment

In this subsection, we provide an outlier identification and treatment method for particle filtering framework. The outlier identification method is given below.

Method 6: If $\forall j \in [M]$, $p^*_{\boldsymbol{y}|\boldsymbol{x}^j}(\boldsymbol{y}) < \varepsilon$ where ε is a threshold, say 5%, then $\boldsymbol{y}$ is an outlier because there exists no any prior state particle that possibly generates this measurement. Alternatively, supposing the weighted mean of particles $\boldsymbol{x}^j$ is $\bar{\boldsymbol{x}} := \sum_{j=1}^{M} u_{\boldsymbol{x}^j} \cdot \boldsymbol{x}^j$, if $p^*_{\boldsymbol{y}|\bar{\boldsymbol{x}}}(\boldsymbol{y}) < \varepsilon$, then $\boldsymbol{y}$ is an outlier. □

The outlier treatment method is given below.

Method 7: The identified outlier can be directly trashed and all prior state particles directly become posterior, during which associated weights keep unchanged. This idea is motivated by re-descending influence functions in M-estimation, e.g., Hampel's influence function [37, Eq. (4.90)]. The outlier can also be replaced by the nearest likelihood particle generated by the prior state particle that has the largest likelihood or replaced by the nearest likelihood particle generated by the weighted mean. This philosophy is motivated by monotonic influence functions in M-estimation, e.g., Huber's influence function [37, Eq. (4.53)]. □

D. Overall Method

The distributionally robust particle filtering framework is summarized in Algorithm 2. Algorithm 2 is a robustified version of the popular canonical particle filter in [18, Algorithm 3]. The used proposal density (i.e., importance density) for importance sampling is the prior state distribution as in [18, Eq. (63)].

Algorithm 2: Distributionally Robust Particle Filtering.

```
Definition: $k$ as discrete time index; $N$ as number of nominal prior state particles; $M$ as number of worst-case prior (and also posterior) state particles; $R$ as number of nominal likelihood particles for every (worst-case) prior state particle, and $T$ as number of worst-case likelihood particles for the same (worst-case) prior state particle; $\boldsymbol{x}_0^i$ as posterior state particles at $k = 0$ and $u_{\boldsymbol{x}_0^i}$ the associated weights, $\forall i \in [N]$; $p^*(\boldsymbol{y}_k|\boldsymbol{x}_k^j)$ as worst-case likelihood of $\boldsymbol{x}_k^j$ given $\boldsymbol{y}_k$; $\hat{N}_{eff}$ as effective sample size and $N_{thres}$ its threshold.
Remark: If measurement noises are additive or multiplicative, ignore Step 3, and use Method 3 in Step 4. If there are no process model uncertainties, ignore Step 2. If resampling is applied at every time $k$, $M$ and $N$ can be different; cf. Line 30. Otherwise, $M$ and $N$ must be identical to guarantee the number of posterior state particles at time $k-1$ is the same as the number of prior state particles at time $k$; cf. Line 5.
Initialization: $N$, $M$, $R$, $T$, $N_{thres}$, and $\{\boldsymbol{x}_0^i, u_{\boldsymbol{x}_0^i}\}_{i\in[N]}$.
Input: $\boldsymbol{y}_k, k = 1, 2, 3, \ldots$
 1: For every $k$, execute the following 5 steps
 2: // Step 1: Generate Nominal Prior State Particles
 3: for $i = 1 : N$ do
 4:   Sample $\boldsymbol{w}_{k-1}^i$ from the distribution of $\boldsymbol{w}_{k-1}$
 5:   $\boldsymbol{x}_k^i = \boldsymbol{f}_k(\boldsymbol{x}_{k-1}^i, \boldsymbol{w}_{k-1}^i)$
 6: end for
 7: // Step 2: Obtain Worst-Case Prior State Particles
 8: Use Method 1 or Method 2 to generate worst-case prior state particles $\{\boldsymbol{x}_k^j\}_{j\in[M]}$ and obtain their weights $\{u_{\boldsymbol{x}_k^j}\}_{j\in[M]}$
 9: // Step 3: Evaluate Worst-Case Likelihood for Every $\boldsymbol{x}_k^j$
10: for $j = 1 : M$ do
11:   //Generate Nominal Likelihood Particles $\boldsymbol{y}_k^r, \forall r \in [R]$
12:   for $r = 1 : R$ do
13:     Sample $\boldsymbol{v}_k^r$ from the distribution of $\boldsymbol{v}_k$
14:     $\boldsymbol{y}_k^r = \boldsymbol{h}_k(\boldsymbol{x}_k^j, \boldsymbol{v}_k^r)$
15:   end for
16:   //Evaluate Worst-Case Likelihood of $\boldsymbol{x}_k^j$ at $\boldsymbol{y}_k$
17:   Use Method 4 or Method 5 for likelihood evaluation
18:   //Outlier Identification and Treatment
19:   Use Method 6 for outlier identification and Method 7 for outlier treatment
20: end for
21: // Step 4: Generate Posterior State Particles $\boldsymbol{x}_k^j$
22: for $j = 1 : M$ do
23:   Keep $\boldsymbol{x}_k^j$ unchanged
24:   Update weights by $u_{\boldsymbol{x}_k^j} \leftarrow u_{\boldsymbol{x}_k^j} \cdot p^*(\boldsymbol{y}_k|\boldsymbol{x}_k^j)$
25: end for
26: Normalize weights $u_{\boldsymbol{x}_k^j}, \forall j \in [M]$
27: // Step 5: Resampling
28: $\hat{N}_{eff} \leftarrow 1/\sum_{j=1}^{M} u_{\boldsymbol{x}_k^j}^2$
29: if $\hat{N}_{eff} < N_{thres}$ then
30:   Resample $N$ times from $\{\boldsymbol{x}_k^j, u_{\boldsymbol{x}_k^j}\}_{j\in[M]}$
31: end if
Output: Worst-case posterior state particles $\{\boldsymbol{x}_k^i\}$ and weights $\{u_{\boldsymbol{x}_k^i}\}$, $\forall i \in [N]$.
```

E. Computational Burden

As we can see, the proposed generic robustified particle filter is computationally intensive: if S in Algorithm 1 and N, M, R, and T in Algorithm 2 are large, the calculation burden is heavy as well. The worst-case complexity order of Algorithm 1 is $\mathcal{O}(S)$. However, the complexity order of Algorithm 2 is hard to be specified because it depends on which sampling method (e.g., the importance sampling and the fundamental theorem of simulation) is used, which resampling method (e.g., systematic and multinomial) is used, and which maximum-entropy method (among Methods 1-5) is used. The burden, however, is unavoidable to robustify particle-based filters and to evaluate likelihoods under non-additive and non-multiplicative measurement noises. If the state equation is exact and measurement noise densities fortunately have closed-form expressions (see, e.g., Method 3), then the computation burden can be limited and the resulted robust particle filter has the same computational complexity as the canonical particle filter (because no extra computation burden is introduced in Step 2 and Step 3).

F. Size of Ambiguity Set

The sizes of ambiguity sets (i.e., θ's in Theorems 3-6) need to be specified in implementing the robustified particle filter. However, this cannot be theoretically conducted because for a real state estimation problem, the true states are unknown. In other words, the training dataset is unavailable so that the sizes of ambiguity sets cannot be tuned to be (nearly) optimal. Therefore, signal processing practitioners are expected to try appropriate values for their specific problems. The general principle is that the sizes can be neither too large nor too small: an extremely large value renders the robust filter being too conservative, while the robust filter with an extremely small value does not have sufficient robustness. For details, see the experiment section.

V. Experiments

All the source codes are available online at GitHub: https://github.com/Spratm-Asleaf/DRSE-Nonlinear. Additional experiments on finding maximum entropy distributions can be found in Appendix G in the online supplementary materials. We first consider a one-dimensional time series example, and then study a target tracking example.

