Robust State Estimation for Linear Systems Under Distributional Uncertainty

Shixiong Wang, *Graduate Student Member, IEEE*, Zhongming Wu, and Andrew Lim

***Abstract*—Modeling uncertainties for real linear systems are unavoidable. These uncertainties can significantly degrade the performance of optimal state estimators designed for nominal system models. The challenge is quantifying such uncertainties and devising robust estimators that are insensitive to them. This paper is therefore concerned with distributionally robust state estimation for linear Markov systems. We propose a new modeling framework that describes uncertainties using a family of distributions so that the worst-case robust estimate in the state space is made over the least-favorable distribution. This framework uses only one or two scalars to express the uncertainty set and does not require the structural information of model uncertainties. Furthermore, the moment-based ambiguity set is suggested to embody the distributional uncertainty family. As a result, the estimation problem transforms into a nonlinear semidefinite program with linear constraints, which can be analytically and efficiently solved. Intensive experiments illustrate the advantages of the proposed framework over existing methods.**

***Index Terms*—State estimation, linear system, distributional robustness, model uncertainty, moment ambiguity set, nonlinear semidefinite programming.**

I. Introduction

A. Subject Matter

STATE estimation for linear systems is critical in several areas, such as target tracking, power system monitoring, geodesy, control and automatics (e.g., robotics), and astronautics (e.g., satellite attitude determination). Mathematically, we are concerned with the state estimation problem for the linear system [1]–[3]

$$\begin{cases} \boldsymbol{x}_k = \boldsymbol{F}_{k-1}\boldsymbol{x}_{k-1} + \boldsymbol{G}_{k-1}\boldsymbol{w}_{k-1}, \\ \boldsymbol{y}_k = \boldsymbol{H}_k\boldsymbol{x}_k + \boldsymbol{v}_k, \end{cases} \tag{1}$$

where k denotes the discrete time, $\boldsymbol{x}_k \in L^2_n(dP_{\boldsymbol{x}_k}) \subset \mathbb{R}^n$ is the state vector, $\boldsymbol{y}_k \in L^2_m(dP_{\boldsymbol{y}_k}) \subset \mathbb{R}^m$ is the measurement vector, and $\boldsymbol{w}_{k-1} \in L^2_p(dP_{\boldsymbol{w}_{k-1}}) \subset \mathbb{R}^p$ and $\boldsymbol{v}_k \in L^2_m(dP_{\boldsymbol{v}_k}) \subset \mathbb{R}^m$ are the process noise and measurement noise, respectively. $L^2_c(dP)$ denotes a c-dimensional L^2 space equipped with the probability measure dP rather than the usual Lebesgue measure. In the canonical settings [1]–[3], the linear system (1) is assumed to hold the following properties:

1) $\boldsymbol{x}_0 \sim \mathcal{N}_n(\bar{\boldsymbol{x}}_0, \boldsymbol{\Pi}_0)$, $\boldsymbol{w}_k \sim \mathcal{N}_p(\boldsymbol{\mu}^{\boldsymbol{w}}_k, \boldsymbol{Q}_k)$, and $\boldsymbol{v}_k \sim \mathcal{N}_m(\boldsymbol{\mu}^{\boldsymbol{v}}_k, \boldsymbol{R}_k)$, where $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a d-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$;
2) $\forall j \neq k$, $\mathbb{E}\boldsymbol{w}_k\boldsymbol{x}^T_0 = \boldsymbol{0}$, $\mathbb{E}\boldsymbol{v}_k\boldsymbol{x}^T_0 = \boldsymbol{0}$, $\mathbb{E}\boldsymbol{w}_k\boldsymbol{w}^T_j = \boldsymbol{0}$, and $\mathbb{E}\boldsymbol{v}_k\boldsymbol{v}^T_j = \boldsymbol{0}$;
3) $\forall k, j$, $\mathbb{E}\boldsymbol{v}_k\boldsymbol{w}^T_j = \boldsymbol{0}$;
4) $\boldsymbol{\mu}^{\boldsymbol{w}}_k, \boldsymbol{\mu}^{\boldsymbol{v}}_k$ are exactly known and typically $\boldsymbol{\mu}^{\boldsymbol{w}}_k \equiv \boldsymbol{0}, \boldsymbol{\mu}^{\boldsymbol{v}}_k \equiv \boldsymbol{0}$;
5) $\boldsymbol{Q}_k$ and $\boldsymbol{R}_k$ are exactly known;
6) $\boldsymbol{F}_{k-1}$, $\boldsymbol{G}_{k-1}$, and $\boldsymbol{H}_k$ are exactly known.

It is well known that the reputed Kalman filter gives the optimal solution to the above problem in the sense of the following: (a) the (linear unbiased) minimum variance estimation [3]; (b) the least/minimum mean square error estimation [2], [4], [5]; (c) the regularized least square estimation [6]; (d) the Bayesian *a posteriori* mean estimation of $\boldsymbol{x}_k$ conditioned on the measurement process $\{\boldsymbol{y}_i\}$ where $0 \leq i \leq k$ [4], [7] (also recall Sherman's theorem); or (e) the orthogonal projection of $\boldsymbol{x}_k$ onto the stochastic Hilbert space spanned by the corresponding innovation process (or equivalently, spanned by the measurement process $\{\boldsymbol{y}_i\}$ where $0 \leq i \leq k$) generated from the linear system (1) [2], [8].

However, for many real problems in engineering, the assumptions are always violated either individually or in batch form. In other words, the system (1) usually suffers from nontrivial and uncertain modeling errors, e.g., uncertain channel characteristics in wireless communication [9], [10], unknown maneuvers in target tracking [11], [12], uncertain attacks/faults in sensor networks [13], unknown noise statistics of sensors [14], and outliers in ultrawideband (UWB) range measurements [15]. The solutions are standard for cases when assumptions 1), 2), and 3) are breached, for example, non-Gaussian (e.g., heavy-tailed)

Manuscript received November 4, 2020; revised May 31, 2021, August 21, 2021, and September 28, 2021; accepted October 4, 2021. Date of publication October 7, 2021; date of current version November 9, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tokunbo Ogunfunmi. The work of Shixiong Wang and Andrew Lim was supported in part by the National Research Foundation of Singapore under Grant NRF-RSS2016-004. The work of Andrew Lim was supported in part by the National Natural Science Foundation of China under Grant 72071139. The work of Zhongming Wu was supported in part by Natural Science Foundation of China under Grant 12001286, and in part by the Startup Foundation for Introducing Talent of NUIST under Grant 2020r003. *(Corresponding author: Zhongming Wu.)*

Shixiong Wang is with the Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore 117576, Singapore (e-mail: s.wang@u.nus.edu).

Zhongming Wu is with the School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: wuzm@nuist.edu.cn).

Andrew Lim was with the Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore 117576, Singapore. He is now with the School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China, and also with the Red Jasper Holdings, Singapore 128424, Singapore (e-mail: isealim@nus.edu.sg).

This article has supplementary downloadable material available at https://doi.org/10.1109/TSP.2021.3118540, provided by the authors.

Digital Object Identifier 10.1109/TSP.2021.3118540

Kalman-like filters [16]–[19] and correlated/colored Kalman filters [1], [3]. However, $\mathbb{E}\boldsymbol{w}_k\boldsymbol{x}_0^T = \mathbf{0}$ and $\mathbb{E}\boldsymbol{v}_k\boldsymbol{x}_0^T = \mathbf{0}$ are always required. Thus, in this paper, we consider only modeling uncertainties 4), 5), and 6) for the linear system (1).

B. Literature Review

Many categories of methods have been developed to address the state estimation problem for the linear system (1) under uncertainties 4), 5), and/or 6). According to the appearance time and philosophical/mathematical complexity of the first inspirational work in each category, we can assign these methodological categories into five generations.

The first-generation methods include representative suboptimal filters, such as fading-memory Kalman-like filters [20]–[22], adaptive Kalman-like filters [23]–[25], multiple-model Kalman bank filters [26], [27], and finite-horizon-memory Kalman-like filters [28, Section V], [29]. These methods represent the first to be considered in practice due to their high computational efficiency (at least for some specific problems) and simplicity.

The second-generation methods include robust Kalman filters for uncertain noise variances [30], [31], $\mathcal{H}_\infty$ filters [28], [32], set-valued Kalman-like filters [33], risk-sensitive (i.e., exponential-cost) Kalman-like filters [8], [34], [35], guaranteed-cost (i.e., upper-bound [11], [36]) Kalman-like filters [37], and their extensions. These filters are robustified by minimizing the worst-case estimation error while sacrificing the estimation performance under nominal conditions. The main disadvantage of this generation is that the existence or stability conditions at every time step must be guaranteed by adjusting some parameters (e.g., γ in Eq. (8) of [28], or α_k in [11]), which prevents online operations [6]. Extensions to these methods involve making a trade-off between robustness and performance [38], [39] or considering a greater number of general uncertainty types [39]–[41].

The third-generation methods include unknown-input Kalman-like filters [42]–[47] and filters for stochastic parametric uncertainties [9], [48], [49]. Specifically, the unknown-input Kalman-like filters treat modeling uncertainties as unknown inputs exerted on the nominal model, while the filters for stochastic parametric uncertainties regard modeling uncertainties as random variables/vectors imposed on nominal system matrices (i.e., $\boldsymbol{F}_k$, $\boldsymbol{G}_k$, and $\boldsymbol{H}_k$). Moreover, in stochastic parametric uncertainty settings, the autocorrelation matrix of the state vector is typically assumed to lie in a predesigned polytope [9], [48]. These two categorical methods are suitable (sometimes highly effective) for some specific settings of system uncertainties when fortunately given the structural information of the system's uncertainties, for example, given G_k in [42] or given Eq. (3) in [9]. Notable extensions include solutions for the case where unknown inputs and measurement outliers exist simultaneously [46], as well as for the case where unknown inputs exist in multiple-model settings [47], etc.

The fourth-generation methods are represented by [6], where the modeling uncertainties are norm-constrained and added to nominal system matrices. Although classic and popular in state-space estimation theory, the framework in [6] has a major limitation in that it is difficult to determine the structural parameters, for example, to select the proper structures of $\boldsymbol{M}_i$, $\boldsymbol{\Delta}_i$, $\boldsymbol{E}_{f,i}$, and $\boldsymbol{E}_{g,i}$ in Eq. (41) of [6], because they are usually matrices/vectors with many entries to be designed. The extensions of this framework include [50]–[52], etc.

This paper studies a new framework that is as general as the third-generation representatives in [9], [42] and the fourth-generation representative in [6]. However, it does not require a filter designer to determine the structure of the modeling uncertainties (e.g., G_k in [42]; $\boldsymbol{F}_{i,k-1}$, $\boldsymbol{G}_{i,k-1}$ in [9]; $\boldsymbol{M}_i$, $\boldsymbol{\Delta}_i$, $\boldsymbol{E}_{f,i}$, and $\boldsymbol{E}_{g,i}$ in [6]), and only a few (typically one to two) scalar parameters are employed to describe the uncertainties. The new framework is termed the *distributionally robust state estimation* for linear Markov systems and is a member of the fifth-generation methods. In this new framework, the modeling uncertainties are expressed using a family of probability distributions. The worst-case state estimator, i.e., the robust estimator, takes effect over the least-favorable distribution.

Note that the literature is listed in perspective, not in strict chronology. Further discussions on the mentioned state-of-the-art frameworks are presented in Section V.

C. Highlights of Contributions

1) We propose a new framework of robust state estimation for the linear system (1), which requires only a few scalar parameters to describe the modeling uncertainties.
2) We use the moment-based ambiguity set to express the modeling uncertainties so that the distributionally robust state estimation problem can be reformulated into a nonlinear semidefinite program (NSDP) with linear constraints.
3) We present an analytical and computationally efficient method to solve the associated NSDP.
4) We prove that the traditional fading-memory Kalman-like filter, which was empirically invented, is a distributionally robust state estimation solution.

D. Paper Structure

In Section II, preliminaries on distributionally robust optimization, optimal estimation, and an optimization trick are reviewed. Section III formulates the distributionally robust state estimation problem for the linear system (1), and the solution under the moment-based ambiguity set is discussed in Section IV. Section V compares the proposed distributionally robust state estimation framework with existing methods, in which the distinctions, advantages, and disadvantages of different frameworks are highlighted. In Section VI, intensive experiments are conducted to compare the performance of the proposed estimator with that of existing state estimators. The conclusions presented in Section VII complete this paper.

E. Notations

$\mathbb{R}^d$ denotes the d-dimensional Euclidean space. $L^2_c(dP)$ denotes a c-dimensional L^2 space equipped with the probability measure dP rather than the usual Lebesgue measure. $\mathbb{E}\boldsymbol{X}$ denotes the expectation of the random matrix $\boldsymbol{X}$. $\mathrm{Tr}[\boldsymbol{A}]$ indicates the trace of a square matrix $\boldsymbol{A}$. Let $\boldsymbol{I}$ and $\boldsymbol{0}$ represent an identity and a null matrix with appropriate dimensions, respectively. Let $[\cdot]^T$ denote the transpose of a matrix. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are deterministic matrices, $\langle \boldsymbol{A}, \boldsymbol{B}\rangle := \mathrm{Tr}[\boldsymbol{A}^T\boldsymbol{B}]$. However, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are random vectors (column by default) in $L^2_c(dP)$, then $\langle \boldsymbol{x}, \boldsymbol{y}\rangle := \mathbb{E}\boldsymbol{x}\boldsymbol{y}^T$ defines the inner product in this stochastic Hilbert space $L^2_c(dP)$. We use $\mathbb{S}^d$ as the collection of all d-dimensional symmetric matrices in $\mathbb{R}^{d\times d}$, and $\mathbb{S}^d_+$ (resp. $\mathbb{S}^d_{++}$) of all d-dimensional symmetric positive semidefinite (resp. positive definite) matrices in $\mathbb{S}^d$. For any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ in $\mathbb{S}^d$, let $\boldsymbol{A} \succeq \boldsymbol{B}$ mean that $\boldsymbol{A} - \boldsymbol{B} \in \mathbb{S}^d_+$ and $\boldsymbol{A} \succ \boldsymbol{B}$ mean that $\boldsymbol{A} - \boldsymbol{B} \in \mathbb{S}^d_{++}$. Let $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote a d-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Let $\mathcal{D}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote any distribution including but not limited to a Gaussian. For any deterministic vector $\boldsymbol{x}$ (column by default), we use $\|\boldsymbol{x}\| := \sqrt{\boldsymbol{x}^T\boldsymbol{x}}$ (resp. $\|\boldsymbol{x}\|_{\boldsymbol{W}} := \sqrt{\boldsymbol{x}^T\boldsymbol{W}\boldsymbol{x}}$) to denote its (resp. weighted) Euclidean norm. Let $\boldsymbol{Y}_k$ denote the set of measurements $\{\boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k\}$. We use $\hat{\boldsymbol{x}}$ to indicate an estimate of the random vector $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}} := \boldsymbol{x} - \hat{\boldsymbol{x}}$ the estimation error. Therefore, the corresponding estimation error covariance is $\langle \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}\rangle$ if $\hat{\boldsymbol{x}}$ is an unbiased estimate. Alternatively, $\langle \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}\rangle$ denotes the mean square estimation error if $\hat{\boldsymbol{x}}$ is biased. See Appendix A for further details.