A. A Time Series Example

In this subsection, we consider a univariate non-stationary growth model [18], [31]. The **true** system model is given as

$$\begin{cases} x_k = \dfrac{x_{k-1}}{2} + \dfrac{25x_{k-1}}{1+x_{k-1}^2} + 8\cos(1.2k) + w_{k-1}, \\ y_k = \dfrac{x_k^2}{20} + 0.5\sin(x_k) + v_k. \end{cases}$$

The process noise w_k and the measurement noise v_k follow zero-mean Gaussian distributions; the variance of w_k is $Q_k = 10$ and of v_k is $R_k = 1$; the initial condition is $x_0 = 0$ [18]. This is an abstract time series problem, and the physical meanings of variables and their units are not specified. In this experiment, we

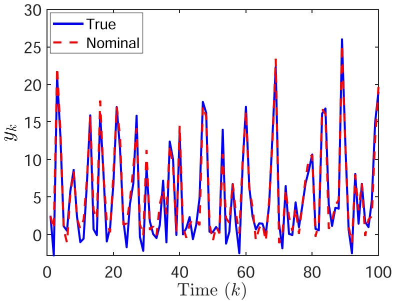


Fig. 2. Given the same system state trajectory (i.e., $\{x_k\}, \forall k$), there do not exist significant differences between measurements from the true model and the nominal model.

suppose the **nominal** system model is uncertain and given as

$$\begin{cases} x_k = \dfrac{x_{k-1}}{2} + \dfrac{25x_{k-1}}{1+x_{k-1}^2} + 8\cos(1.2k) + w_{k-1}, \\ y_k = \dfrac{x_k^2}{20} + v_k. \end{cases}$$

Namely, there is a model mismatch for the measurement equation: the term "$0.5\sin(x_k)$" in the true measurement model was not identified by the model designer. In the measurement equation, the term "$x_k^2/20$" is dominating because it has significantly larger values than those of the term "$0.5\sin(x_k)$". Hence, the model designer can hardly get aware of this model mismatch; cf. Fig. 2. Till now, none of existing filters for nonlinear system models can handle such a model mismatch.

We implement the canonical particle filter (PF) in [18, Algorithm 3], the Gaussian approximation method, specifically the unscented Kalman filter (UKF), in [72, pp. 448–450], and the robust particle filter (RPF) in Algorithm 2 for comparison. In this example, since the measurement noise v_k is additive and Gaussian, Method 3 is used to evaluate worst-case likelihoods for the proposed robust particle filter. For all methods, we assume that the initial state particles are sampled from a one-dimensional Gaussian distribution with mean of 0 and variance of 1. Since there do not exist model uncertainties in the state equation, Algorithm 1 is not used. In Algorithm 2, we do not initialize R and T because for this closed-form likelihood evaluation case, they are not used.

First, we investigate the performance of the PF and the RPF with different numbers of particles. Let the number of particles be N for both the PF and the RPF. For every given N, we conduct 100 independent Monte Carlo episodes and each episode runs 100 time steps. The performance of each method, in each episode, is measured by the root time-averaged mean square error (RTAMSE) along 100 time steps, i.e.,

$$\sqrt{\frac{1}{100}\sum_{k=1}^{100}(x_k - \hat{x}_k)^2}$$

where $\hat{x}_k$ denotes the estimate of x_k. The overall performance of each method is measured by the **averaged RTAMSE** of the 100 episodes, which is shown in Table I. In Method 3 for the RPF, we use $\theta = 5$ for demonstration.

TABLE I
AVERAGED RTAMSE VERSUS NUMBER OF PARTICLES

N	**50**	**100**	**150**	**200**	**250**
PF	5.33	4.98	4.77	4.76	4.74
UKF	9.59	9.59	9.59	9.59	9.59
RPF	**4.99**	**4.71**	**4.66**	**4.62**	**4.64**
N	**300**	**350**	**400**	**450**	**500**
PF	4.69	4.72	4.71	4.69	4.70
UKF	9.59	9.59	9.59	9.59	9.59
RPF	**4.64**	**4.62**	**4.62**	**4.62**	**4.61**

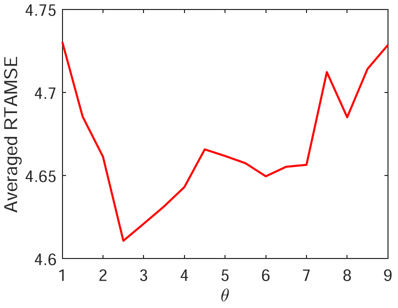


Fig. 3. The averaged RTAMSEs of the RPF versus the values of θ.

As we can see from Table I, as the number of particles increase, the averaged RTAMSEs of both the PF and the RPF decrease. However, the RPF always outperforms the standard PF because the RPF is robust against model uncertainties. In this case, the UKF performs badly because as a Gaussian filter, it cannot handle the high nonlinearity of the system model.

Second, we investigate whether the RPF is robust against the value of θ; cf. Method 3. Since both the PF and the RPF can work satisfactorily with $N = 200$, we use $N = 200$ for every possible θ. All other settings remain unchanged. The averaged RTAMSEs of the RPF versus the values of θ are shown in Fig. 3.

As we can see from Fig. 3, the RPF is sensitive to the value of θ. If we use $\theta = 2.5$, the RPF has smaller averaged RTAMSE than that under $\theta = 5$. Hence, the results of the RPF in Table I can be further refined with $\theta = 2.5$. We used $\theta = 5$ for the RPF in Table I just for the purpose of illustration, which, however, does not lose the generality. To sum up, the value of θ can neither be too large nor be too small. Otherwise, the performance of the RPF is not satisfactory. This is because a large value of θ renders the RPF being too conservative, while a small value cannot provide sufficient robustness for the RPF. This conclusion is consistent with that for the linear system case [39], [40]. Nevertheless, the optimal or convincing tuning method for the value of θ is open. This is because for a real state estimation problem, the true state is unknown, and therefore, there does not exist the training dataset to convincingly determine the value of θ. Thus, in practice, readers should try different values for θ to achieve the satisfactory filtering performance for their specific problems.

Third, we investigate whether the performance of the RPF degrades if the nominal model is exactly the same as the true

TABLE II
AVERAGED RTAMSE VERSUS NUMBER OF PARTICLES (EXACT MODEL)

N	50	100	150	200	250
PF	**4.22**	**3.76**	**3.69**	**3.68**	**3.63**
UKF	9.29	9.29	9.29	9.29	9.29
RPF	4.62	4.44	4.36	4.35	4.32
N	**300**	**350**	**400**	**450**	**500**
PF	**3.63**	**3.63**	**3.64**	**3.62**	**3.61**
UKF	9.29	9.29	9.29	9.29	9.29
RPF	4.33	4.33	4.33	4.33	4.33

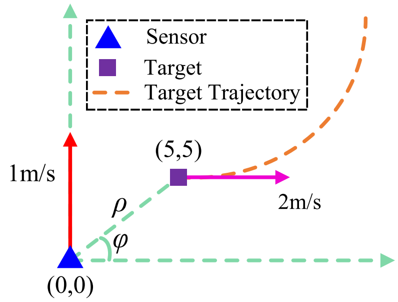


Fig. 4. A target tracking diagram. The initial position of the target is (5, 5) and of the sensor is (0, 0). The trajectory is a quarter of a circle but every point on this circle is contaminated by noise; i.e., the trajectory is not ideally round.

model. In Method 3 for the RPF, we fix $\theta = 5$ for demonstration. The results are shown in Table II.

As we can see from Table II, as the number of particles increase, the averaged RTAMSEs of both the PF and the RPF decrease. However, in this case (i.e., when the nominal model is exactly the same as the true model), the RPF cannot outperform the standard PF. This is because the RF is the optimal method for the exact model in the sense that it gives the best approximates to the posterior state distributions. Hence, there is a trade-off between the optimality under perfect conditions and the robustness under uncertain conditions: the robustness under uncertain conditions comes with the cost of sacrificing the optimality under perfect conditions. To be specific, the PF is optimal when there do not exist model uncertainties but it has no robustness when model uncertainties exist. In contrast, the RPF is robust against model uncertainties but it is not optimal when the nominal model is exactly true. This conclusion is consistent with that for the linear system case; see [39], [40] for details. This is intuitively understandable from a basic life philosophy: nothing is free, although some are cheap.

B. A Target Tracking Example

In this subsection, we consider a target tracking problem under uncertain conditions; see Fig. 4 for an illustration. The target moves along the curved-orange-dotted trajectory and its true (but unknown) speed is $v = 2\,m/s$. The sensor is able to obtain the real-time distance ρ and relative orientation φ from the target to itself; the sensor moves along the vertical axis from the origin and its speed is $v_0 = 1\,m/s$.