II. Preliminaries

A. Distributionally Robust Optimization

Distributionally robust optimization, originating from statistical game theory (cf. mixed strategy) [53] and robust statistics [54], is currently popular in academic communities, such as the fields of operations research [55], machine learning [56], and systems control [57]. Suppose the domain of the decision vector $\boldsymbol{x}$ is $\mathcal{X}$ and the parameter vector of an optimization problem is $\boldsymbol{\xi}$ with its support Ξ. In many application scenarios, we do not know the real distribution $\mathbb{P}_{\boldsymbol{\xi}}$ of $\boldsymbol{\xi}$. However, we can assume that $\mathbb{P}_{\boldsymbol{\xi}}$ lies in a family of distributions $\mathcal{F}$ with some properties. Therefore, we have a robust optimization problem over $\mathcal{F}$ that considers the parameters' uncertainties as

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}_{\boldsymbol{\xi}}\in\mathcal{F}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{\xi}}}[f(\boldsymbol{x}, \boldsymbol{\xi})], \tag{2}$$

where $f(\cdot,\cdot)$ is the objective function. Here, $\mathcal{F}$ is termed as a **ambiguity set**, given as

$$\mathcal{F} = \left\{ \mathbb{P}_{\boldsymbol{\xi}} \,\middle|\, \begin{array}{l} \boldsymbol{\xi} \sim \mathbb{P}_{\boldsymbol{\xi}} \\ \mathbb{P}_{\boldsymbol{\xi}}(\boldsymbol{\xi} \in \Xi) = 1 \\ \text{other requirements} \end{array} \right\}.$$

Thus, the ambiguity set $\mathcal{F}$ forms a distributional uncertainty space for the modeling uncertainties of the objective function $f(\cdot,\cdot)$. Typically, the said "other requirements" could be some constraints imposed on the moments of $\boldsymbol{\xi}$ [58] or a metric/divergence of distributions such as the Kullback–Leibler (KL) divergence [59]

$$\mathcal{F}_{KL} = \left\{ \mathbb{P}_{\boldsymbol{\xi}} \,\middle|\, \begin{array}{l} \boldsymbol{\xi} \sim \mathbb{P}_{\boldsymbol{\xi}} \\ \mathbb{P}_{\boldsymbol{\xi}}(\boldsymbol{\xi} \in \Xi) = 1 \\ \mathrm{KL}(\mathbb{P}_{\boldsymbol{\xi}} \| \bar{\mathbb{P}}_{\boldsymbol{\xi}}) \le \theta \end{array} \right\}, \tag{3}$$

or the Wasserstein metric [60]

$$\mathcal{F}_{W} = \left\{ \mathbb{P}_{\boldsymbol{\xi}} \,\middle|\, \begin{array}{l} \boldsymbol{\xi} \sim \mathbb{P}_{\boldsymbol{\xi}} \\ \mathbb{P}_{\boldsymbol{\xi}}(\boldsymbol{\xi} \in \Xi) = 1 \\ \mathrm{W}(\mathbb{P}_{\boldsymbol{\xi}}, \bar{\mathbb{P}}_{\boldsymbol{\xi}}) \le \theta \end{array} \right\}, \tag{4}$$

or others including the τ-divergence [35], ϕ-divergence [59], $\alpha/\beta/\gamma$-divergence [61], etc., where $\mathrm{KL}(\cdot\|\cdot)$ defines the KL divergence, $\mathrm{W}(\cdot,\cdot)$ defines the Wasserstein metric, and we suppose that the nominal distribution of $\boldsymbol{\xi}$ is $\bar{\mathbb{P}}_{\boldsymbol{\xi}}$. Intuitively, $\mathcal{F}_{KL}$ and $\mathcal{F}_W$ mean that although we do not know the real underlying distribution $\mathbb{P}_{\boldsymbol{\xi}}$, we believe that $\mathbb{P}_{\boldsymbol{\xi}}$ lies in a ball centered at $\bar{\mathbb{P}}_{\boldsymbol{\xi}}$ with a radius of θ.

Suppose that $\boldsymbol{x}^*$ and $\mathbb{P}^*_{\boldsymbol{\xi}}$ solve the distributionally robust optimization problem (2). We term $\boldsymbol{x}^*$ the worst-case robust solution and $\mathbb{P}^*_{\boldsymbol{\xi}}$ the least-favorable (i.e., worst-case) distribution.

B. Optimal Estimation

The linear system model (1) induces two stochastic vector processes $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{y}_k\}$, where $k = 0, 1, 2, \cdots$ and $\forall k$, $\boldsymbol{x}_k \in L^2_n(dP_{\boldsymbol{x}_k})$, $\boldsymbol{y}_k \in L^2_m(dP_{\boldsymbol{y}_k})$. Let $\mathcal{H}'_{\boldsymbol{Y}_k}$ denote the stochastic Hilbert space generated by $\{\boldsymbol{y}_k\}$ up to and including k [62]

$$\mathcal{H}'_{\boldsymbol{Y}_k} := \left\{ \boldsymbol{g}(\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) \,\middle|\, \boldsymbol{g} : \underbrace{\mathbb{R}^m \times \cdots \times \mathbb{R}^m}_{k+1} \longmapsto \mathbb{R}^n \right\}, \tag{5}$$

where $\boldsymbol{g}(\cdot)$ is any second-moment-finite Borel measurable function, which might be nonlinear. Meanwhile, let $\mathcal{H}_{\boldsymbol{Y}_k}$ denote the stochastic Hilbert space spanned by $\{\boldsymbol{1}, \boldsymbol{Y}_k\}$ [62]

$$\mathcal{H}_{\boldsymbol{Y}_k} := \left\{ \boldsymbol{b}_k + \sum_{i=0}^{k} \boldsymbol{A}_i \boldsymbol{y}_i \,\middle|\, \boldsymbol{b}_k \in \mathbb{R}^n; \boldsymbol{A}_0, \ldots, \boldsymbol{A}_k \in \mathbb{R}^{n\times m} \right\}. \tag{6}$$

It is well known that the optimal estimate of $\boldsymbol{x}_k$ given $\boldsymbol{Y}_k$ in the sense of minimum mean square error is the unique orthogonal projection of $\boldsymbol{x}_k$ onto $\mathcal{H}'_{\boldsymbol{Y}_k}$. For the special case when $\{\boldsymbol{x}_k\} \cup \{\boldsymbol{y}_k\}$ are jointly Gaussian, the optimal estimate of $\boldsymbol{x}_k$ given $\boldsymbol{Y}_k$ in the sense of minimum mean square error is the unique orthogonal projection of $\boldsymbol{x}_k$ onto $\mathcal{H}_{\boldsymbol{Y}_k}$. However, regardless of whether it is Gaussian or not, the unique orthogonal projection of $\boldsymbol{x}_k$ onto $\mathcal{H}_{\boldsymbol{Y}_k}$ gives the optimal **linear** estimation [62]. In view of the optimal Bayesian posterior estimation theory [4], [7] (cf. Sherman's theorem), this projection point is the same as the conditional mean of $\boldsymbol{x}_k$ given $\boldsymbol{Y}_k$, i.e., $\hat{\boldsymbol{x}}_k = \mathbb{E}(\boldsymbol{x}_k \mid \boldsymbol{Y}_k)$, which minimizes the mean square estimation error [8]

$$\hat{\boldsymbol{x}}_k = \operatorname*{arginf}_{\boldsymbol{\psi}(\cdot)\in\mathcal{H}'_{\boldsymbol{Y}_k}} \langle \boldsymbol{x}_k - \boldsymbol{\psi}(\boldsymbol{Y}_k), \boldsymbol{x}_k - \boldsymbol{\psi}(\boldsymbol{Y}_k)\rangle \tag{7}$$

whose joint state-measurement distribution is $\mathbb{P}_k(\boldsymbol{x}_k, \boldsymbol{Y}_k)$.

In particular, in the linear case (e.g., jointly Gaussian), this optimal Bayesian estimator admits a linear form [8]

$$\hat{\boldsymbol{x}}_k = \bar{\boldsymbol{x}}_k + \boldsymbol{\Sigma}_{xY,k}\boldsymbol{\Sigma}^{-1}_{YY,k}\left[\boldsymbol{Y}_k - \bar{\boldsymbol{Y}}_k\right], \tag{8}$$

where $\bar{\boldsymbol{x}}_k$ and $\bar{\boldsymbol{Y}}_k$ are *a priori* expectations of $\boldsymbol{x}_k$ and $\boldsymbol{Y}_k$, respectively, $\boldsymbol{\Sigma}_{xY,k} := \langle \boldsymbol{x}_k - \bar{\boldsymbol{x}}_k, \boldsymbol{Y}_k - \bar{\boldsymbol{Y}}_k \rangle$, and $\boldsymbol{\Sigma}_{YY,k} := \langle \boldsymbol{Y}_k - \bar{\boldsymbol{Y}}_k, \boldsymbol{Y}_k - \bar{\boldsymbol{Y}}_k \rangle$. With a slight abuse of notation, we note that in (8), $\boldsymbol{Y}_k - \bar{\boldsymbol{Y}}_k := \operatorname{col}\{\boldsymbol{y}_i - \bar{\boldsymbol{y}}_i\}_{0 \le i \le k}$,

$$\langle \boldsymbol{x}_k, \boldsymbol{Y}_k \rangle := [\langle \boldsymbol{x}_k, \boldsymbol{y}_0 \rangle, \langle \boldsymbol{x}_k, \boldsymbol{y}_1 \rangle, \ldots, \langle \boldsymbol{x}_k, \boldsymbol{y}_k \rangle],$$

and

$$\langle \boldsymbol{Y}_k, \boldsymbol{Y}_k \rangle := [\langle \boldsymbol{y}_i, \boldsymbol{y}_j \rangle]_{0 \le i, j \le k}.$$

In other words, $\langle \boldsymbol{Y}_k, \boldsymbol{Y}_k \rangle$ is a block matrix, and the block-type entry at the i^{th} row and j^{th} column is defined by $\langle \boldsymbol{y}_i, \boldsymbol{y}_j \rangle$. As a result, the minimum mean square estimation error is given as

$$\langle \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k, \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k \rangle = \boldsymbol{\Sigma}_{xx,k} - \boldsymbol{\Sigma}_{xY,k} \boldsymbol{\Sigma}_{YY,k}^{-1} \boldsymbol{\Sigma}_{Yx,k}, \tag{9}$$

where $\boldsymbol{\Sigma}_{Yx,k} = \boldsymbol{\Sigma}_{xY,k}^T$ and $\boldsymbol{\Sigma}_{xx,k} := \langle \boldsymbol{x}_k - \bar{\boldsymbol{x}}_k, \boldsymbol{x}_k - \bar{\boldsymbol{x}}_k \rangle$. Eq. (9) implies that introduction of the information of $\boldsymbol{x}_k$ from $\boldsymbol{Y}_k$ helps reduce (resp. improve) the estimation error (resp. performance) of $\boldsymbol{x}_k$. In contrast, if $\boldsymbol{x}_k$ is statistically independent of $\boldsymbol{Y}_k$, we have $\boldsymbol{\Sigma}_{xY,k} \equiv \boldsymbol{0}$, admitting $\hat{\boldsymbol{x}} = \bar{\boldsymbol{x}}$ and $\langle \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k, \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k \rangle = \boldsymbol{\Sigma}_{xx,k}$; i.e., there is no improvement in estimation performance after introducing $\boldsymbol{Y}_k$.

However, as a state estimation problem, the measurements $\boldsymbol{y}_k$ are available in sequence one by one, not in block as $\boldsymbol{Y}_k$. Therefore, we need to design a time-incremental version [63] (i.e., recursive form [8]) of the optimal estimator (8). Namely,

$$\inf_{\boldsymbol{\psi}_k(\cdot) \in \mathcal{H}'_{\boldsymbol{y}_k}} \langle \boldsymbol{x}_k - \boldsymbol{\psi}_k(\boldsymbol{y}_k), \boldsymbol{x}_k - \boldsymbol{\psi}_k(\boldsymbol{y}_k) \rangle \tag{10}$$

with the joint state-measurement distribution conditioned on the previous measurements $\mathbb{P}_k(\boldsymbol{x}_k, \boldsymbol{y}_k \mid \boldsymbol{Y}_{k-1})$. Note that $\mathcal{H}'_{\boldsymbol{y}_k}$ is different from $\mathcal{H}'_{\boldsymbol{Y}_k}$.