We use the (nearly) constant-velocity (CV) model [73] to track the target: the state equation is $\boldsymbol{x}_k = \boldsymbol{F}\boldsymbol{x}_{k-1} + \boldsymbol{G}\boldsymbol{w}_{k-1}$ and

$$\boldsymbol{x}_k := \begin{bmatrix} x_{1,k} \\ s_{1,k} \\ x_{2,k} \\ s_{2,k} \end{bmatrix}, \boldsymbol{F} := \begin{bmatrix} 1 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \Delta t \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{G} := \begin{bmatrix} \Delta t^2/2 & 0 \\ \Delta t & 0 \\ 0 & \Delta t^2/2 \\ 0 & \Delta t \end{bmatrix},$$

where $\Delta t := 0.5\,s$ is the sampling time; x_1 and s_1 (resp. x_2 and s_2) denote the real-time position and velocity of the target in the horizontal (resp. vertical) axis, respectively; white-Gaussian-distributed $\boldsymbol{w}_{k-1}$ is the acceleration noise vector whose mean is zero and covariance is $\boldsymbol{Q}_{k-1} := \text{diag}\{5, 5\}$. When we use the CV model, we assume that the true velocity of the target in each axis is a slowly-changing (i.e., almost constant) time function. Another choice is to use the (nearly) constant-acceleration (CA) model [73], which assumes that the true acceleration of the target in each axis is a slowly-changing time function. Yet another choice is to use the coordinated-turn (CT) model [73], which assumes that the true angular velocity of the target is a slowly-changing time function. For this target tracking example, the CV model, the CA model, the CT model, and also other suitable models [73] are applicable although they may have different tracking performance. However, none of these models are exact models because, for example, when we adopt the CT model, we do not know the exact values of the angular velocity over time. That is, there exist model mismatches no matter which model is used. Since this article is studying the advantage of the proposed distributionally robust particle filter over the standard particle filter when modeling uncertainties exist, rather than investigating which model is best for this specific target tracking example, it is sufficient to use the CV model for demonstration.

On the other hand, the nominal measurement model is $\boldsymbol{y}_k := [\rho_k, \varphi_k]^T$ and

$$\rho_k = \sqrt{(x_{1,k} - x^0_{1,k})^2 + \left(x_{2,k} - x^0_{2,k}\right)^2} + v_{1,k},$$

$$\varphi_k = \tan^{-1}\left(x_{2,k} - x^0_{2,k}, x_{1,k} - x^0_{1,k}\right) + v_{2,k},$$

where $x^0_{1,k}$ and $x^0_{2,k}$ denote the real-time position of the sensor in the horizontal axis and the vertical axis, respectively; v_1 is the ranging error with unit of m and v_2 is the heading error with unit of rad; $\tan^{-1}(\cdot,\cdot)$ is the two-argument inverse tangent function [74]. Both v_1 and v_2 are white Gaussian with zero mean. The measurement noise covariance is $\boldsymbol{R}_k := [\begin{smallmatrix} 1 & 0 \\ 0 & 0.001 \end{smallmatrix}]$. (Namely, the error range of v_1 is $\pm 3\sqrt{1}\,m = \pm 3\,m$ and of v_2 is $\pm 3\sqrt{0.001}\,rad = \pm 0.095\,rad = \pm 5.44\,deg$.) The unit of all position variables is meter (m), the unit of all speed and velocity variables is meter per second (m/s), and the unit of all angle variables is radian (rad).

However, in practice, there may exist positioning errors for the moving sensor; the nominal values of $x^0_{1,k}$ and $x^0_{2,k}$ (might from GPS etc.) are uncertain. Specifically, the true governing (but unknown) measurement model might be

$$\rho_k = \sqrt{\left(x_{1,k} - x^0_{1,k} - \eta_{1,k}\right)^2 + \left(x_{2,k} - x^0_{2,k} - \eta_{2,k}\right)^2} + v_{1,k},$$

$$\varphi_k = \tan^{-1}\left(x_{2,k} - x^0_{2,k} - \eta_{2,k}, x_{1,k} - x^0_{1,k} - \eta_{1,k}\right) + v_{2,k},$$

where $\eta_{1,k}$ and $\eta_{2,k}$ are the sensor's positioning errors. In this experiment, they are assumed to be Gaussian having the same mean of zero and the same variance of $1\,m^2$; i.e., the error range is $\pm 3\sqrt{1}\,m = \pm 3\,m$.

We implement the canonical particle filter (PF) in [18, Algorithm 3], the Gaussian approximation method, specifically the unscented Kalman filter, (GA-UKF) in [72, pp. 448–450], and the robust particle filter (RPF) in Algorithm 2 for comparison. In this example, since the measurement noises $v_{1,k}$ and $v_{2,k}$ are additive and Gaussian, Method 3 with $\theta := 5$ is used to evaluate worst-case likelihoods for the proposed robust particle filter. For both the PF and the RPF, the number of particles are all set to 1000. For all methods, we assume that the initial state particles are sampled from a 4-dimensional Gaussian distribution with mean of $[5, 0, 5, 0]^T$ and covariance of $\operatorname{diag}\{0.1, 0.1, 0.1, 0.1\}$. In Algorithm 1, $S := 500$, $\alpha := 0.05$, $\epsilon := 1 \times 10^{-4}$, $\lambda_0 := 2$, and $\lambda_1 := 0$. In Algorithm 2, $N = M := 1000$. (We do not initialize R and T because for this closed-form likelihood evaluation case, they are not used.) Just for the demonstration purpose and without loss of generality, all the involved parameter values in this experiment are arbitrarily set but the conclusions remain consistent. One may try other values for comparison using the shared source codes at GitHub.

We conduct 50 independent episodes of Monte Carlo simulations and each episode runs 100 discrete time steps.

For each episode, the position estimation error is measured by the root time-averaged mean square error (RTAMSE) along 100 time steps, i.e.,

$$\sqrt{\frac{1}{100}\sum_{k=1}^{100}(x_{1,k} - \hat{x}_{1,k})^2 + (x_{2,k} - \hat{x}_{2,k})^2}$$

where $\hat{x}_{1,k}$ (resp. $\hat{x}_{2,k}$) denotes the estimate of the position $x_{1,k}$ (resp. $x_{2,k}$). Likewise, the velocity estimation error for each episode is given by

$$\sqrt{\frac{1}{100}\sum_{k=1}^{100}(s_{1,k} - \hat{s}_{1,k})^2 + (s_{2,k} - \hat{s}_{2,k})^2}$$

where $\hat{s}_{1,k}$ (resp. $\hat{s}_{2,k}$) denotes the estimate of the velocity $s_{1,k}$ (resp. $s_{2,k}$). The position and velocity RTAMSEs of each episode are shown in Fig. 5.

The overall target tracking error is measured by the **averaged RTAMSE** of the 50 episodes, which is shown in Table III, where "R" stands for averaged RTAMSEs (unit: m for the position case or m/s for the velocity case), while "T" denotes average execution time at each time step (unit: second).

At each time step k, the position estimation error is measured by the averaged root mean square errors (RMSE) of the 50 episodes, i.e.,

$$\frac{1}{50}\sum_{l=1}^{50}\sqrt{\left[x^{(l)}_{1,k} - \hat{x}^{(l)}_{1,k}\right]^2 + \left[x^{(l)}_{2,k} - \hat{x}^{(l)}_{2,k}\right]^2},$$

where $\hat{x}^{(l)}_{1,k}$ (resp. $\hat{x}^{(l)}_{2,k}$) denotes the estimate of the position $x_{1,k}$ (resp. $x_{2,k}$) in the l^{th} episode. Likewise, at time k, the velocity estimation error is measured by the averaged root mean square errors (RMSE) of the 50 episodes, i.e.,

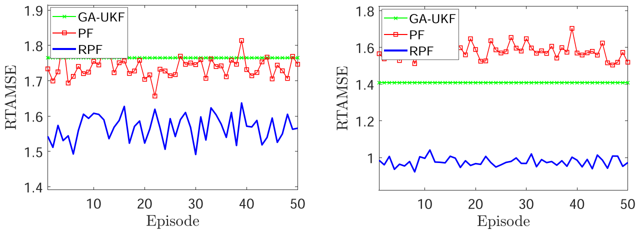
$$\frac{1}{50}\sum_{l=1}^{50}\sqrt{\left[s^{(l)}_{1,k} - \hat{s}^{(l)}_{1,k}\right]^2 + \left[s^{(l)}_{2,k} - \hat{s}^{(l)}_{2,k}\right]^2},$$

where $\hat{s}^{(l)}_{1,k}$ (resp. $\hat{s}^{(l)}_{2,k}$) denotes the estimate of the velocity $s_{1,k}$ (resp. $s_{2,k}$) in the l^{th} episode. The averaged position and velocity RMSEs at each time step k are shown in Fig. 6.

As we can see from Fig. 5 and Table III, when there exist modeling uncertainties (i.e., when the non-exact CV model is used and when both $\eta_{1,k}$ and $\eta_{2,k}$ are non-zero),

1) the GA-UKF and the PF have the roughly same position estimation error, while the RPF has a significantly smaller position estimation error;
2) the velocity estimation error of the GA-UKF is significantly smaller than that of the PF, while the velocity estimation error of the RPF is significantly smaller than those of both the PF and the GA-UKF.