According to [8], (10) with $\mathbb{P}_k(\boldsymbol{x}_k, \boldsymbol{y}_k \mid \boldsymbol{Y}_{k-1})$ is equivalent to (7) with $\mathbb{P}_k(\boldsymbol{x}_k, \boldsymbol{Y}_k)$. Therefore, $\hat{\boldsymbol{x}}_k$ in (8) also reads

$$\hat{\boldsymbol{x}}_k = \bar{\boldsymbol{x}}_k + \boldsymbol{\Sigma}_{xy,k} \boldsymbol{\Sigma}_{yy,k}^{-1} [\boldsymbol{y}_k - \bar{\boldsymbol{y}}_k], \tag{11}$$

where $\bar{\boldsymbol{x}}_k$ and $\bar{\boldsymbol{y}}_k$ are **conditional** *a priori* expectations of $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ given $\boldsymbol{Y}_{k-1}$, respectively; $\boldsymbol{\Sigma}_{xy,k} := \langle \boldsymbol{x}_k - \bar{\boldsymbol{x}}_k, \boldsymbol{y}_k - \bar{\boldsymbol{y}}_k \rangle$; $\boldsymbol{\Sigma}_{yy,k} := \langle \boldsymbol{y}_k - \bar{\boldsymbol{y}}_k, \boldsymbol{y}_k - \bar{\boldsymbol{y}}_k \rangle$. In this case, $\bar{\boldsymbol{x}}_k = \mathbb{E}(\boldsymbol{x}_k | \boldsymbol{Y}_{k-1}) = \boldsymbol{F}_{k-1} \hat{\boldsymbol{x}}_{k-1}$ and $\bar{\boldsymbol{y}}_k = \mathbb{E}(\boldsymbol{y}_k | \boldsymbol{Y}_{k-1}) = \boldsymbol{H}_k \boldsymbol{F}_{k-1} \hat{\boldsymbol{x}}_{k-1}$, leading (11) to

$$\hat{\boldsymbol{x}}_k = \boldsymbol{F}_{k-1} \hat{\boldsymbol{x}}_{k-1} + \boldsymbol{\Sigma}_{xy,k} \boldsymbol{\Sigma}_{yy,k}^{-1} [\boldsymbol{y}_k - \boldsymbol{H}_k \boldsymbol{F}_{k-1} \hat{\boldsymbol{x}}_{k-1}], \tag{12}$$

which has a recursive form from $\hat{\boldsymbol{x}}_{k-1}$ to $\hat{\boldsymbol{x}}_k$. In addition, the minimum mean square estimation error reads

$$\langle \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k, \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k \rangle = \boldsymbol{\Sigma}_{xx,k} - \boldsymbol{\Sigma}_{xy,k} \boldsymbol{\Sigma}_{yy,k}^{-1} \boldsymbol{\Sigma}_{yx,k}, \tag{13}$$

where $\boldsymbol{\Sigma}_{yx,k} = \boldsymbol{\Sigma}_{xy,k}^T$ and $\boldsymbol{\Sigma}_{xx,k} := \langle \boldsymbol{x}_k - \bar{\boldsymbol{x}}_k, \boldsymbol{x}_k - \bar{\boldsymbol{x}}_k \rangle$. Since the estimator (12) is unbiased, the minimum mean square estimation error matrix coincides with the minimum estimation error covariance matrix. The filter (12) is obviously the canonical Kalman filter. Explicit expressions of $\boldsymbol{\Sigma}_{xx,k}$, $\boldsymbol{\Sigma}_{xy,k}$, and $\boldsymbol{\Sigma}_{yy,k}$ are straightforward to derive and can also be obtained from the canonical Kalman filter.

C. An Optimization Equivalence

Given a matrix variable $\boldsymbol{X}$ and its convex and compact domain $\mathcal{X}$, the optimization problems $\min_{\boldsymbol{X} \in \mathcal{X}} \boldsymbol{X}$ and $\min_{\boldsymbol{X} \in \mathcal{X}} \operatorname{Tr}[\boldsymbol{X}]$ have the same optimal solution $\boldsymbol{X}^*$ since the trace operator is monotonically increasing. Therefore, in this paper, when we mention minimizing a matrix (recall that $\boldsymbol{A} \succeq \boldsymbol{B}$ means $\boldsymbol{A} - \boldsymbol{B} \succeq \boldsymbol{0}$), we mean minimizing its trace. See also Appendix B.

III. Distributionally Robust State Estimation

A. Motivations

If the nominal system model (1) is exact, the nominal joint state-measurement distribution conditioned on the previous measurements $\mathbb{P}_k(\boldsymbol{x}_k, \boldsymbol{y}_k | \boldsymbol{Y}_{k-1})$ induced from (1) is also exact. Thus, the state estimation in (10) is optimal. However, when modeling uncertainties exist in (1), the true joint state-measurement distribution conditioned on previous measurements $\mathbb{Q}_k(\boldsymbol{x}_k, \boldsymbol{y}_k | \boldsymbol{Y}_{k-1})$ would deviate from the nominal distribution $\mathbb{P}_k(\boldsymbol{x}_k, \boldsymbol{y}_k | \boldsymbol{Y}_{k-1})$. Since we do not exactly know $\mathbb{Q}_k(\boldsymbol{x}_k, \boldsymbol{y}_k | \boldsymbol{Y}_{k-1})$, motivated by distributionally robust optimization theory, we consider the distributionally robust state estimation counterpart of (10) at the time step k

$$\inf_{\boldsymbol{\psi}_k \in \mathcal{H}'_{\boldsymbol{y}_k}} \sup_{\mathbb{Q}_k \in \mathcal{P}_{\boldsymbol{x}_k, \boldsymbol{y}_k | \boldsymbol{Y}_{k-1}}} \langle \boldsymbol{x}_k - \boldsymbol{\psi}_k(\boldsymbol{y}_k), \boldsymbol{x}_k - \boldsymbol{\psi}_k(\boldsymbol{y}_k) \rangle, \tag{14}$$

which minimizes the mean square estimation error over the worst-case distribution, where $\mathcal{P}_{\boldsymbol{x}_k, \boldsymbol{y}_k | \boldsymbol{Y}_{k-1}}$ denotes the ambiguity set of $\mathbb{Q}_k(\boldsymbol{x}_k, \boldsymbol{y}_k \mid \boldsymbol{Y}_{k-1})$. If $\mathcal{P}_{\boldsymbol{x}_k, \boldsymbol{y}_k | \boldsymbol{Y}_{k-1}}$ contains only the nominal distribution $\mathbb{P}_k(\boldsymbol{x}_k, \boldsymbol{y}_k \mid \boldsymbol{Y}_{k-1})$; i.e., we believe that the nominal distribution is exact, the robust estimation problem (14) reduces to the nominal estimation problem (10). Since robust estimation (14) is obtained in the worst-case scenario, the associated robust estimator would be insensitive to modeling uncertainties.

With the robust estimation model (14) on hand, the next steps are 1) to identify the explicit expression of the nominal distribution $\mathbb{P}_k(\boldsymbol{x}_k, \boldsymbol{y}_k \mid \boldsymbol{Y}_{k-1})$, 2) to explicitly define a proper form of the ambiguity set $\mathcal{P}_{\boldsymbol{x}_k, \boldsymbol{y}_k | \boldsymbol{Y}_{k-1}}$ around $\mathbb{P}_k(\boldsymbol{x}_k, \boldsymbol{y}_k \mid \boldsymbol{Y}_{k-1})$, and 3) to derive tractable reformulation(s) of (14) based on $\mathcal{P}_{\boldsymbol{x}_k, \boldsymbol{y}_k | \boldsymbol{Y}_{k-1}}$. We will progressively work on the three problems in the next subsection.

B. Moment-Based Distributionally Robust State Estimation

First, we find the nominal distribution $\mathbb{P}_k(\boldsymbol{x}_k, \boldsymbol{y}_k \mid \boldsymbol{Y}_{k-1})$. For notation brevity, let $\boldsymbol{z}_k := [\boldsymbol{x}_k^T, \boldsymbol{y}_k^T]^T$. From (1), the nominal distribution conditioned on $\boldsymbol{x}_{k-1}$ is known as

$$\mathbb{P}_k(\boldsymbol{z}_k \mid \boldsymbol{x}_{k-1}) = \mathcal{N}_{n+m} \left(\begin{bmatrix} \boldsymbol{F}_{k-1} \\ \boldsymbol{H}_k \boldsymbol{F}_{k-1} \end{bmatrix} \boldsymbol{x}_{k-1}, \boldsymbol{\Sigma}_k^{\circ} \right), \tag{15}$$

where

$$\boldsymbol{\Sigma}_k^{\circ} = \begin{bmatrix} \boldsymbol{G}_{k-1} \boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{0} \\ \boldsymbol{H}_k \boldsymbol{G}_{k-1} \boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{R}_k^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \boldsymbol{G}_{k-1} \boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{0} \\ \boldsymbol{H}_k \boldsymbol{G}_{k-1} \boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{R}_k^{\frac{1}{2}} \end{bmatrix}^T,$$

in which we note that the notation of the square root of a positive semidefinite matrix is $\boldsymbol{Q}^{\frac{1}{2}} (\boldsymbol{Q}^{\frac{1}{2}})^T = \boldsymbol{Q}$. For details of derivation, see Appendix C. The extension of $\boldsymbol{\Sigma}_k^{\circ}$ to the case where $\boldsymbol{w}_{k-1}$ and $\boldsymbol{v}_k$ are correlated is straightforward. We do not discuss it here. Suppose that the conditional distribution of $\boldsymbol{x}_{k-1}$ given $\boldsymbol{Y}_{k-1}$ is $\mathbb{Q}_{k-1}(\boldsymbol{x}_{k-1} \mid \boldsymbol{Y}_{k-1}) = \mathcal{D}_n(\hat{\boldsymbol{x}}_{k-1}, \boldsymbol{V}_{k-1})$, where the optimal (robust) estimate of $\boldsymbol{x}_{k-1}$ is $\hat{\boldsymbol{x}}_{k-1}$ and the corresponding estimation error covariance is $\boldsymbol{V}_{k-1}$. Note that the system (1) is not guaranteed to be exact so that the true distribution $\mathbb{Q}_{k-1}(\boldsymbol{x}_{k-1} \mid \boldsymbol{Y}_{k-1})$ may not be Gaussian. This is because, for example, if $\boldsymbol{F}_k$ contains a random variable at one entry, even though $\boldsymbol{x}_k$ and $\boldsymbol{w}_k$ are white (i.e., mutually independent) Gaussian and $\boldsymbol{G}_k$ is deterministically constant, $\boldsymbol{x}_{k+1}$ will no longer be Gaussian. However, for simplicity, we may limit our estimation problem within the Gaussian filter framework [64] (cf. the unscented [65]/cubature [66] Kalman filter for nonlinear system filtering problem). That is, we use a Gaussian distribution $\mathcal{N}_n(\hat{\boldsymbol{x}}_{k-1}, \boldsymbol{V}_{k-1})$ to approximate $\mathcal{D}_n(\hat{\boldsymbol{x}}_{k-1}, \boldsymbol{V}_{k-1})$ in the state estimation procedure. By using the nominal system model (1), we can obtain the nominal joint state-measurement distribution conditioned on the previous measurements as

$$\mathbb{P}_k(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1}) = \int_{\mathbb{R}^n} \mathbb{P}_k(\boldsymbol{z}_k \mid \boldsymbol{x}_{k-1})\, \mathbb{Q}_{k-1}(d\boldsymbol{x}_{k-1} \mid \boldsymbol{Y}_{k-1}), \tag{16}$$

giving the **time-update step** in the estimation procedure as

$$\mathbb{P}_k(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1}) \sim \mathcal{N}_{n+m}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{17}$$

where

$$\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_{x,k} \\ \boldsymbol{\mu}_{y,k} \end{bmatrix} = \begin{bmatrix} \boldsymbol{F}_{k-1} \\ \boldsymbol{H}_k \boldsymbol{F}_{k-1} \end{bmatrix} \hat{\boldsymbol{x}}_{k-1} \tag{18}$$

and

$$\begin{aligned} \boldsymbol{\Sigma}_k = & \begin{bmatrix} \boldsymbol{F}_{k-1} \\ \boldsymbol{H}_k \boldsymbol{F}_{k-1} \end{bmatrix} \boldsymbol{V}_{k-1} \begin{bmatrix} \boldsymbol{F}_{k-1} \\ \boldsymbol{H}_k \boldsymbol{F}_{k-1} \end{bmatrix}^T + \\ & \begin{bmatrix} \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{0} \\ \boldsymbol{H}_k \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{R}_k^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{0} \\ \boldsymbol{H}_k \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{R}_k^{\frac{1}{2}} \end{bmatrix}^T . \end{aligned} \tag{19}$$

Specifically, in (18), we let $\boldsymbol{\mu}_{x,k} := \boldsymbol{F}_{k-1}\hat{\boldsymbol{x}}_{k-1}$ and $\boldsymbol{\mu}_{y,k} := \boldsymbol{H}_k \boldsymbol{F}_{k-1}\hat{\boldsymbol{x}}_{k-1}$, respectively.

Remark 1: For concepts of time update and measurement update, see, e.g., [3, Chapter 5.1] or [8, Theorem 4]. The measurement-update step of the proposed filter in this paper will be explained later in Corollary 1. The complete estimation procedure will be given in Algorithm 1. □

Second, we define the ambiguity set $\mathcal{P}_{\boldsymbol{z}_k|\boldsymbol{Y}_{k-1}}$ for the real distribution $\mathbb{Q}_k(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1})$. Suppose the true distribution

$$\mathbb{Q}_k(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1}) \sim \mathcal{D}_{n+m}\left(\boldsymbol{c}_k, \boldsymbol{S}_k\right), \tag{20}$$

where $\boldsymbol{c}_k = [\boldsymbol{c}_{x,k}^T, \boldsymbol{c}_{y,k}^T]^T$, $\boldsymbol{c}_{x,k} = \mathbb{E}(\boldsymbol{x}_k|\boldsymbol{Y}_{k-1})$, and $\boldsymbol{c}_{y,k} = \mathbb{E}(\boldsymbol{y}_k|\boldsymbol{Y}_{k-1})$. If we adopt the moment-based ambiguity set proposed in [58], we have the ambiguity set of $\mathbb{Q}_k(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1})$ as (21), where $\gamma_3 \geq 0$ and $\gamma_2 \geq 1 \geq \gamma_1 \geq 0$. Eq (21) means that $\boldsymbol{c}_k$ lies in a ball centered at $\boldsymbol{\mu}_k$ and the real covariance $\boldsymbol{S}_k$ is linearly bounded by the nominal covariance. This ambiguity set describes the trust level that we have towards the nominal distribution (17) and is parameterized by $\boldsymbol{c}_k$ and $\boldsymbol{S}_k$. The trust level is quantified by γ_3, γ_2, and γ_1. The smaller γ_3 is and the closer γ_2 and γ_1 are to one, the more trust we have towards the nominal distribution. Note that when $\gamma_3 = 0$ and $\gamma_2 = \gamma_1 = 1$, the ambiguity set contains only the nominal distribution $\mathbb{P}_k(\boldsymbol{z}_k|\boldsymbol{Y}_{k-1})$ whose mean is $\boldsymbol{\mu}_k$ and covariance is $\boldsymbol{\Sigma}_k$ [cf. (17)]. In highlights, the set (21) defines a space for distributional model uncertainties (cf. the norm-based model uncertainties in [6]).

Third, we derive tractable reformulation(s) of (14). Recall from (6) that under the linear estimation case (i.e., the Gaussian approximation framework [64] is used regardless of whether $\mathbb{P}_k$ and $\mathbb{Q}_k$ are Gaussian), the second-moment-finite Borel measurable optimal estimator $\boldsymbol{\psi}_k(\cdot)$ has an affine form, i.e.,

$$\hat{\boldsymbol{x}}_k = \boldsymbol{\psi}_k(\boldsymbol{y}_k) = \boldsymbol{A}_k \boldsymbol{y}_k + \boldsymbol{b}_k, \tag{22}$$

where $\boldsymbol{A}_k \in \mathbb{R}^{n\times m}$ is a matrix (hence Borel measurable) and $\boldsymbol{b}_k \in \mathbb{R}^n$ is a vector to be determined. Therefore, we have the following theorem.

Theorem 1: With the optimal estimator (22), the distributionally robust state estimation problem (14) admits von Neumann's min-max theorem (i.e., saddle point theorem), i.e.,

$$\begin{aligned} & \inf_{\boldsymbol{A}_k, \boldsymbol{b}_k} \sup_{\boldsymbol{c}_k, \boldsymbol{S}_k} \langle \boldsymbol{x}_k - (\boldsymbol{A}_k \boldsymbol{y}_k + \boldsymbol{b}_k), \boldsymbol{x}_k - (\boldsymbol{A}_k \boldsymbol{y}_k + \boldsymbol{b}_k)\rangle \\ = & \\ & \sup_{\boldsymbol{c}_k, \boldsymbol{S}_k} \inf_{\boldsymbol{A}_k, \boldsymbol{b}_k} \langle \boldsymbol{x}_k - (\boldsymbol{A}_k \boldsymbol{y}_k + \boldsymbol{b}_k), \boldsymbol{x}_k - (\boldsymbol{A}_k \boldsymbol{y}_k + \boldsymbol{b}_k)\rangle. \end{aligned} \tag{23}$$

In addition, if $\boldsymbol{\Sigma}_k \succ \boldsymbol{0}$, this optimization problem is equivalent to a nonlinear semidefinite program (NSDP)

$$\sup_{\boldsymbol{S}_k} \boldsymbol{S}_{xx,k} - \boldsymbol{S}_{xy,k}\boldsymbol{S}_{yy,k}^{-1}\boldsymbol{S}_{yx,k}, \tag{24}$$

Subject to

$$\begin{cases} \boldsymbol{S}_k \preceq \gamma_2 \boldsymbol{\Sigma}_k, \\ \boldsymbol{S}_k \succeq \gamma_1 \boldsymbol{\Sigma}_k, \\ \boldsymbol{S}_k = \begin{bmatrix} \boldsymbol{S}_{xx,k} & \boldsymbol{S}_{xy,k} \\ \boldsymbol{S}_{yx,k} & \boldsymbol{S}_{yy,k} \end{bmatrix} \succ \boldsymbol{0}, \\ \boldsymbol{S}_{xx,k} \succ \boldsymbol{0}, \\ \boldsymbol{S}_{yy,k} \succ \boldsymbol{0}. \end{cases} \tag{25}$$

Proof: Since the optimal estimator $\boldsymbol{\psi}_k(\cdot)$ is parameterized by $\boldsymbol{A}_k$ and $\boldsymbol{b}_k$ and the distributions $\mathbb{Q}_k$ in (21) are parameterized by $\boldsymbol{c}_k$ and $\boldsymbol{S}_k$, (14) is equivalent to the left-hand side of the equality in (23).

Let $\boldsymbol{S}_{xx,k} := \langle \boldsymbol{x}_k - \boldsymbol{c}_{x,k}, \boldsymbol{x}_k - \boldsymbol{c}_{x,k}\rangle$, $\boldsymbol{S}_{yx,k}^T = \boldsymbol{S}_{xy,k} := \langle \boldsymbol{x}_k - \boldsymbol{c}_{x,k}, \boldsymbol{y}_k - \boldsymbol{c}_{y,k}\rangle$, $\boldsymbol{S}_{yy,k} := \langle \boldsymbol{y}_k - \boldsymbol{c}_{y,k}, \boldsymbol{y}_k - \boldsymbol{c}_{y,k}\rangle$, and

$$\boldsymbol{S}_k := \begin{bmatrix} \boldsymbol{S}_{xx,k} & \boldsymbol{S}_{xy,k} \\ \boldsymbol{S}_{yx,k} & \boldsymbol{S}_{yy,k} \end{bmatrix}.$$

Since $\boldsymbol{\Sigma}_k \succ \boldsymbol{0}$, we have $\boldsymbol{S}_k \succ \boldsymbol{0}$. By Schur complement, we further have $\boldsymbol{S}_{xx,k} \succ \boldsymbol{0}$ and $\boldsymbol{S}_{yy,k} \succ \boldsymbol{0}$. This means that (25) is equivalent to

$$\begin{cases} \boldsymbol{S}_k \preceq \gamma_2 \boldsymbol{\Sigma}_k, \\ \boldsymbol{S}_k \succeq \gamma_1 \boldsymbol{\Sigma}_k. \end{cases} \tag{26}$$

With the affine optimal estimator (22), straightforward algebraic manipulations on the objective function of (14), i.e.,

$$\langle \boldsymbol{x}_k - (\boldsymbol{A}_k \boldsymbol{y}_k + \boldsymbol{b}_k), \boldsymbol{x}_k - (\boldsymbol{A}_k \boldsymbol{y}_k + \boldsymbol{b}_k)\rangle$$

gives the objective function in (27).

$$
\begin{aligned}
\inf_{\boldsymbol{A}_k,\boldsymbol{b}_k}\sup_{\boldsymbol{c}_k,\boldsymbol{S}_k}\ & \langle \boldsymbol{I}, \boldsymbol{S}_{xx,k}+\boldsymbol{c}_{x,k}\boldsymbol{c}_{x,k}^T\rangle + \langle \boldsymbol{A}_k^T\boldsymbol{A}_k, \boldsymbol{S}_{yy,k}+\boldsymbol{c}_{y,k}\boldsymbol{c}_{y,k}^T\rangle \\
& - \langle \boldsymbol{A}_k, \boldsymbol{S}_{xy,k}+\boldsymbol{c}_{x,k}\boldsymbol{c}_{y,k}^T\rangle \\
& - \langle \boldsymbol{A}_k^T, \boldsymbol{S}_{yx,k}+\boldsymbol{c}_{y,k}\boldsymbol{c}_{x,k}^T\rangle + 2\langle \boldsymbol{b}_k, \boldsymbol{A}_k\boldsymbol{c}_{y,k}-\boldsymbol{c}_{x,k}\rangle \\
& + \langle \boldsymbol{b}_k, \boldsymbol{b}_k\rangle \qquad (27)
\end{aligned}
$$

For details of derivation, see Appendix D. Hence, problem (14) can be reformulated as solving (27) subject to (21). Since (27) is constraint-free, quadratic and convex in terms of $\boldsymbol{b}_k$, the optimal solution of $\boldsymbol{b}_k$ is obtained by the first-order optimality condition, i.e.,

$$\boldsymbol{b}_k^\star = \boldsymbol{c}_{x,k} - \boldsymbol{A}_k\boldsymbol{c}_{y,k}. \qquad (28)$$

This equality simplifies (27) to

$$
\begin{aligned}
\inf_{\boldsymbol{A}_k}\sup_{\boldsymbol{S}_k}\ & \langle \boldsymbol{I}, \boldsymbol{S}_{xx,k}\rangle + \langle \boldsymbol{A}_k^T\boldsymbol{A}_k, \boldsymbol{S}_{yy,k}\rangle \\
& - \langle \boldsymbol{A}_k, \boldsymbol{S}_{xy,k}\rangle - \langle \boldsymbol{A}_k^T, \boldsymbol{S}_{yx,k}\rangle,
\end{aligned} \qquad (29)
$$

during which the following fact is used: for any deterministic matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$, we have

$$\langle \boldsymbol{A}, \boldsymbol{B}+\boldsymbol{C}\rangle = \langle \boldsymbol{A}, \boldsymbol{B}\rangle + \langle \boldsymbol{A}, \boldsymbol{C}\rangle.$$

The objective function (29) can be further written in a compact form as

$$\inf_{\boldsymbol{A}_k}\sup_{\boldsymbol{S}_k}\langle \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{A}_k \\ -\boldsymbol{A}_k^T & \boldsymbol{A}_k^T\boldsymbol{A}_k \end{bmatrix}, \boldsymbol{S}_k\rangle, \qquad (30)$$

which is subject to (21). To avoid notation clutter, we rewrite (21) as

$$
\begin{cases}
(\boldsymbol{c}_k-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{c}_k-\boldsymbol{\mu}_k) \le \gamma_3, \\
\boldsymbol{S}_k + (\boldsymbol{c}_k-\boldsymbol{\mu}_k)(\boldsymbol{c}_k-\boldsymbol{\mu}_k)^T \preceq \gamma_2\boldsymbol{\Sigma}_k, \\
\boldsymbol{S}_k + (\boldsymbol{c}_k-\boldsymbol{\mu}_k)(\boldsymbol{c}_k-\boldsymbol{\mu}_k)^T \succeq \gamma_1\boldsymbol{\Sigma}_k.
\end{cases} \qquad (31)
$$

Since the ambiguity set (21) is convex and compact in terms of $(\boldsymbol{c}_k, \boldsymbol{S}_k)$ and the objective function in (30)

$$\langle \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{A}_k \\ -\boldsymbol{A}_k^T & \boldsymbol{A}_k^T\boldsymbol{A}_k \end{bmatrix}, \boldsymbol{S}_k\rangle$$

is linear (thus concave) in $\boldsymbol{S}_k$ and positive-definite quadratic (thus convex) in $\boldsymbol{A}_k$, von Neumann's min-max theorem (i.e., saddle point theorem) holds, i.e.,

$$
\begin{aligned}
& \inf_{\boldsymbol{A}_k}\sup_{\boldsymbol{S}_k}\langle \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{A}_k \\ -\boldsymbol{A}_k^T & \boldsymbol{A}_k^T\boldsymbol{A}_k \end{bmatrix}, \boldsymbol{S}_k\rangle, \\
= \ & \\
& \sup_{\boldsymbol{S}_k}\inf_{\boldsymbol{A}_k}\langle \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{A}_k \\ -\boldsymbol{A}_k^T & \boldsymbol{A}_k^T\boldsymbol{A}_k \end{bmatrix}, \boldsymbol{S}_k\rangle.
\end{aligned}
$$

This gives the min-max equality (23). In view that the optimization problem (30) over $\boldsymbol{A}_k$ is constraint-free, differentiable, and convex, the first-order optimality condition, i.e.,

$$\boldsymbol{A}_k\boldsymbol{S}_{yy,k} - \boldsymbol{S}_{xy,k} = \boldsymbol{0},$$

gives the optimal solution of $\boldsymbol{A}_k$ as

$$\boldsymbol{A}_k^\star = \boldsymbol{S}_{xy,k}\cdot\boldsymbol{S}_{yy,k}^{-1}. \qquad (32)$$

This equality simplifies (30) to (24). Note that the objective function (24) is irrelevant to $\boldsymbol{c}_k$. Therefore, to maximize (24), the larger the feasible set of $\boldsymbol{S}_k$, the better. This gives the optimal solution of $\boldsymbol{c}_k$ as

$$\boldsymbol{c}_k^\star = \boldsymbol{\mu}_k. \qquad (33)$$

This equality simplifies (31) to (26), which is equivalent to (25). This completes the proof.