This is because the GA-UKF and the RPF are distributionally robust against model uncertainties: when model uncertainties exist, the GA-UKF and the RPF have the ability to withstand them. In addition, the RPF outperforms the GA-UKF since



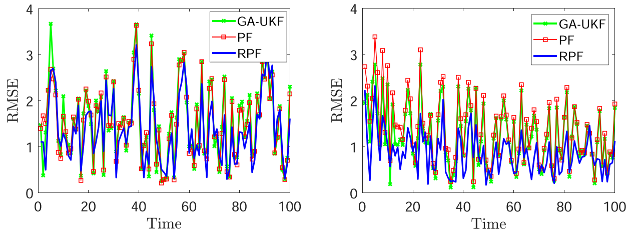
(a) Position RTAMSE (b) Velocity RTAMSE

Fig. 5. Position and velocity RTAMSEs of each episode.

TABLE III
TARGET TRACKING RESULTS

	PF		GA-UKF		RPF	
	R	T	R	T	R	T
Position	1.74	0.0080	1.76	0.000025	**1.57**	0.0085
Velocity	1.58		1.41		**0.98**	

R: averaged RTAMSEs (unit: m for position; m/s for velocity).
T: Time (unit: second).



(a) Position RMSE (b) Velocity RMSE

Fig. 6. Averaged position and velocity RMSEs at each time step k. In particular, the velocity RMSE of the RPF is significantly smaller than those of the PF and the GA-UKF.

the GA-UKF restrictively assumes Gaussianity of prior state distributions and likelihood distributions. However, the benefit of using GA-UKF is that it is computationally efficient.[6]

In the experiment, just for the illustration purpose, we use $\theta = 5$ in the proposed RPF. However, the RPF is sensitive to the value of θ: this value can neither be too large nor be too small. Otherwise, the performance of the RPF is not satisfactory. A large value of θ renders the RPF being too conservative, while a small value cannot provide sufficient robustness for the RPF. This phenomenon has been reported in Fig. 3 and deeply studied for the linear system case [39], [40]. Hence, we do not discuss much here. One may also use the shared source codes to verify this claim. Nevertheless, the tuning method for the value of θ is open. This is because for a real-world state estimation problem, the true state is unknown, and therefore, there does not exist the training dataset to convincingly determine the value of θ. Thus, in practice, readers should try different values for θ to achieve the satisfactory filtering performance for their specific problems.

VI. Conclusion

This article studies the distributionally robust state estimation scheme for nonlinear systems subject to model uncertainties. Attention has been paid to the particle filtering framework. The maximum entropy prior state distributions and the maximum entropy likelihood distributions are leveraged to robustify the particle filter. The proposed maximum-entropy strategies can also provide weight-balancing mechanism to reduce particle degeneracy and new-sample-generating mechanism to diminish particle impoverishment. The existing Gaussian approximation framework is proven to be distributionally robust but it may have limited ability to handle high nonlinearity of nonlinear system models. In addition, a generic likelihood evaluation method is presented under non-additive and non-multiplicative measurement noises. However, extra computation burden is required to obtain worst-case prior state particles even when worst-case likelihoods can be analytically evaluated. Another issue is to properly choose the radii of ambiguity sets, i.e., θ's in Theorems 3-6. Nevertheless, these radii cannot be trained to be (nearly) optimal because for real state estimation problems, true states (i.e., training dataset) are unknown. Therefore, in practice, practitioners have to try appropriate values for their specific problems.

References

[1] N. Wahlström and E. özkan, "Extended target tracking using Gaussian processes," *IEEE Trans. Signal Process.*, vol. 63, no. 16, pp. 4165–4178, Aug. 2015.

[2] J. Zhao et al., "Power system dynamic state estimation: Motivations, definitions, methodologies, and future work," *IEEE Trans. Power Syst.*, vol. 34, no. 4, pp. 3188–3198, Jul. 2019.

[3] W. Peng, Z.-S. Ye, and N. Chen, "Joint online RUL prediction for multivariate deteriorating systems," *IEEE Trans. Ind. Inform.*, vol. 15, no. 5, pp. 2870–2878, May 2019.

[4] Y. Tian, M. Ge, and F. Neitzel, "Particle filter-based estimation of inter-frequency phase bias for real-time glonass integer ambiguity resolution," *J. Geodesy*, vol. 89, no. 11, pp. 1145–1158, 2015.

[5] W. Song, Z. Wang, J. Wang, F. E. Alsaadi, and J. Shan, "Particle filtering for nonlinear/non-Gaussian systems with energy harvesting sensors subject to randomly occurring sensor saturations," *IEEE Trans. Signal Process.*, vol. 69, pp. 15–27, 2020.

[6] C. Papachristos et al., *Modeling, Control, State Estimation and Path Planning Methods for Autonomous Multirotor Aerial Robots*. Norwell, MA, USA: Now Publishers, 2018.

[7] L. Zhang, S. Qian, S. Zhang, and H. Cai, "Federated nonlinear predictive filtering for the gyroless attitude determination system," *Adv. Space Res.*, vol. 58, no. 9, pp. 1671–1681, 2016.

[8] K. Fujii, "Extended Kalman filter," *Refernce Manual*, pp. 14–22, 2013. [Online]. Available: https://www-jlc.kek.jp/2004sep/subg/offl/kaltest/doc/ReferenceManual.pdf

[9] N. Pletschen and K. J. Diepold, "Nonlinear state estimation for suspension control applications: A Takagi-Sugeno Kalman filtering approach," *Control Eng. Pract.*, vol. 61, pp. 292–306, 2017.

[10] H. H. Afshari, S. A. Gadsden, and S. Habibi, "Gaussian filters for parameter and state estimation: A general review of theory and recent trends," *Signal Process.*, vol. 135, pp. 218–238, 2017.

[11] E. A. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Proc. IEEE Adaptive Syst. Signal Process. Commun. Control Symp.*, 2000, pp. 153–158.

[12] I. Arasaratnam and S. Haykin, "Cubature Kalman filters," *IEEE Trans. Autom. Control*, vol. 54, no. 6, pp. 1254–1269, Jun. 2009.

[13] M. Katzfuss, J. R. Stroud, and C. K. Wikle, "Understanding the ensemble Kalman filter," *Amer. Statistician*, vol. 70, no. 4, pp. 350–357, 2016.

[14] H. Wang, H. Li, J. Fang, and H. Wang, "Robust Gaussian Kalman filter with outlier detection," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1236–1240, Aug. 2018.

[15] H. Li, D. Medina, J. Vilà-Valls, and P. Closas, "Robust variational-based Kalman filter for outlier rejection with correlated measurements," *IEEE Trans. Signal Process.*, vol. 69, pp. 357–369, 2020.

[16] K. Li, S. Zhao, and F. Liu, "Joint state estimation for nonlinear state-space model with unknown time-variant noise statistics," *Int. J. Adaptive Control Signal Process.*, vol. 35, no. 4, pp. 498–512, 2021.

[17] J. Courts, A. Wills, and T. B. Schon, "Gaussian variational state estimation for nonlinear state-space models," *IEEE Trans. Signal Process.*, vol. 69, pp. 5979–5993, 2021.

[18] 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.

[19] A. Doucet et al., "A tutorial on particle filtering and smoothing: Fifteen years later," *Handbook Nonlinear Filtering*, vol. 12, no. 656-704, 2009, Art. no. 3.

[20] F. Gustafsson, "Particle filter theory and practice with positioning applications," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 25, no. 7, pp. 53–82, Jul. 2010.

[21] J. Elfring, E. Torta, and R. van de Molengraft, "Particle filters: A hands-on tutorial," *Sensors*, vol. 21, no. 2, 2021, Art. no. 438.

[22] D. Crisan and A. Doucet, "A survey of convergence results on particle filtering methods for practitioners," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 736–746, Mar. 2002.

[23] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 131–146, Nov. 2008.

[24] T. Li, M. Bolic, and P. M. Djuric, "Resampling methods for particle filtering: Classification, implementation, and strategies," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 70–86, May 2015.

[25] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djuric, "Adaptive importance sampling: The past, the present, and the future," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 60–79, Jul. 2017.

[26] Y. El-Laham, V. Elvira, and M. F. Bugallo, "Robust covariance adaptation in adaptive importance sampling," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 1049–1053, Jul. 2018.

[27] V. Stojanovic and N. Nedic, "Robust Kalman filtering for nonlinear multivariable stochastic systems in the presence of non-gaussian noise," *Int. J. Robust Nonlinear Control*, vol. 26, no. 3, pp. 445–460, 2016.

[28] X. Kai, C. Wei, and L. Liu, "Robust extended Kalman filtering for nonlinear systems with stochastic uncertainties," *IEEE Trans. Syst., Man, Cybern.-Part A: Syst. Hum.*, vol. 40, no. 2, pp. 399–405, Mar. 2010.

[6] Note that in the time-series example, the UKF is unsatisfactory when there exists measurement model uncertainty, although it is distributionally robust. This is because the system model there is highly nonlinear but the UKF, as a Gaussian filter, cannot handle high nonlinearity in nonlinear system models.

[29] R. B. Abdallah, G. Pagès, D. Vivet, J. Vilà-Valls, and E. Chaumette, "Robust linearly constrained square-root cubature Kalman filter for mismatched nonlinear dynamic systems," *IEEE Control Syst. Lett.*, vol. 6, pp. 2335–2340, 2022.

[30] Y. Wang, Z. Yang, Y. Wang, V. Dinavahi, J. Liang, and K. Wang, "Robust dynamic state estimation for power system based on adaptive cubature Kalman filter with generalized correntropy loss," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[31] L. Chang and K. Li, "Unified form for the robust Gaussian information filtering based on M-estimate," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 412–416, Apr. 2017.

[32] O. Cappé and E. Moulines, "On-line expectation–maximization algorithm for latent data models," *J. Roy. Stat. Society: Ser. B. Stat. Methodol.*, vol. 71, no. 3, pp. 593–613, 2009.

[33] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *J. Roy. Stat. Soc.: Ser. B Stat. Methodol.*, vol. 72, no. 3, pp. 269–342, 2010.

[34] S. Gillijns and B. De Moor, "Unbiased minimum-variance input and state estimation for linear discrete-time systems," *Automatica*, vol. 43, no. 1, pp. 111–116, 2007.

[35] K. Myers and B. Tapley, "Adaptive sequential estimation with unknown noise statistics," *IEEE Trans. Autom. Control*, vol. AC-21, no. 4, pp. 520–523, Aug. 1976.

[36] I. Urteaga, M. F. Bugallo, and P. M. Djurić, "Sequential Monte Carlo methods under model uncertainty," in *Proc. IEEE Stat. Signal Process. Workshop*, 2016, pp. 1–5.

[37] P. J. Huber, *Robust Statistics*, 2nd. ed. Hoboken, NJ, USA: Wiley, 2009.

[38] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *Proc. Operations Res. Manage. Sci. Age Analytics*, 2019, pp. 130–166.

[39] S. Wang, Z. Wu, and L. Andrew, "Robust state estimation for linear systems under distributional uncertainty," *IEEE Trans. Signal Process.*, vol. 69, pp. 5963–5978, 2021.

[40] S. Wang and Z. Ye, "Distributionally robust state estimation for linear systems subject to uncertainty and outlier," *IEEE Trans. Signal Process.*, vol. 70, pp. 452–467, 2021.

[41] M. Fauß, A. M. Zoubir, and H. V. Poor, "Minimax robust detection: Classic results and recent advances," *IEEE Trans. Signal Process.*, vol. 69, pp. 2252–2283, 2021.

[42] A. H. Sayed, "A framework for state-space estimation with uncertain models," *IEEE Trans. Autom. Control*, vol. 46, no. 7, pp. 998–1013, Jul. 2001.

[43] F. Wang and V. Balakrishnan, "Robust Kalman filters for linear time-varying systems with stochastic parametric uncertainties," *IEEE Trans. Signal Process.*, vol. 50, no. 4, pp. 803–813, Apr. 2002.

[44] Y. Zhang, Z. Zhang, A. Lim, and M. Sim, "Robust data-driven vehicle routing with time windows," *Operations Res.*, vol. 69, no. 2, pp. 469–485, 2021.

[45] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, "Regularization via mass transportation," *J. Mach. Learn. Res.*, vol. 20, no. 103, pp. 1–68, 2019.

[46] I. Yang, "A dynamic game approach to distributionally robust safety specifications for stochastic systems," *Automatica*, vol. 94, pp. 94–101, 2018.

[47] R. Chen and I. C. Paschalidis, "Distributionally robust learning," *Found. Trends Optim.*, vol. 4, no. 1-2, pp. 1–243, 2020.

[48] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[49] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1991.

[50] J. O. Berger et al., "An overview of robust Bayesian analysis," *Test*, vol. 3, no. 1, pp. 5–124, 1994.

[51] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd. ed. Berlin, Germany: Springer, 1985.

[52] C. P. Robert et al., *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, 2nd. ed, vol. 2. Berlin, Germany: Springer, 2007.

[53] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inf. Theory*, vol. IT-26, no. 1, pp. 26–37, Jan. 1980.

[54] P. D. Grünwald and A. P. Dawid, "Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory," *Ann. Statist.*, vol. 32, no. 4, pp. 1367–1433, 2004.

[55] I. R. Petersen and A. V. Savkin, *Robust Kalman Filtering for Signals and Systems With Large Uncertainties*. Berlin, Germany: Springer, 1999.

[56] K. Zhou, J. Doyle, and K. Glover, *Robust and Optimal Control*. Englewood Cliffs, NJ, USA: Prentice Hall, 1996.

[57] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations Res.*, vol. 58, no. 3, pp. 595–612, 2010.

[58] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Operations Res.*, vol. 62, no. 6, pp. 1358–1376, 2014.

[59] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen, "Robust solutions of optimization problems affected by uncertain probabilities," *Manage. Sci.*, vol. 59, no. 2, pp. 341–357, 2013.

[60] K. P. Murphy, *Machine Learning: A. Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[61] L. V. Kantorovich and S. Rubinshtein, "On a space of totally additive functions," *Vestnik St Petersburg Univ. Math.*, vol. 13, no. 7, pp. 52–59, 1958.

[62] C. Villani, *Topics in Optimal Transportation*, vol. 58. Providence, RI, USA: American Mathematical Society, 2003.

[63] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations," *Math. Program.*, vol. 171, no. 1, pp. 115–166, 2018.

[64] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[65] J. G. Carlsson and R. Devulapalli, "Dividing a territory among several facilities," *INFORMS J. Comput.*, vol. 25, no. 4, pp. 730–742, 2013.

[66] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[67] L. F. Shampine, "Vectorized adaptive quadrature in MATLAB," *J. Comput. Appl. Math.*, vol. 211, no. 2, pp. 131–140, 2008.

[68] R. E. Caflisch, "Monte Carlo and quasi-Monte Carlo methods," *Acta Numerica*, vol. 7, pp. 1–49, 1998.

[69] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*. Philadelphia, PA, USA: SIAM, 1992.

[70] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[71] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.

[72] D. Simon, *Optimal State Estimation: Kalman, $H\infty$, and Nonlinear Approaches*. Hoboken, NJ, USA: Wiley, 2006.

[73] 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.

[74] M. Mallick, "A note on bearing measurement model," 2018. [Online]. Available: https://www.researchgate.net/publication/325214760_A_Note_on_Bearing_Measurement_Model

Shixiong Wang (Member, IEEE) received the B.Eng. degree in detection, guidance and control technology, the M.Eng. degree in systems and control engineering from the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China, in 2016 and 2018, respectively and the Ph.D. degree from the Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore, in 2022. He is currently a Postdoctoral Research Fellow with the Institute of Data Science, National University of Singapore. His research interest includes statistics and optimization theories with applications in signal processing (especially optimal estimation theory) and control technology.

Supplementary Materials

Appendix A
On Model Uncertainties

In the signal processing [55, Chapter 1], [41] and automatic control [56, Chapter 9] communities (and also many other fields), a nominal model $\boldsymbol{O} = \mathcal{M}(\boldsymbol{I})$ is said to be uncertain if it is not guaranteed to be exactly the same as the true governing model $\boldsymbol{O} = \mathcal{M}_0(\boldsymbol{I})$, where $\boldsymbol{O}$ denotes the output and $\boldsymbol{I}$ the input. Other equivalent terms to "uncertain model" that are widely used include "mismatched model", "deviated model", and "perturbed model", etc. Possible cases are as follows.

1) **Parameter Uncertainty**. Suppose the nominal model $\boldsymbol{O} = \mathcal{M}(\boldsymbol{I}; \boldsymbol{\beta})$ is parameterized by $\boldsymbol{\beta}$. If the model type is exact and only the parameter $\boldsymbol{\beta}$ is uncertain, the model uncertainty is reflected by "parameter uncertainty". In state estimation contexts, a possible example is that the true system model is guaranteed to be linear and the noises are guaranteed to be Gaussian, but we do not exactly know the system matrices and/or noise statistics.
2) **Type Uncertainty**. In state estimation contexts, an example might be the case that the true system model is nonlinear but we might use a linear nominal model. Another example might be the case that the true system model is known to be the one among candidate models. However, at one time instant, we do not exactly know which candidate model is governing the true plant [15], [36]. In this case, one may also call it "mode uncertainty".
3) **Measurement Outlier**. If outliers unexpectedly exist in measurements, the nominal measurement distribution might deviate from the true measurement distribution. In linear-system state estimation contexts, a possible example is that the nominal measurement noise model is Gaussian, whereas the true measurement noise model is fat-tailed (e.g., Laplacian, Student's t).