Remark 2: Note that when there are no uncertainties in (1), the ambiguity set contains only the nominal distribution $\mathbb{P}_k(\boldsymbol{z}_k|\boldsymbol{Y}_{k-1})$. Hence, $\boldsymbol{c}_k$ and $\boldsymbol{S}_k$ would be fixed, and $\boldsymbol{c}_k = \boldsymbol{\mu}_k$ and $\boldsymbol{S}_k = \boldsymbol{\Sigma}_k$ always hold. This observation reduces the distributionally robust state estimator (14) to the canonical Kalman filter (10). Moreover, the worst-case estimation error covariance (24) becomes the nominal estimation error covariance (13).□

To further lower the number of parameters of the uncertainty set, motivated by the reputed restricted isometry property [67], we may consider an alternative as

$$
\begin{cases}
\boldsymbol{S}_k \preceq (1+\gamma)\boldsymbol{\Sigma}_k, \\
\boldsymbol{S}_k \succeq (1-\gamma)\boldsymbol{\Sigma}_k, \\
\boldsymbol{S}_k = \begin{bmatrix} \boldsymbol{S}_{xx,k} & \boldsymbol{S}_{xy,k} \\ \boldsymbol{S}_{yx,k} & \boldsymbol{S}_{yy,k} \end{bmatrix} \succ \boldsymbol{0}, \\
\boldsymbol{S}_{xx,k} \succ \boldsymbol{0}, \\
\boldsymbol{S}_{yy,k} \succ \boldsymbol{0},
\end{cases} \qquad (34)
$$

in which $0 \le \gamma < 1$. However, (34) is not equivalent to (25).

Corollary 1 (Measurement-update Step): Suppose that $\boldsymbol{S}_k^\star$ solves the optimization problem (24) and (25). By recalling (22), (28), (32), and (33), the distributionally robust estimator

$$
\begin{aligned}
\mathcal{P}_{\boldsymbol{z}_k|\boldsymbol{Y}_{k-1}} = \mathcal{P}_{\boldsymbol{z}_k|\boldsymbol{Y}_{k-1}}(\boldsymbol{c}_k,\boldsymbol{S}_k) &= \left\{ \mathbb{Q}_k(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1}) \,\middle|\, \begin{array}{l} \boldsymbol{z}_k|\boldsymbol{Y}_{k-1} \sim \mathbb{Q}_k \\ \mathbb{Q}_k(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1}) = \mathcal{D}_{n+m}(\boldsymbol{c}_k,\boldsymbol{S}_k) \\ [\mathbb{E}(\boldsymbol{z}_k|\boldsymbol{Y}_{k-1})-\boldsymbol{\mu}_k]^T\boldsymbol{\Sigma}_k^{-1}[\mathbb{E}(\boldsymbol{z}_k|\boldsymbol{Y}_{k-1})-\boldsymbol{\mu}_k] \le \gamma_3 \\ \mathbb{E}\left[(\boldsymbol{z}_k-\boldsymbol{\mu}_k)(\boldsymbol{z}_k-\boldsymbol{\mu}_k)^T|\boldsymbol{Y}_{k-1}\right] \preceq \gamma_2\boldsymbol{\Sigma}_k \\ \mathbb{E}\left[(\boldsymbol{z}_k-\boldsymbol{\mu}_k)(\boldsymbol{z}_k-\boldsymbol{\mu}_k)^T|\boldsymbol{Y}_{k-1}\right] \succeq \gamma_1\boldsymbol{\Sigma}_k \end{array} \right\} \qquad (21) \\
&= \left\{ \mathbb{Q}_k(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1}) \,\middle|\, \begin{array}{l} (\boldsymbol{c}_k-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{c}_k-\boldsymbol{\mu}_k) \le \gamma_3 \\ \boldsymbol{S}_k + (\boldsymbol{c}_k-\boldsymbol{\mu}_k)(\boldsymbol{c}_k-\boldsymbol{\mu}_k)^T \preceq \gamma_2\boldsymbol{\Sigma}_k \\ \boldsymbol{S}_k + (\boldsymbol{c}_k-\boldsymbol{\mu}_k)(\boldsymbol{c}_k-\boldsymbol{\mu}_k)^T \succeq \gamma_1\boldsymbol{\Sigma}_k \end{array} \right\}
\end{aligned}
$$

in the sense of linear minimum mean square estimation error is given as

$$\begin{aligned}\hat{\boldsymbol{x}}_k &= \boldsymbol{\psi}_k^\star(\boldsymbol{y}_k) = \boldsymbol{A}_k^\star \boldsymbol{y}_k + \boldsymbol{b}_k^\star \\ &= \boldsymbol{\mu}_{x,k} + \boldsymbol{S}_{xy,k}^\star \cdot (\boldsymbol{S}_{yy,k}^\star)^{-1}(\boldsymbol{y}_k - \boldsymbol{\mu}_{y,k}),\end{aligned} \tag{35}$$

and according to Theorem 1, the worst-case estimation error covariance is as

$$\boldsymbol{V}_k = \boldsymbol{S}_{xx,k}^\star - \boldsymbol{S}_{xy,k}^\star(\boldsymbol{S}_{yy,k}^\star)^{-1}\boldsymbol{S}_{yx,k}^\star. \tag{36}$$

Note that $\boldsymbol{\mu}_{x,k}$ and $\boldsymbol{\mu}_{y,k}$ in (35) are defined in (18). Moreover, the least-favorable (i.e., worst-case) conditional distribution of $\boldsymbol{z}_k$ given $\boldsymbol{Y}_{k-1}$ is $\mathbb{Q}_k^\star(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1}) = \mathcal{D}_{n+m}(\boldsymbol{\mu}_k, \boldsymbol{S}_k^\star)$ and the worst-case conditional distribution of $\boldsymbol{x}_k$ given $\boldsymbol{Y}_k$ is $\mathbb{Q}_k^\star(\boldsymbol{x}_k \mid \boldsymbol{Y}_k) = \mathcal{D}_n(\hat{\boldsymbol{x}}_k, \boldsymbol{V}_k)$ [cf. $\mathbb{Q}_{k-1}(\boldsymbol{x}_{k-1} \mid \boldsymbol{Y}_{k-1}) = \mathcal{D}_n(\hat{\boldsymbol{x}}_{k-1}, \boldsymbol{V}_{k-1})$ in (16)]. In the Gaussian filter framework, we have approximately $\mathbb{Q}_k^\star(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1}) = \mathcal{N}_{n+m}(\boldsymbol{\mu}_k, \boldsymbol{S}_k^\star)$ and $\mathbb{Q}_k^\star(\boldsymbol{x}_k \mid \boldsymbol{Y}_k) = \mathcal{N}_n(\hat{\boldsymbol{x}}_k, \boldsymbol{V}_k)$. □

C. Other Types of Ambiguity Sets

This subsection discusses the scenarios when we do not adopt the moment-based ambiguity set. We consider the metrics/divergences of distributions, such as the Kullback–Leibler divergence and the Wasserstein distance. Note that the Kullback–Leibler divergence is not a statistical metric since it does not meet the metric axioms. We do not explicitly discuss the τ-divergence [35] because the conclusions under the Kullback–Leibler divergence remain the same as those under the τ-divergence. When $\tau = 0$, the τ-divergence degenerates to the Kullback–Leibler divergence.

1) Kullback–Leibler Divergence: In this case, the ambiguity set is as (3). See also [63]. When we consider the distributionally robust estimation problem (14), (3) is specified into

$$\mathcal{P}_{\boldsymbol{z}_k|\boldsymbol{Y}_{k-1}} = \{\mathbb{Q}_k(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1}) \,|\mathrm{KL}(\mathbb{Q}_k\|\mathbb{P}_k) \le \theta\}. \tag{37}$$

In general, if we use the τ-divergence, $\mathrm{KL}(\mathbb{Q}_k\|\mathbb{P}_k) \le \theta$ should be replaced with $D_\tau(\mathbb{Q}_k\|\mathbb{P}_k) \le \theta$, where $D_\tau(\mathbb{Q}_k\|\mathbb{P}_k)$ denotes the τ-divergence [35]. Supposing $\mathbb{Q}_k(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1})$ is also Gaussian, Eq. (37) can be explicitly expressed as

$$\begin{aligned}&\mathrm{KL}(\mathbb{Q}_k\|\mathbb{P}_k) \\ &= \frac{1}{2}\left[\|\boldsymbol{c}_k - \boldsymbol{\mu}_k\|^2_{\boldsymbol{\Sigma}_k^{-1}} + \mathrm{Tr}\left[\boldsymbol{\Sigma}_k^{-1}\boldsymbol{S}_k - \boldsymbol{I}\right] - \ln\det\left(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{S}_k\right)\right] \\ &\le \theta.\end{aligned} \tag{38}$$

The corresponding worst-case conditional distribution of $\boldsymbol{z}_k$ given $\boldsymbol{Y}_{k-1}$ is

$$\mathbb{Q}_k^\star(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1}) = \mathcal{N}_{n+m}\left(\boldsymbol{\mu}_k, \boldsymbol{S}_k^\star\right), \tag{39}$$

where

$$\boldsymbol{S}_k^\star = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{xx,k} & \boldsymbol{\Sigma}_{xy,k} \\ \boldsymbol{\Sigma}_{yx,k} & \boldsymbol{\Sigma}_{yy,k}\end{bmatrix}, \tag{40}$$

and $\tilde{\boldsymbol{\Sigma}}_{xx,k}$ is determined by the boundary condition $\mathrm{KL}(\mathbb{Q}_k\|\mathbb{P}_k) = \theta$ [63]. In the τ-divergence case, the forms of the corresponding $\mathbb{Q}_k^\star(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1})$ and $\boldsymbol{S}_k^\star$ are the same as those in (39) and (40), respectively, but $\tilde{\boldsymbol{\Sigma}}_{xx,k}$ is determined instead from the boundary condition $\mathcal{D}_\tau(\mathbb{Q}_k\|\mathbb{P}_k) = \theta$ [35].

Eq. (40) admits that the distributionally robust state estimation under the Kullback–Leibler divergence (in general, the τ-divergence) can be written as

$$\hat{\boldsymbol{x}}_k = \boldsymbol{\mu}_{x,k} + \boldsymbol{\Sigma}_{xy,k} \cdot (\boldsymbol{\Sigma}_{yy,k})^{-1}(\boldsymbol{y}_k - \boldsymbol{\mu}_{y,k}), \tag{41}$$

which is in the same form as the optimal estimation under the nominal distribution, i.e., (11). This means that under the Kullback–Leibler divergence or the τ-divergence, the worst-case conditional distribution of $\boldsymbol{z}_k$ given $\boldsymbol{Y}_{k-1}$ at the current time step **does not directly influence** the optimal robust estimation at the same time step. It is worth mentioning that the τ-divergence (including Kullback–Leibler) distributionally robust estimator generalizes the risk-sensitive estimator (i.e., the exponential-cost estimator) in the sense of allowing the time-varying sensitivity parameter [35], [63].

2) Wasserstein Distance: In this case, the ambiguity set is as (4). See also [68]. When we consider the distributionally robust estimation problem (14), Eq. (4) is specified into

$$\mathcal{P}_{\boldsymbol{z}_k|\boldsymbol{Y}_{k-1}} = \{\mathbb{Q}_k(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1}) \,|\mathrm{W}(\mathbb{Q}_k, \mathbb{P}_k) \le \theta\}. \tag{42}$$

If we suppose $\mathbb{Q}_k(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1})$ is also Gaussian, Eq. (42) can be explicitly expressed as

$$\begin{aligned}&\mathrm{W}(\mathbb{Q}_k, \mathbb{P}_k) \\ &= \sqrt{\|\boldsymbol{c}_k - \boldsymbol{\mu}_k\|^2 + \mathrm{Tr}\left[\boldsymbol{S}_k + \boldsymbol{\Sigma}_k - 2\left(\boldsymbol{\Sigma}_k^{\frac{1}{2}}\boldsymbol{S}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]} \\ &\le \theta.\end{aligned} \tag{43}$$

The corresponding worst-case conditional distribution of $\boldsymbol{z}_k$ given $\boldsymbol{Y}_{k-1}$ is

$$\mathbb{Q}_k^\star(\boldsymbol{z}_k \mid \boldsymbol{Y}_{k-1}) = \mathcal{N}_{n+m}\left(\boldsymbol{\mu}_k, \boldsymbol{S}_k^\star\right), \tag{44}$$

where $\boldsymbol{S}_k^\star$ solves a NSDP subject to (43) [68].

Eq. (44) suggests that the distributionally robust state estimation under the Wasserstein ambiguity set is (35), which is generally not guaranteed to have the same form as (41). This means that, under the Wasserstein distance, the worst-case conditional distribution of $\boldsymbol{z}_k$ given $\boldsymbol{Y}_{k-1}$ at the current time step **directly influences** the optimal robust estimation at the same time step.

3) Comparisons With the Moment Ambiguity Set: Three points must be highlighted. First, note that both the Kullback–Leibler (in general, the τ-divergence) ambiguity set and the Wasserstein ambiguity set require that the real conditional distribution of $\boldsymbol{z}_k$ given $\boldsymbol{Y}_{k-1}$ is Gaussian. Otherwise, there is no explicit equivalence between (37) and (38) and between (42) and (43). This requirement is difficult to satisfy for a linear system under unknown uncertainties. For example, if $\boldsymbol{F}_{k-1}$ contains a random variable at one entry, even though $\boldsymbol{x}_{k-1}$ and $\boldsymbol{w}_{k-1}$ are white (i.e., mutually independent) Gaussian and $\boldsymbol{G}_{k-1}$ is deterministically constant, $\boldsymbol{x}_k$ will no longer be Gaussian. Second, although Gaussian, the Kullback–Leibler (in general, the τ-divergence) ambiguity set and the Wasserstein ambiguity set are highly nonlinear, whereas our moment ambiguity set is linear. Note that an optimization problem over a linear feasible set is generally easier to solve. Specifically, compared with the extremely nonlinear semidefinite program under the Wasserstein ambiguity set [i.e., an NSDP *s.t.* (43)], the nonlinearity of our NSDP under the moment ambiguity set [i.e., (24) *s.t.* (25)] is considerably more moderate, and fortunately, our new NSDP can be analytically (and therefore computationally efficiently) solved. This feature saves a substantial amount of running time. Third, under the Wasserstein ambiguity set, the worst-case conditional distribution of $\boldsymbol{z}_k$ given $\boldsymbol{Y}_{k-1}$ at the current time step (i.e., k) **directly influences** the optimal estimation at the same time step, while under the Kullback–Leibler (in general, the τ-divergence) ambiguity set and the moment ambiguity set [see (48)], it **does not directly influence** the optimal estimation at the same time step. However, this does not mean that the Kullback–Leibler (in general, the τ-divergence) distributionally robust estimator and the moment distributionally robust estimator do nothing to robustify the state estimation. Rather, the effect is indirect: they influence the filter gains in the future instead of the gains at the current time step. More specifically, note that each filter type has a different associated $\boldsymbol{V}_k$ at the time step k. Therefore, according to (19), they have different $\boldsymbol{\Sigma}_{k+1}$ values, which lead to different state estimates at the time step $k+1$.