The list is not exhaustive, however, most common in practice.

Appendix B
Proof of Lemma 1

This lemma is a special case of [62, Theorem 1.3]. With the facts in [62, Remark 1.12], the statements in this lemma can be obtained. However, the proof of [62, Theorem 1.3] is rather complicated because it dealt with a more general problem and conducted many advanced analyses; it is not motivational for the contexts of this article. Below gives a new and concise proof because it is necessary for insights in Fig. 1.

First, by noting that $p(\boldsymbol{x}_{\mathbb{Q}}) = q(\boldsymbol{x}) = \sum_{i=1}^{N} q_i \delta_{\boldsymbol{x}^i}(\boldsymbol{x})$ and $\int q_i \delta_{\boldsymbol{x}^i}(\boldsymbol{x}) d\boldsymbol{x} = q_i$, we have

$$\begin{array}{rl} & \inf_{\pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}) d\boldsymbol{x}_{\mathbb{P}} d\boldsymbol{x}_{\mathbb{Q}} \\ = & \inf_{I(\boldsymbol{x}_{\mathbb{Q}}|\boldsymbol{x}_{\mathbb{P}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \frac{I(\boldsymbol{x}_{\mathbb{Q}}|\boldsymbol{x}_{\mathbb{P}}) p(\boldsymbol{x}_{\mathbb{P}})}{p(\boldsymbol{x}_{\mathbb{Q}})} p(\boldsymbol{x}_{\mathbb{Q}}) d\boldsymbol{x}_{\mathbb{P}} d\boldsymbol{x}_{\mathbb{Q}} \\ = & \inf_{I(\boldsymbol{x}^i|\boldsymbol{x}_{\mathbb{P}})} \sum_{i=1}^{N} \int \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}^i\| \frac{I(\boldsymbol{x}^i|\boldsymbol{x}_{\mathbb{P}}) p(\boldsymbol{x}_{\mathbb{P}})}{p(\boldsymbol{x}_{\mathbb{Q}})|_{\boldsymbol{x}_{\mathbb{Q}} = \boldsymbol{x}^i}} q_i d\boldsymbol{x}_{\mathbb{P}} \\ = & \inf_{I(\boldsymbol{x}^i|\boldsymbol{x}_{\mathbb{P}})} \sum_{i=1}^{N} \int \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}^i\| I(\boldsymbol{x}^i|\boldsymbol{x}_{\mathbb{P}}) p(\boldsymbol{x}_{\mathbb{P}}) d\boldsymbol{x}_{\mathbb{P}} \\ = & \inf_{I(\boldsymbol{x}^i|\boldsymbol{x})} \sum_{i=1}^{N} \int \|\boldsymbol{x} - \boldsymbol{x}^i\| I(\boldsymbol{x}^i|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}. \end{array}$$

The first equality holds because when reformulating the Wasserstein distance, the marginals $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{Q}_{\boldsymbol{x}}$ are fixed. The infimum optimization problem above has a clear physical meaning in transport theory: we aim to move all the resources (that are continuously distributed) in the whole region to some fixed facilities $\{\boldsymbol{x}^i\}_{i=1,2,\ldots,N}$. At every point $\boldsymbol{x}$, the normalized amount of resources are $p(\boldsymbol{x})$. The proportion of $p(\boldsymbol{x})$ to be moved from $\boldsymbol{x}$ to the facility $\boldsymbol{x}^i$ is $I(\boldsymbol{x}^i|\boldsymbol{x})$. The cost to move every unit of resources from $\boldsymbol{x}$ to $\boldsymbol{x}^i$ is $\|\boldsymbol{x} - \boldsymbol{x}^i\|$. Therefore, the Wasserstein distance denotes the minimum transport cost to move a distribution from one support set to another. Since $I(\boldsymbol{x}^i|\boldsymbol{x})$ are conditional distributions, implicit constraints are

$$\left\{ \begin{array}{rll} \int I(\boldsymbol{x}^i|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} & = q_i, & \forall i \in [N], \\ \sum_{i=1}^{N} I(\boldsymbol{x}^i|\boldsymbol{x}) & = 1, & \forall \boldsymbol{x}, \\ I(\boldsymbol{x}^i|\boldsymbol{x}) & \geq 0, & \forall i \in [N], \forall \boldsymbol{x}. \end{array} \right.$$

Second, we write the Lagrange dual problem

$$\begin{array}{cl} \sup_{\lambda_i} \inf_{I(\boldsymbol{x}^i|\boldsymbol{x})} & \sum_{i=1}^{N} \int \|\boldsymbol{x} - \boldsymbol{x}^i\| I(\boldsymbol{x}^i|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} + \\ & \sum_{i=1}^{N} \lambda_i \left[q_i - \int p(\boldsymbol{x}) I(\boldsymbol{x}^i|\boldsymbol{x}) d\boldsymbol{x} \right] \\ s.t. & \sum_{i=1}^{N} I(\boldsymbol{x}^i|\boldsymbol{x}) = 1, \qquad \forall \boldsymbol{x}, \\ & I(\boldsymbol{x}^i|\boldsymbol{x}) \geq 0, \qquad \forall i \in [N], \forall \boldsymbol{x}. \end{array}$$

The sup-inf objective function also writes

$$\begin{array}{cl} \sup_{\lambda_i} \inf_{I(\boldsymbol{x}^i|\boldsymbol{x})} & \int \sum_{i=1}^{N} (\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i) I(\boldsymbol{x}^i|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} + \\ & \sum_{i=1}^{N} \lambda_i q_i. \end{array}$$

Now we recall the physical meaning of $I(\boldsymbol{x}^i|\boldsymbol{x})$ from perspective of optimal transport: it denotes the proportion of $p(\boldsymbol{x})$ to be moved to $\boldsymbol{x}^i$; i.e., $I(\boldsymbol{x}^i|\boldsymbol{x})$ are weights. As a result, we have

$$\min_i \{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\} \leq \sum_{i=1}^{N} (\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i) I(\boldsymbol{x}^i|\boldsymbol{x}), \quad \forall \boldsymbol{x},$$

where $I(\boldsymbol{x}^i|\boldsymbol{x}) = 1$ for the i letting the equality strictly hold, and $I(\boldsymbol{x}^i|\boldsymbol{x}) = 0$ otherwise. The above inequality holds because the weighted mean of a vector is no less than the minimum element in this vector. This gives the dual problem

$$\sup_{\lambda_i} \int \min_{i \in [N]} \{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\} p(\boldsymbol{x}) d\boldsymbol{x} + \sum_{i=1}^{N} \lambda_i q_i.$$

Note that the strong duality holds because the primal optimization problem is convex, and the relative interior point $p(\boldsymbol{x}_{\mathbb{Q}})$ satisfies the Slater's condition: when $p(\boldsymbol{x}_{\mathbb{P}}) := p(\boldsymbol{x}_{\mathbb{Q}})$, the optimal solution $I(\boldsymbol{x}^i|\boldsymbol{x}^i) = 1$ and $I(\boldsymbol{x}^i|\boldsymbol{x}^j) = 0, \forall j \neq i$. Since the value of $I(\boldsymbol{x}^i|\boldsymbol{x})$ is either one or zero, all $p(\boldsymbol{x})$ near $\boldsymbol{x}^i$ are moved to $\boldsymbol{x}^i$, and the cumulative at $\boldsymbol{x}^i$ is q_i (n.b., $\int I(\boldsymbol{x}^i|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = q_i$). This implies a region-partition operation: the sub-region C_i is defined by such a set of $\boldsymbol{x}$ that satisfies $\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i \leq \|\boldsymbol{x} - \boldsymbol{x}^j\| - \lambda_j, \ \forall j \neq i$. In other words, $\int_{C_i} p(\boldsymbol{x}) d\boldsymbol{x} = q_i, \ \forall i \in [N]$. □

APPENDIX C
PROOF OF THEOREM 3

We first consider the case when $\theta > 0$. Let $g(\boldsymbol{x}, \boldsymbol{\lambda}) := \min_{i\in[N]} \left\{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\right\}$. The Lagrange dual problem is

$$\min_{v_0\geq 0, v_1} \max_{p(\boldsymbol{x})} \int -p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} + v_0 \cdot$$
$$\left\{\theta - \max_{\boldsymbol{\lambda}} \left[\int p(\boldsymbol{x}) \min_{i\in[N]} \left\{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\right\} d\boldsymbol{x} + \sum_{i=1}^{N} q_i\lambda_i\right]\right]\right\}$$
$$+ v_1 \left[1 - \int p(\boldsymbol{x}) d\boldsymbol{x}\right]$$
$$= \min_{v_0\geq 0, v_1} \max_{p(\boldsymbol{x})} \min_{\boldsymbol{\lambda}} \quad v_0 \cdot \left(\theta - \sum_{i=1}^{N} q_i\lambda_i\right) + v_1 +$$
$$\int -[\ln p(\boldsymbol{x}) + v_0 g(\boldsymbol{x}, \boldsymbol{\lambda}) + v_1] p(\boldsymbol{x}) d\boldsymbol{x}.$$

For every two bounded functions f_1 and f_2 that have the same support, $\min(f_1 + f_2) \geq \min f_1 + \min f_2$. Therefore, it is easy to verify that the objective function is convex in terms of $\boldsymbol{\lambda}$ and concave in terms of $p(\boldsymbol{x})$ by the original definitions of convexity and concavity. Since the objective function is concave and constraint-free in terms of $p(\boldsymbol{x})$, we use the varitional method to maximize it over $p(\boldsymbol{x})$. Let $\mathcal{L}[p(\boldsymbol{x})] := \int -[\ln p(\boldsymbol{x}) + v_0 g(\boldsymbol{x}, \boldsymbol{\lambda}) + v_1] p(\boldsymbol{x}) d\boldsymbol{x}$ be a functional of $p(\boldsymbol{x})$. The variation of $\mathcal{L}[p(\boldsymbol{x})]$ is

$$\delta\mathcal{L}[p(\boldsymbol{x})] = \left.\frac{\partial\mathcal{L}[p(\boldsymbol{x}) + \epsilon h(\boldsymbol{x})]}{\partial\epsilon}\right|_{\epsilon=0}$$
$$= \int - [\ln p(\boldsymbol{x}) + 1 + v_0 g(\boldsymbol{x}, \boldsymbol{\lambda}) + v_1] h(\boldsymbol{x}) d\boldsymbol{x},$$

where $h(\boldsymbol{x}) \in L^1$ is an arbitrary function.

Let $\delta\mathcal{L}[p(\boldsymbol{x})] = 0$ and according to the fundamental lemma of calculus of variations, we have

$$[\ln p(\boldsymbol{x}) + 1 + v_0 g(\boldsymbol{x}, \boldsymbol{\lambda}) + v_1] \equiv 0,$$

almost everywhere. This gives the form of $p(\boldsymbol{x})$ in (16). Substituting $p(\boldsymbol{x})$ back into the objective of the Lagrange dual problem gives (17). The strong duality holds because (15) is concave and $\mathbb{Q}_{\boldsymbol{x}}$ is a relative interior point at which the inequality constraint in (15) is strictly satisfied (due to $\theta > 0$) and the equality constraint in (15) simultaneously holds (i.e., the Slater's conditions are met).

When $\theta = 0$, the gradient in (19) vanishes if and only if $\mathbb{P}_{\boldsymbol{x}} = \mathbb{Q}_{\boldsymbol{x}}$. Therefore, (16) and (17) also work for $\theta = 0$. In summary, this theorem works for all $\theta \geq 0$. □

APPENDIX D
PROOF OF LEMMA 2

This lemma is a special case of [62, Theorem 1.3]. One can also prove it using the standard Lagrange dual theory (cf. Appendix B). We do not give details due to necessity. □

APPENDIX E
PROOF OF THEOREM 4

The proof is straightforward by writing the Lagrange dual problem and differentiating with respect to P_{ij}. The strong duality holds: (25) is concave and $\{P^0_{ij}\}_{\forall i, \forall j}$ is assumed to be a relative interior point satisfying the Slater's conditions. In the special case when $M = N$, and $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{Q}_{\boldsymbol{x}}$ have the same support, P^0_{ij} can be constructed as follow:

$$P^0_{ij} = \left\{ \begin{array}{ll} q_i, & \text{if } i = j, \\ 0, & \text{otherwise}, \end{array} \right.$$

which is resulted from letting $\mathbb{P}_{\boldsymbol{x}} := \mathbb{Q}_{\boldsymbol{x}}$. In a general case when $M \neq N$ or they have different supports, to guarantee the existence of P^0_{ij}, we must let θ be strictly larger than $\min_{P_{ij}} \sum_{i=1}^{N}\sum_{j=1}^{M} \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \cdot P_{ij}$ over all P_{ij} such that $\sum_{j=1}^{M} P_{ij} = q_i, \ \forall i \in [N]$. Note that unlike Theorem 3, we additionally require the existence of P^0_{ij}. This is because the reference distribution $\mathbb{Q}_{\boldsymbol{x}}$ in this case is no longer guaranteed to be a relative interior point that satisfies the Slater's conditions. □

APPENDIX F
PROOF OF THEOREM 6

If $\theta = 0$, the maximum entropy distribution solving (34) is $\boldsymbol{q}$ itself. Below discusses the case when $\theta > 0$. The Lagrange dual problem of (34) is

$$\begin{array}{rl} \displaystyle\min_{\lambda_0\geq 0, \lambda_1} \max_{p_i} & \sum_{i=1}^{N} -p_i \ln p_i \\ & +\lambda_0 \cdot \left[\theta - \sum_{i=1}^{N} p_i \ln\left(\frac{p_i}{q_i}\right)\right] \\ & +\lambda_1 \cdot \left[1 - \sum_{i=1}^{N} p_i\right]. \end{array}$$

It is concave, smooth, and constraint-free with respect to p_i. Therefore, the optimal solution of p_i is obtained by the first-order optimality condition, i.e.,

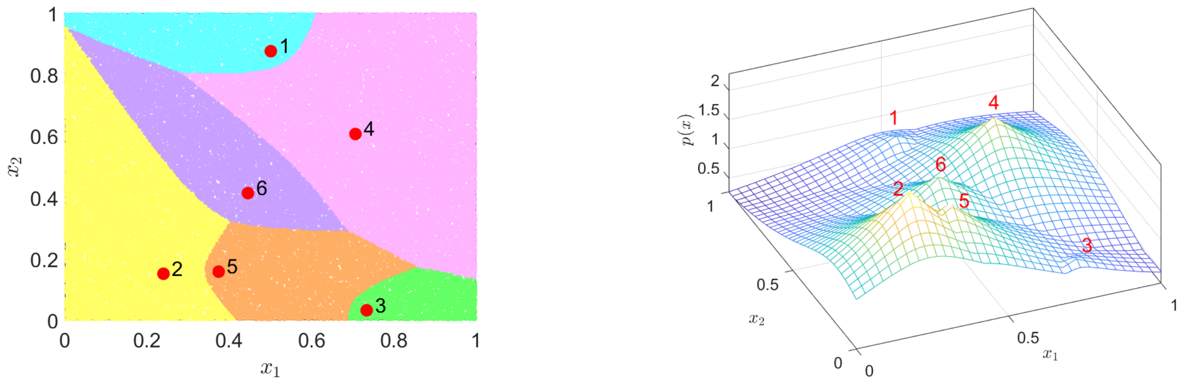
$$-(\lambda_0 + 1) \cdot [\ln(p_i) + 1] + \lambda_0 \ln(q_i) - \lambda_1 = 0.$$

This gives (35). Substituting (35) back into the objective of the Lagrange dual problem, we have (36). Since (34) is concave, and $\boldsymbol{q}$ is a relative interior point in the feasible region of (34) such that the inequality is strictly satisfied (due to $\theta > 0$) and the equality is met, the strong duality holds due to the Slater's condition. Namely, if λ_0 and λ_1 solve (36), p_i in (35) solves (34). When $\theta = 0$, the gradient (37) vanishes if and only if $\boldsymbol{p} = \boldsymbol{q}$; i.e., (35) and (36) also work for the case when $\theta = 0$. In summary, this theorem works for all $\theta \geq 0$. □

APPENDIX G
MAXIMUM ENTROPY DISTRIBUTIONS

A. Continuous Maximum Entropy Distribution Using Wasserstein Distance

We consider a two-dimensional continuous rectangular region $[0, 1] \times [0, 1]$. Let $\boldsymbol{x}$ be a 2-dimensional prior state particle: x_1 denote the horizontal axis and x_2 the vertical axis. Suppose the reference discrete prior state distribution $\boldsymbol{q}$ is supported on six points, which are randomly sampled from the rectangle.



(a) Optimal Partition.

(b) Maximum Entropy Distribution.