Algorithm 1: Moment-Based Distributionally Robust Estimator.

```
Definition: $\hat{\boldsymbol{x}}_k$ as the robust state estimate; $\boldsymbol{V}_k$ as the state
  estimation error covariance.
Initialize: $\hat{\boldsymbol{x}}_0$, $\boldsymbol{V}_0$, $\gamma$.
Remark: In (35), $\boldsymbol{c}_k^\star$ has already been replaced with $\boldsymbol{\mu}_k$. In
  general, $\gamma_1$, $\gamma_2$ can be independently initialized without $\gamma$.
  By (47), the robust state estimation results only depend on
  $\gamma_2$. Therefore, we do not initialize $\gamma_1$.
Input: measurement $\boldsymbol{y}_k$, $k = 1, 2, 3, \ldots$
  $\gamma_2 \leftarrow 1 + \gamma$.          // See (34)
  while true do
    // Time-Update Step, i.e., Prior Estimation
    Use (18) and (19) to obtain $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$;

    // Obtain the Worst-Case Scenario
    Use (33) to obtain $\boldsymbol{c}_k^\star$;
    Solve (24) and (25) with (47) to obtain $\boldsymbol{S}_k^\star$;

    // Measurement-Update Step, i.e., Posterior
       Estimation
    Use (35) and (36) to obtain $\hat{\boldsymbol{x}}_k$ and $\boldsymbol{V}_k$;

    // Next Time Step
    $k \leftarrow k + 1$;
  end while
Output:$\hat{\boldsymbol{x}}_k$
```

IV. Solving the Moment-Based Distributionally Robust Estimation Problem

In this section, we first discuss the solution to the nonlinear semidefinite programming problem (24) subject to (25) and then summarize the overall moment-based distributional robust estimation scheme.

A. Solution to the Nonlinear Semidefinite Program

Theorem 2: The NSDP (24) subject to (25) is analytically solved by $\boldsymbol{S}_k^\star = \gamma_2 \boldsymbol{\Sigma}_k$.

Proof: The NSDP (24) subject to (25) is equivalent to

$$\sup_{\boldsymbol{S}_k} \operatorname{Tr}\left[\boldsymbol{S}_{xx,k} - \boldsymbol{S}_{xy,k}\boldsymbol{S}_{yy,k}^{-1}\boldsymbol{S}_{yx,k}\right] \tag{45}$$

subject to (26). Let $f(\boldsymbol{S}_k) := \operatorname{Tr}[\boldsymbol{S}_{xx,k} - \boldsymbol{S}_{xy,k}\boldsymbol{S}_{yy,k}^{-1}\boldsymbol{S}_{yx,k}]$.

The gradient of $f(\boldsymbol{S}_k)$ with respect to $\boldsymbol{S}_k$ admits

$$\nabla_{\boldsymbol{S}_k} f(\boldsymbol{S}_k) = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{S}_{xy,k}\boldsymbol{S}_{yy,k}^{-1} \\ -\boldsymbol{S}_{yy,k}^{-1}\boldsymbol{S}_{yx,k} & \boldsymbol{S}_{yy,k}^{-1}\boldsymbol{S}_{yx,k}\boldsymbol{S}_{xy,k}\boldsymbol{S}_{yy,k}^{-1} \end{bmatrix}. \tag{46}$$

Since the top left block of $\nabla_{\boldsymbol{S}_k} f(\boldsymbol{S}_k)$ (i.e., $\boldsymbol{I}$) is positive definite and its Schur complement is

$$\boldsymbol{S}_{yy,k}^{-1}\boldsymbol{S}_{yx,k}\boldsymbol{S}_{xy,k}\boldsymbol{S}_{yy,k}^{-1} - \boldsymbol{S}_{yy,k}^{-1}\boldsymbol{S}_{yx,k}\boldsymbol{I}^{-1}\boldsymbol{S}_{xy,k}\boldsymbol{S}_{yy,k}^{-1} = \boldsymbol{0} \succeq \boldsymbol{0},$$

we have $\nabla_{\boldsymbol{S}_k} f(\boldsymbol{S}_k) \succeq \boldsymbol{0}$, i.e., positive semidefinite. This means that $f(\boldsymbol{S}_k)$ is a nondecreasing function with respect to $\boldsymbol{S}_k$. Therefore, if we assume that $\boldsymbol{S}_k^\star$ solves the NSDP (45) subject to (26), we must have

$$\boldsymbol{S}_k^\star = \gamma_2 \boldsymbol{\Sigma}_k. \tag{47}$$

This completes the proof.

Remark 3: In the proof of Theorem 2, the following facts are involved.

1) For a block matrix $\boldsymbol{M} := [\begin{smallmatrix}\boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D}\end{smallmatrix}]$, if $\boldsymbol{A}$ is invertible, then the Schur complement of block $\boldsymbol{A}$ of matrix $\boldsymbol{M}$ is defined as $\boldsymbol{M}/\boldsymbol{A} := \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}$. Further, if $\boldsymbol{M}$ is symmetric (i.e., $\boldsymbol{C} = \boldsymbol{B}^T$) and $\boldsymbol{A} \succ 0$, then the matrix $\boldsymbol{M} \succeq \boldsymbol{0}$ if and only if $\boldsymbol{M}/\boldsymbol{A} \succeq \boldsymbol{0}$.
2) If $\boldsymbol{S}$ is a symmetric and invertible variable matrix and $\boldsymbol{A}$ is constant with respect to $\boldsymbol{S}$, then the following identities hold: $\nabla_{\boldsymbol{S}} \operatorname{Tr}[\boldsymbol{S}] = \boldsymbol{I}$; $\nabla_{\boldsymbol{S}} \operatorname{Tr}[\boldsymbol{A}\boldsymbol{S}] = \nabla_{\boldsymbol{S}} \operatorname{Tr}[\boldsymbol{S}\boldsymbol{A}] = \boldsymbol{A}^T$; and $\nabla_{\boldsymbol{S}} \operatorname{Tr}[\boldsymbol{A}^T\boldsymbol{S}^{-1}\boldsymbol{A}] = \nabla_{\boldsymbol{S}} \operatorname{Tr}[\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{A}^T] = -(\boldsymbol{S}^{-1})^T(\boldsymbol{A}\boldsymbol{A}^T)^T(\boldsymbol{S}^{-1})^T = -\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{S}^{-1}$. □

Theorem 2 reveals that $\boldsymbol{S}_{xy,k}^\star \cdot (\boldsymbol{S}_{yy,k}^\star)^{-1}$ equals $\boldsymbol{\Sigma}_{xy,k} \cdot (\boldsymbol{\Sigma}_{yy,k})^{-1}$ so that (35) admits

$$\hat{\boldsymbol{x}}_k = \boldsymbol{\mu}_{x,k} + \boldsymbol{\Sigma}_{xy,k} \cdot (\boldsymbol{\Sigma}_{yy,k})^{-1}(\boldsymbol{y}_k - \boldsymbol{\mu}_{y,k}), \tag{48}$$

which is the same as (41). This implies that under the moment ambiguity set, the optimal robust state estimate is not directly influenced by the worst-case distribution at the current step. In addition, by comparing (47) with [21], we can conclude that the traditional fading-memory Kalman-like filter is a distributionally robust state estimation solution under moment-based ambiguity.

B. Moment-Based Distributionally Robust Estimator

The overall moment-based distributionally robust estimator to the linear system (1) is summarized in Algorithm 1.

C. Computational Complexity

From Remark 2, we know that $\boldsymbol{S}_k^\star \equiv \boldsymbol{\Sigma}_k$ gives the canonical Kalman filter. Since the moment-based distributionally robust state estimator is solved by $\boldsymbol{S}_k^\star = \gamma_2 \boldsymbol{\Sigma}_k$, where γ_2 is just a scalar [cf. (47)], it has the same order of computational complexity as the canonical Kalman filter. Specifically, at each time instant k, the computational complexity is $O(n^3)$ because for a state estimation problem, we usually have $n \geq m$ and $n \geq p$. For detailed analysis, see the online supplementary materials. This means that the moment-based distributionally robust state estimator is computationally as efficient as the canonical Kalman filter.

V. Comparison With Existing Robust State Estimation Frameworks

Regarding modeling uncertainties in (1), the first-generation methods actually do not address the problem from the perspective of robustness. Instead, they adaptively adjust the filter parameters/structures so that the state estimation is consistent with the measurements and the divergences of filters are avoided. For example, the adaptive Kalman filter assumes that modeling uncertainties perturb the process noise covariance $\boldsymbol{Q}_{k-1}$ and/or the measurement noise covariance $\boldsymbol{R}_k$ (i.e., we do not exactly know the true $\boldsymbol{Q}_{k-1}$ or $\boldsymbol{R}_k$) and then estimates $\boldsymbol{Q}_{k-1}$ or $\boldsymbol{R}_k$ when estimating the state. One issue with the adaptive Kalman filter is that addressing the fast-changing statistics of noises is hard (i.e., when the true $\boldsymbol{Q}_{k-1}$ or $\boldsymbol{R}_k$ changes quickly). Likewise, unknown-input filters try to improve the state estimation performance, for example, by estimating the unknown input in the sense of unbiased minimum variance (see [42], [44]), in the sense of maximum likelihood (see [45]), or by leveraging an auxiliary term (see [43]).

The successive four generations (except unknown-input filters in the third generation) are essentially robust filters (i.e., robust state estimators). The worst-case state estimation error covariance matrix (i.e., the upper bound of the state estimation error covariance matrix [11], [37]) is minimized to achieve robustness so that the filter is insensitive to modeling uncertainties.

When modeling uncertainties exist, filter designers must explicitly describe their structures and parameters. For example, in unknown-input filters [42], we study the linear system

$$\begin{cases} \boldsymbol{x}_k = \boldsymbol{F}_{k-1}\boldsymbol{x}_{k-1} + \boldsymbol{\Gamma}_{k-1}\boldsymbol{d}_{k-1} + \boldsymbol{G}_{k-1}\boldsymbol{w}_{k-1}, \\ \boldsymbol{y}_k = \boldsymbol{H}_k\boldsymbol{x}_k + \boldsymbol{v}_k, \end{cases} \tag{49}$$

where $\boldsymbol{d}_{k-1} \in \mathbb{R}^q$ is the unknown input used to describe the modeling uncertainties. Note that the unknown-input $\boldsymbol{d}_k$ may also exist in the measurement dynamics [44]–[47]. Obviously, in this case, the modeling uncertainties are limited to the range space of $\boldsymbol{\Gamma}_{k-1}$. To achieve good estimation performance, the filter designer must carefully determine the structure and entries of $\boldsymbol{\Gamma}_{k-1}$. For another example, in [6], we are concerned with the linear system

$$\begin{cases} \boldsymbol{x}_k = (\boldsymbol{F}_{k-1} + \delta\boldsymbol{F}_{k-1})\boldsymbol{x}_{k-1} + (\boldsymbol{G}_{k-1} + \delta\boldsymbol{G}_{k-1})\boldsymbol{w}_{k-1}, \\ \boldsymbol{y}_k = \boldsymbol{H}_k\boldsymbol{x}_k + \boldsymbol{v}_k, \end{cases} \tag{50}$$

where $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$ are used to model the perturbations imposed on the nominal system matrices $\boldsymbol{F}_{k-1}$ and $\boldsymbol{G}_{k-1}$, respectively. In addition, $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$ are assumed to satisfy the following structure:

$$\left[\, \delta\boldsymbol{F}_{k-1} \; \delta\boldsymbol{G}_{k-1} \,\right] = \boldsymbol{M}_{k-1}\Delta_{k-1}\left[\, \boldsymbol{E}_{f,k-1} \; \boldsymbol{E}_{g,k-1} \,\right], \tag{51}$$

where Δ_{k-1} is an arbitrary contraction operator (i.e., the operator norm is less than one). $\boldsymbol{M}_{k-1}$, $\boldsymbol{E}_{f,k-1}$, and $\boldsymbol{E}_{g,k-1}$ are structure matrices that must be carefully designed. For the third example, we refer to [9], in which the focused linear system is the same as (50), but $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$ are modeled as

$$\begin{cases} \delta\boldsymbol{F}_{k-1} = \sum_{i=1}^{l} \boldsymbol{F}_{i,k-1} \cdot \zeta_{i,k-1} \\ \delta\boldsymbol{G}_{k-1} = \sum_{i=1}^{l} \boldsymbol{G}_{i,k-1} \cdot \zeta_{i,k-1}, \end{cases} \tag{52}$$

where $\zeta_{i,k-1}$ is a random variable with assumed-known statistics; l, $\boldsymbol{F}_{i,k-1}$, and $\boldsymbol{G}_{i,k-1}$ are assumed to be exactly known. For the fourth example, we shall recall the framework introduced in this paper where the modeling uncertainties are described by a family of distributions; see (2), (14), (21), and (25).