Fig. 7. Optimal partition and maximum entropy distribution. The whole rectangular region is partitioned into six sub-regions. Red-filled circles in (a) indicate the supports of the reference distribution $\boldsymbol{q}$. Peaks in (b) correspond to the the supporting points of $\boldsymbol{q}$.

TABLE IV
THE REFERENCE DISTRIBUTION

	$\boldsymbol{x}^1$	$\boldsymbol{x}^2$	$\boldsymbol{x}^3$	$\boldsymbol{x}^4$	$\boldsymbol{x}^5$	$\boldsymbol{x}^6$
Points	0.5007	0.2397	0.7338	0.7065	0.3739	0.4450
	0.8763	0.1513	0.0323	0.6066	0.1581	0.4139
Weights	0.0583	0.2695	0.0340	0.3496	0.1453	0.1433

Their weights are also randomly determined. The points and their weights are displayed in Table IV.

We use Theorem 3 and its corresponding projected gradient descent method to find the continuous maximum entropy distribution. The uncertainty budget θ is set to $\theta := 0.025$ (only for a possible demonstration; other values also applicable). In the projected gradient descent procedure, the step size $\alpha := 0.05$ and the maximum allowed iteration steps $S := 500$. The results are shown in Fig. 7. The Monte Carlo integration method is used to evaluate integrals in (19), (20), and (21); for every Monte Carlo sample $\boldsymbol{x}$, it belongs to C_i if

$$\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i \leq \|\boldsymbol{x} - \boldsymbol{x}^j\| - \lambda_j, \ \forall j \neq i.$$

B. Discrete Maximum Entropy Distribution Using Kullback-Leibler Divergence

The reference distribution $\boldsymbol{q}$ and the induced maximum entropy distribution $\boldsymbol{p}$ are displayed in Table V and Fig. 8. $\boldsymbol{p}$ is calculated by Theorem 6. Since they have the same support set, we do not explicitly demonstrate what the particles $\boldsymbol{x}^i$ are. The uncertainty budget θ is set to $\theta := 0.0075$ (only for a possible demonstration; other values also applicable). In the projected gradient descent procedure, the step size $\alpha := 0.05$ and the maximum allowed iteration steps $S := 500$. From Table V and Fig. 8, we can see that $\boldsymbol{p}$ are more balanced than $\boldsymbol{q}$: the minimum of $\boldsymbol{p}$ is larger than that of $\boldsymbol{q}$ (when $i = 4$), while the maximum of $\boldsymbol{p}$ is smaller than that of $\boldsymbol{q}$ (when $i = 2$).

C. Discrete Maximum Entropy Distribution Using Wasserstein Distance

We let the reference discrete distribution $\boldsymbol{q}$ explicitly be a likelihood distribution of one (worst-case) prior state particle $\boldsymbol{x}$. Suppose $\boldsymbol{q}$ and its induced maximum entropy distribution $\boldsymbol{p}$ have different support sets, as displayed in Fig. 9. The support set $\{\boldsymbol{y}^r|\boldsymbol{x}\}_{r\in[R]}$ of $\boldsymbol{q}$ consists of particles propagated from a 2-dimensional nonlinear measurement equation

$$\begin{cases} y_1^r &= |\sin{(x_1 + x_2 + v_1^r)}|, \\ y_2^r &= |\cos{(e^{x_1 \times x_2 + v_2^r})}|, \end{cases} \quad \forall r \in [4]$$

where $\boldsymbol{x} := [x_1, x_2]^T := [0, 0]^T$ is the fixed prior state particle, and measurement noises v_1^r and v_2^r are sampled from a uniform distribution $\mathcal{U}[0, 1]$. The support set of $\boldsymbol{p}$, however, is constructed by five uniformly sampled points (i.e., green-filled squares No. $1 \sim 5$) and a new measurement (i.e., green-filled square No. 6). Randomly setting the reference distribution

$$\boldsymbol{q} := [0.3700, \ 0.3194, \ 0.0610, \ 0.2496]^T,$$

then the induced maximum entropy distribution $\boldsymbol{p}$ is given as

$$\boldsymbol{p} = [0.2641, \ 0.1272, \ 0.3440, \ 0.2513, \ 0.0071, \ 0.0064]^T,$$

TABLE V
THE REFERENCE DISTRIBUTION AND ITS INDUCED MAXIMUM ENTROPY DISTRIBUTION (USING KULLBACK-LEIBLER DIVERGENCE)

	$\boldsymbol{x}^1$	$\boldsymbol{x}^2$	$\boldsymbol{x}^3$	$\boldsymbol{x}^4$	$\boldsymbol{x}^5$	$\boldsymbol{x}^6$
$\boldsymbol{q}$	0.1993	0.2907	0.0974	0.0492	0.1505	0.2128
$\boldsymbol{p}$	0.1934	0.2492	0.1196	0.0756	0.1602	0.2021

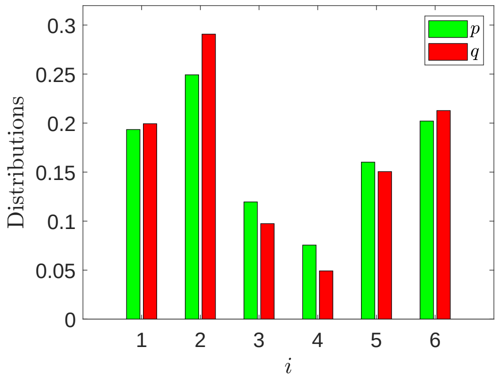


Fig. 8. The maximum entropy distribution $\boldsymbol{p}$ (left bar at each i) induced by the reference distribution $\boldsymbol{q}$ (right bar at each i) using the Kullback-Leibler Divergence.

where $\boldsymbol{p}$ is obtained by Theorem 5. The uncertainty budget θ is set to $\theta := 0.325$ (only for a possible demonstration; other values also applicable). In the projected gradient descent procedure, the step size $\alpha := 0.05$ and the maximum allowed iteration steps $S := 500$. As expected, although the support sets are different, we can still calculate the weights of new supporting points of $\boldsymbol{p}$, and the worst-case likelihood of the new measurement is evaluated as 0.0064. This small-valued likelihood result coincides with our intuition because the new point No. 6 is far away from the supports (i.e., red-filled circles) of $\boldsymbol{q}$.

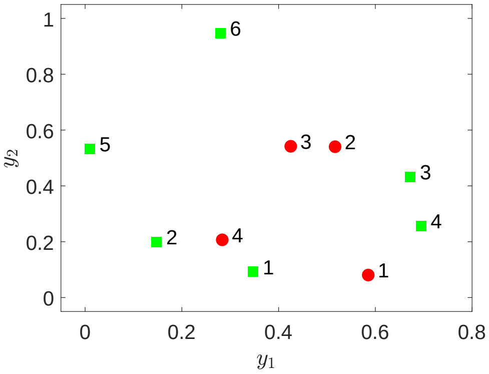


Fig. 9. The maximum entropy distribution $\boldsymbol{p}$ induced by the reference distribution $\boldsymbol{q}$ using the Wasserstein distance. Red-filled circles are supports of $\boldsymbol{q}$, while green-filed squares are supports of $\boldsymbol{p}$.

Alternatively, we may suppose the support set of $\boldsymbol{p}$ is constructed by the union of the support set of $\boldsymbol{q}$ and the new measurement. The supporting points of $\boldsymbol{q}$ are uniformly sampled from $[0, 1] \times [0, 1]$. We have the results in Table VI, in which $\boldsymbol{y}^5$ is a new measurement uniformly sampled from $[0, 1] \times [0, 1]$ as well. The uncertainty budget θ is set to $\theta := 0.01$ (only for a possible demonstration; other values also applicable). In the projected gradient descent procedure, the step size $\alpha := 0.05$ and the maximum allowed iteration steps $S := 500$. From Table VI, it can be seen that the likelihood (of the associated worst-case prior state particle) at this new measurement is 0.0260.

TABLE VI

THE REFERENCE DISTRIBUTION AND ITS INDUCED MAXIMUM ENTROPY DISTRIBUTION (USING WASSERSTEIN DISTANCE)

	$\boldsymbol{y}^1$	$\boldsymbol{y}^2$	$\boldsymbol{y}^3$	$\boldsymbol{y}^4$	$\boldsymbol{y}^5$
Points	0.4314	0.6146	0.0059	0.5459	**0.6206**
	0.5779	0.2699	0.8958	0.1993	**0.3924**
Weights ($\boldsymbol{q}$)	0.3438	0.1316	0.3191	0.2055	/
Weights ($\boldsymbol{p}$)	0.3372	0.1327	0.3191	0.1850	**0.0260**