In summary, all the exemplified robust estimation frameworks minimize the worst-case state estimation error covariance (viz., the upper bound of the state estimation error covariance), although the uncertainties are described, structured, parameterized, and bounded in different ways. However, the magic of the proposed framework is that only a few scalars [e.g., two scalars γ_1 and γ_2 in (25) or only one scale γ in (34)] rather than subtly designed matrices [e.g., $\boldsymbol{\Gamma}_{k-1}$ in (49); $\boldsymbol{M}_{k-1}$, $\boldsymbol{E}_{f,k-1}$, and $\boldsymbol{E}_{g,k-1}$ in (51); and $\boldsymbol{F}_{i,k-1}$ and $\boldsymbol{G}_{i,k-1}$ in (52)] are required to describe the modeling uncertainties. This means that when ONLY the nominal model (1) is available and we do not know how uncertainties exist, our framework takes the least risk of failure. This is because if the structure matrices in (49), (51), and (52) are inappropriately provided, the estimation performance degrades significantly. However, to design proper structure matrices, additional information on real system perturbations is required. From the perspective of information, additional information (e.g., structures and values) on modeling uncertainties helps improve the estimation performance. As we can expect, if we can exactly model the system in the form of (49), (51), or (52), the specifically designed frameworks might outperform our new distributional framework. The claims in this section will be validated in the experiments.

VI. Experiments

This section compares the state estimation performance of the existing filters with our newly proposed filter. All the source data and codes are available online at GitHub: https://github.com/Spratm-Asleaf/DRSE. Interested readers can reproduce and/or verify the claims of this paper by changing the parameters or codes themselves. To ensure clarity regarding figures, we distinguish different results only by different colors. Readers who have problems identifying colors could change the codes to generate different line types and markers to display the results. Additional experiments can be found in the online supplementary materials.

We continue studying the classical instance discussed in [6], [63], [68], i.e.,

$$\boldsymbol{F}_k^{real} = \begin{bmatrix} 0.9802 & 0.0196 + \alpha \cdot \Delta_k \\ 0 & 0.9802 \end{bmatrix},$$

$$\boldsymbol{G}_k = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{H}_k = \begin{bmatrix} 1 & -1 \end{bmatrix},$$

$$\boldsymbol{Q}_k = \begin{bmatrix} 1.9608 & 0.0195 \\ 0.0195 & 1.9605 \end{bmatrix}, \boldsymbol{R}_k = \begin{bmatrix} 1 \end{bmatrix},$$

where the random scalar $\Delta_k \in \mathcal{U} := [-1, 1]$ denotes the real perturbations imposed on the system and $\mathcal{U}$ defines its support; α is a multiplicative coefficient (in [6], α was fixed as 0.099). In this state estimation problem, the nominal system matrix is

$$\boldsymbol{F}_k = \begin{bmatrix} 0.9802 & 0.0196 \\ 0 & 0.9802 \end{bmatrix}.$$

A. Candidate Filters

According to Section I-B and Section V, we are motivated to implement the following filters for comparison.

1) **TMKF**: the canonical Kalman filter with the true model. Note that in the simulation, we know the underlying true model over time (viz., $\boldsymbol{F}_k^{real}$). Therefore, this method theoretically gives the best estimate of state in the sense of linear unbiased minimum estimation error covariance.
2) **KF**: the Kalman filter (with the nominal model $\boldsymbol{F}_k$).
3) **Adaptive**: the adaptive Kalman-like filter [25] (cf. [24]).
4) **Fading**: the adaptive fading-memory Kalman-like filter [22].
5) H_∞ : the H_∞ filter [28].
6) **UB**: the upper-bound Kalman-like filter [11].
7) **UI**: the unknown-input Kalman-like filter [42].
8) **SPU**: the filter for stochastic parametric uncertainties [9].
9) **SNKF**: Sayed's norm-constrained Kalman-like filter [6].
10) τ**-KF**: the τ-divergence Kalman-like filter [35].
11) **WKF**: the Wasserstein Kalman-like filter [68].
12) **MKF**: the moment-based distributionally robust state estimator introduced in this paper.

The twelve methods above are representatives of the five filter generations beginning with the canonical Kalman filter in the 1960 s and ending with the filter proposed in this paper. We do not consider the set-valued Kalman-like filter [33], the guaranteed-cost Kalman-like filter [37], and the traditional risk-sensitive Kalman-like filter [34] because in [6], they have been substantially studied and compared. Note that the τ-divergence Kalman-like filter [35] generalizes the Kullback–Leibler Kalman-like filter [63] (when $\tau = 0$, the τ-divergence gives the Kullback–Leibler divergence). Note also that the traditional risk-sensitive Kalman-like filter is a special case of the τ-divergence Kalman-like filter [35], [63].

B. Results With Exactly Known Structures of Uncertainties

In this illustration, we first assume that the structural information of the modeling uncertainties is known. Namely, all the filtering frameworks know that the uncertainties impact the first entry of the state vector.

In all methods, the initial state estimate is set as $\hat{\boldsymbol{x}}_0 = [0, 0]^T$ and its corresponding state estimation error covariance $\boldsymbol{V}_0$ is set as $\mathrm{diag}\{1, 1\}$, where $\mathrm{diag}\{\cdot\}$ denotes a diagonal matrix [6], [63], [68]. All the parameters of each filter are tuned to perform (nearly) optimally for the studied instance (when Δ_k randomly changes and $\alpha = 1$). The details of the parameter settings are available in the disclosed codes at GitHub.

In the H_∞ filter, we select γ (see [28]) such that the existence condition of the H_∞ filter is guaranteed. From simulation validation, we select $\gamma = 102$.

In Sayed's norm-constrained Kalman-like filter [6], we set $\boldsymbol{M}_{k-1} = [0.0198, 0]^T$, $\boldsymbol{E}_{f,k-1} = [0, \alpha/0.0198]$, and $\boldsymbol{E}_{g,k-1} = [0, 0]^T$ in (51), such that

$$\boldsymbol{M}_{k-1}\boldsymbol{E}_{f,k-1} = \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix}.$$

Namely, we assume that we know exactly the structural information of the modeling uncertainties.

In the unknown-input Kalman-like filter [42], we set $\boldsymbol{\Gamma}_k = [1, 0]^T$ in (49) because, as supposed before, we know that the modeling uncertainties influence the first entry of the state vector, and we need to guarantee Assumption 1 of [42].

In the filter for stochastic parametric uncertainties [9], we have $l = 1$ in (52),

$$\boldsymbol{F}_{1,k-1} = \begin{bmatrix} 0 & \sqrt{3}\alpha \\ 0 & 0 \end{bmatrix},$$

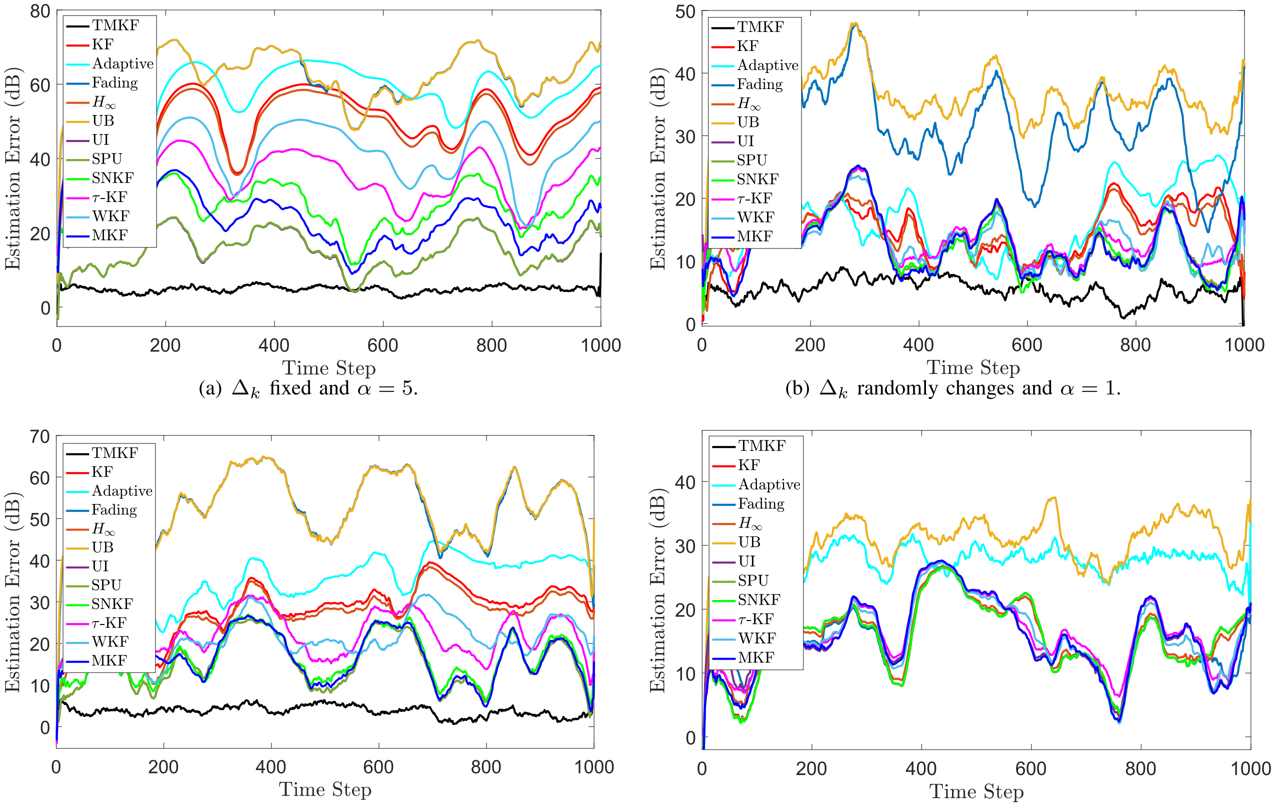
and $\boldsymbol{G}_{1,k-1} = \boldsymbol{0}$. Note that $\zeta_{i,k-1}$ is assumed to have unit variance in [9]. However, in the studied instance, the variance of Δ_k is $[1 - (-1)]^2/12 = 1/3$ if uniformly distributed. Thus, the right-top entry of $\boldsymbol{F}_{1,k-1}$ is $\sqrt{3}\alpha$ rather than α. The initial polytope is constructed as a hypercube centered at $\mathrm{diag}\{1, 1\}$ with an edge length of 1. Namely, the vertexes of this polytope are $\mathrm{diag}\{0.5, 0.5\}$, $\mathrm{diag}\{0.5, 1.5\}$, $\mathrm{diag}\{1.5, 0.5\}$, and $\mathrm{diag}\{1.5, 1.5\}$ (i.e., $p = 4$). In other words, we construct the initial polytope for the autocorrelation matrix (of the state vector) around the initial state estimation error covariance (recall that the initial state estimation error covariance has been set as $\mathrm{diag}\{1, 1\}$).

In the τ-divergence Kalman-like filter [35], we let $\tau = 0$ (therefore, the τ-divergence Kalman-like filter specifies the Kullback–Leibler Kalman-like filter [63]) and the radius of the ambiguity set be 1.5×10^{-4}.

In the Wasserstein Kalman-like filter [68], the radius of the ambiguity set is set to 0.1.

In our moment-based distributionally robust filter, $\gamma = 0.02$, and therefore, $\gamma_2 = 1.02$ (see Algorithm 1).

Suppose each simulation episode runs $T = 1000$ discrete-time steps. The estimation error at each time step k (shown in figures) is measured in decibels (dB) by $10\log_{10}[(x_{1,k} - \hat{x}_{1,k})^2 + (x_{2,k} - \hat{x}_{2,k})^2]$, where $x_{1,k}$ (resp.



(a) Δ_k fixed and $\alpha = 5$.

(b) Δ_k randomly changes and $\alpha = 1$.

(c) Δ_k randomly changes and $\alpha = 5$.

(d) $\alpha = 0$.

Fig. 1. The results with prior known structural information (for H_∞, the prior parametric information is known, i.e., $\gamma = 102$). In (a), SPU and UI coincide.

$x_{2,k}$) is the first (resp. second) component of the state vector $\boldsymbol{x}_k$ and $\hat{x}_{1,k}$ (resp. $\hat{x}_{2,k}$) denotes its estimate. The overall estimation error of each episode (shown in tables) is measured by the root mean square error (RMSE) as

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

In principle, we should repeat the experiment independently several times and compare the average estimation performance, just as [6] and [68] did where 500 independent episodes were run. However, from the simulations, it is evident that the relative estimation performance of each filter compared to other filters is the same for every independent episode. Therefore, without loss of generality, we display only the estimation results of each filter for a single episode. Interested readers could validate this claim with the disclosed codes themselves. We conduct each of the following four experiments once (rather than many as explained).

- First, we fix $\Delta_k = 1$ for all k and let $\alpha = 5$; i.e., the modeling uncertainty is constant but unknown over time. The results are shown in Fig. 1(a) and Table I .
- Second, let Δ_k randomly take its value with the uniform distribution from its support $\mathcal{U}$ at each step k and let $\alpha = 1$; i.e., the modeling uncertainty is a stochastic process over time, but with relatively small magnitude. The results are shown in Fig. 1(b) and Table II.

TABLE I
THE RESULTS WHEN $\Delta_k = 1$ FIXED AND $\alpha = 5$

Filter	RMSE	Avg Time	Filter	RMSE	Avg Time
TMKF	2.59	2.83e-5	UI	9.57	3.45e-5
KF	562.28	1.91e-5	SPU	9.66	17364.00e-5
Adaptive	1172.48	3.02e-5	SNKF	37.27	80.07e-5
Fading	2347.60	3.98e-5	τ-KF	91.18	28.17e-5
H_∞	474.88	16.41e-5	WKF	192.28	425.60e-5
UB	2349.63	68.91e-5	MKF	40.48	11.16e-5

Note: The results are obtained by a laptop with 8 G RAM and an Intel(R) Core(TM) i7-8850H CPU @ 2.60 GHz. **Avg Time**: Average Execution Time at each time step (unit: seconds); **1e-5**: 1×10^{-5}.

TABLE II
THE RESULTS WHEN Δ_k RANDOMLY CHANGES AND $\alpha = 1$

Filter	RMSE	Avg Time	Filter	RMSE	Avg Time
TMKF	3.19	1.91e-5	UI	8.19	3.48e-5
KF	8.56	1.87e-5	SPU	8.10	16958.00e-5
Adaptive	11.38	2.42e-5	SNKF	8.41	80.07e-5
Fading	163.84	3.52e-5	τ-KF	8.33	27.72e-5
H_∞	8.33	13.34e-5	WKF	7.53	425.00e-5
UB	187.70	24.67e-5	MKF	7.83	11.33e-5

See Table I for table notes.

- Third, let Δ_k randomly take its value with the uniform distribution from its support $\mathcal{U}$ at each step k and let $\alpha = 5$; i.e., the modeling uncertainty is a stochastic process over time, but with relatively large magnitude. The results are shown in Fig. 1(c) and Table III.

TABLE III
The Results When Δ_k Randomly Changes and $\alpha = 5$

Filter	RMSE	Avg Time	Filter	RMSE	Avg Time
TMKF	2.51	2.44e-5	UI	11.11	3.27e-5
KF	45.21	2.06e-5	SPU	11.10	17548.00e-5
Adaptive	94.64	3.10e-5	SNKF	18.30	83.14e-5
Fading	1429.59	4.07e-5	τ-KF	26.25	29.96e-5
H_∞	38.67	15.51e-5	WKF	20.56	420.94e-5
UB	1426.57	328.47e-5	MKF	14.88	11.46e-5

See Table I for table notes.

TABLE IV
The Results When $\alpha = 0$

Filter	RMSE	Avg Time	Filter	RMSE	Avg Time
TMKF	9.61	2.04e-5	UI	11.03	3.63e-5
KF	9.61	1.93e-5	SPU	9.61	16914.00e-5
Adaptive	44.96	2.55e-5	SNKF	18.30	83.14e-5
Fading	11.97	3.02e-5	τ-KF	10.79	29.87e-5
H_∞	9.78	14.53e-5	WKF	10.33	416.44e-5
UB	77.04	11.99e-5	MKF	10.84	11.19e-5

See Table I for table notes.

TABLE V
The Results Without Prior structural/parametric Information

Filter	RMSE	Avg Time	Filter	RMSE	Avg Time
TMKF	2.35	1.85e-5	UI	269.71	3.33e-5
KF	29.35	1.95e-5	SPU	Fail to Work	
Adaptive	39.95	2.55e-5	SNKF	69.53	79.96e-5
Fading	1116.28	3.20e-5	τ-KF	20.11	28.78e-5
H_∞	243.15	6.50e-5	WKF	14.73	422.05e-5
UB	1118.04	185.13e-5	MKF	10.57	10.75e-5

See Table I for table notes.

- Fourth, we let $\alpha = 0$; i.e., there are no modeling uncertainties. The results are shown in Fig. 1(d) and Table IV .

Note that the UB filter [11], which is in essence a kind of fading-memory Kalman-like filter (cf. [21]) is inappropriate for the instance discussed in this paper because Assumption (19) of [11] requires that $\operatorname{rank}(\boldsymbol{H}_k) = n$. However, the instance that we are working on admits that $\operatorname{rank}(\boldsymbol{H}_k) = 1 \neq n = 2$. Therefore, the UB filter produces extremely unsatisfactory experimental results.

From Fig. 1 and Tables I–IV, the conclusions below can be outlined.

1) The TMKF always gives the best performance because it works with the true system model.
2) The UB filter does not work well for the instance that we are studying.
3) The traditional adaptive Kalman-like filter and the adaptive fading-memory Kalman-like filter perform worse than the canonical Kalman filter on the studied instance.
4) The H_∞ filter can be a choice because it at least outperforms the KF when modeling uncertainties exist.
5) The MKF is essentially the traditional fading-memory Kalman-like filter with a fixed fading factor γ_2. However, it outperforms the adaptive-factor fading-memory Kalman-like filters in [22] and [11]. This phenomenon is interesting and exists for the conventional risk-sensitive Kalman-like filter (which has a fixed risk-sensitive parameter) and the Kullback–Leibler divergence-based Kalman-like filter (which has an adaptive risk-sensitive parameter) [63, Fig. 5]. Therefore, it is not always beneficial to adaptively adjust the risk-sensitive parameter of a risk-sensitive Kalman-like filter and the fading factor of a fading-memory Kalman-like filter.
6) When we know the structural information of the modeling uncertainties, the UI filter and SPU filter are two powerful solutions. However, the computational efficiency of the SPU filter is extremely low since at each time step, the SPU filter needs to numerically solve a semidefinite program (it is well known that solving a semidefinite program is generally challenging).
7) The SNKF is another good choice when we know the structural information of the modeling uncertainties.
8) Although the structural information of the modeling uncertainties is not used, the distributionally robust state estimators are still promising. In addition, compared with the τ-KF and WKF, the newly proposed MKF is attractive due to its high computational efficiency and estimation performance.
9) When there are no modeling uncertainties, i.e., when the nominal model is the true model, the KF works best compared with any robust filtering frameworks (see Table IV). This is because the KF is theoretically optimal for an exact system model. Therefore, the cost of robustness under uncertain conditions is to sacrifice optimality under perfect conditions. More specifically, robust filters are robust under uncertain conditions, but they are not optimal under perfect conditions; the canonical Kalman filter is optimal under perfect conditions, but it is not robust under uncertain conditions.

C. Results Without Exactly Known Structures of Uncertainties

For experiments in this subsection, we no longer assume that the structural information of the modeling uncertainties is known. In other words, we know neither the perturbation structure existing as $[\begin{smallmatrix} 0 & \alpha \\ 0 & 0 \end{smallmatrix}]$, nor the exact value of α. Thus, we may give improper structure matrices for different filtering frameworks. For example, we may instead (mistakenly) set $\boldsymbol{E}_{f,k-1} = [5, 0]$ in (51), $\boldsymbol{\Gamma}_k = [0, 1]^T$ in (49), and $\boldsymbol{F}_{1,k-1} = [\begin{smallmatrix} 0 & 0 \\ 3 & 0 \end{smallmatrix}]$ in (52). To clarify further, all the frameworks no longer know that the uncertainties impact the first entry of the state vector. Instead, they might assume that uncertainties impact the second entry of the state vector. In addition, for the H_∞ filter, we do not select a large enough γ (see [28]) in advance to guarantee the existence of the H_∞ filter. Alternatively, we arbitrarily select $\gamma = 25$ (rather than minimally required 102). As we can expect, this incorrect structural/parametric information will mislead the filters and degrade the estimation performance. In this experiment, we set $\alpha = 5$ and let Δ_k take random uniformly distributed values from its support. The results are given in Fig. 2 and Table V.

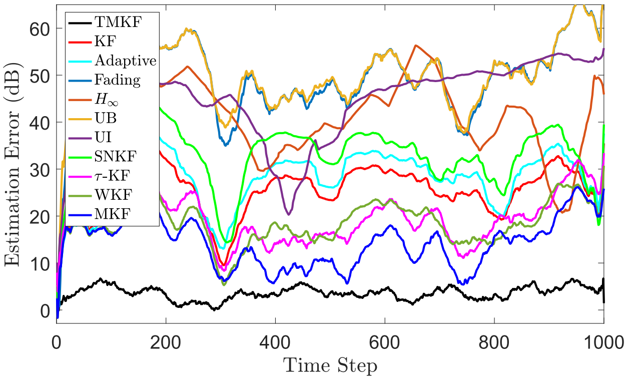


Fig. 2. The results without prior structural/parametric information. In this case, only the distributionally robust estimators can outperform the true-model Kalman filter. Structural/parametric-information-aware filters such as UI, SNKF, and H_∞ perform poorly. Moreover, SPU even fails to work (and therefore is not plotted).

From the results, we can observe the potential of the newly proposed distributionally robust estimation framework. Namely, even if we do not know the correct structural information of the modeling uncertainties, we do not have a risk of encountering a disaster. However, compared with Fig. 1, we can see that the cost of this powerful robustness is that the distributionally robust estimation framework never accounts for the (partially) known information of the modeling uncertainties. Therefore, when given some exact structural information of the modeling uncertainties, the distributionally robust estimation framework would perform worse than the specifically designed structure-information-aware filtering frameworks. The discrepancy between absolute robustness and optimality, however, is unavoidable from the perspective of information.

D. Suggestions on Tuning the Size of the Ambiguity Set

The size of the ambiguity set (21) is controlled by three scalars, namely, γ_1, γ_2, and γ_3. To include the nominal values of the mean (i.e., $\boldsymbol{\mu}_k$) and covariance (i.e., $\boldsymbol{\Sigma}_k$) in the ambiguity set (21), we must have $\gamma_3 \geq 0$ and $\gamma_2 \geq 1 \geq \gamma_1 \geq 0$. Note that when $\gamma_3 = 0$ and $\gamma_2 = \gamma_1 = 1$, the ambiguity set (21) contains only the nominal distribution whose mean is $\boldsymbol{\mu}_k$ and covariance is $\boldsymbol{\Sigma}_k$. However, the moment-based distributionally robust state estimator requires $\gamma_3 \equiv 0$ [see (33)], is irrelevant to γ_1 [see (47)], and only depends on γ_2. Therefore, γ_1 can be any value in [0,1], and we only investigate how to tune γ_2. The caption of Fig. 3 lists the RMSEs of the candidate filters.

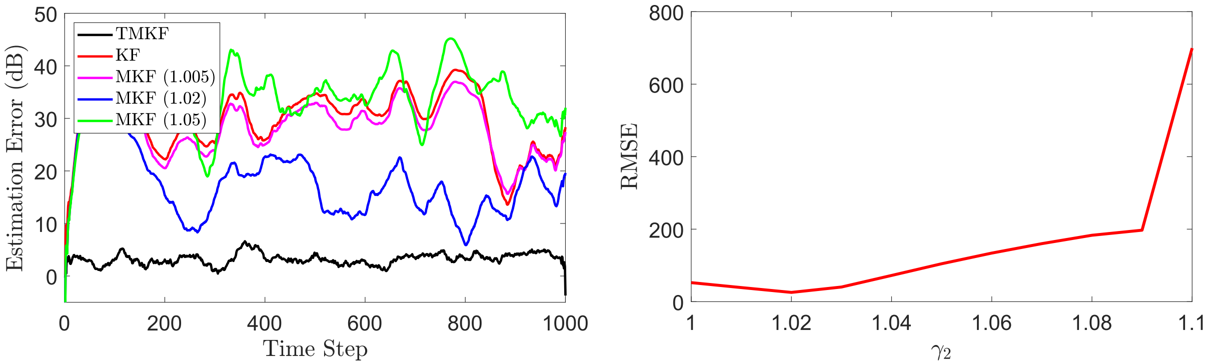
From Fig. 3, it is evident that γ_2 can be neither too large nor too small to obtain satisfactory estimation performance. The robust state estimator with a too small γ_2 value has insufficient robustness (i.e., insufficient ability against uncertainties), while that with a too large γ_2 value is too conservative to produce satisfactory estimation performance. Unfortunately, the optimal tuning method for γ_2 is unknown (unless γ_2 can be directly given in the model identification stage that defines $\boldsymbol{F}_k$, $\boldsymbol{G}_k$, and $\boldsymbol{H}_k$). At present, the authors can only suggest that readers try appropriate values for their specific problems. Nevertheless, we believe that tuning a scalar γ_2 is significantly easier than tuning structural matrices $\boldsymbol{\Gamma}_{k-1}$ in (49), $\boldsymbol{M}_{k-1}$, $\boldsymbol{E}_{f,k-1}$ and $\boldsymbol{E}_{g,k-1}$ in (51), and $\boldsymbol{F}_{i,k-1}$ and $\boldsymbol{G}_{i,k-1}$ in (52).

A possible tuning method of γ_2 for a real system involves leveraging a controller. This approach is reasonable because a natural purpose for state estimation is to design a state-feedback controller. In this case, the controller is parameterized by γ_2. Hence, we can choose the value with which the controller works best, e.g., for high-accuracy output tracking (i.e., the real output is close enough to the expected output). However, controller design is not the unique reason for state estimation. Sometimes, we are only concerned with monitoring the state of a system without adjusting its quantities (i.e., state and output). In this case, the rule of thumb is to choose the value that makes the estimated state [or some transform(s) of it] be consistent, as much as possible, with subjective (e.g., qualitative) or objective (e.g., quantitative) evidence collected somehow from somewhere else.

VII. Conclusions

In this paper, the distributionally robust state estimation method for linear Markov systems is proposed. We integrate the existing Kullback–Leibler-divergence robust state estimation method [63], the τ-divergence robust state estimation method [35], the Wasserstein-distance robust state estimation method [68], and the newly proposed moment-based robust state estimation method into a unified framework. The characteristics are outlined below.

1) The proposed framework uses only a few scalars (i.e., the radius/scale of the ambiguity set) rather than structured matrices with many entries to describe the modeling uncertainties. Therefore, it does not require *a priori* structural information of modeling uncertainties.
2) Our framework uses a family of distributions to describe the modeling uncertainties, after which the state estimation is performed over the worst-case distribution. In essence, borrowing phrasings from existing frameworks, the upper bound of the estimation error covariance is minimized.
3) The family of distributions [i.e., the ambiguity set, see (2)] can be described by several means, such as the τ-divergence, the Kullback–Leibler divergence (37), the Wasserstein distance (42), and the proposed moment-based ambiguity set (21). The detailed comparisons among those different ambiguity sets can be revisited in Section III-C. The newly proposed moment-based filter in this paper is most attractive due to it having the highest computational efficiency, which can be attributed to the analytical tractability of the linearly constrained NSDP (recall Section IV-A). In addition, the state estimation performance of the moment-based filter is better than that of the τ-divergence filter (when $\tau = 0$, i.e., the Kullback–Leibler divergence) and the Wasserstein-distance filter for the studied instance.
4) The distributionally robust estimation framework outperforms other existing structural-information-aware frameworks when we do not have *a priori* structural information of modeling uncertainties. However, when we



(a) The results with $\gamma_2 = 1.005$, $\gamma_2 = 1.02$, and $\gamma_2 = 1.05$.

(b) RMSE versus γ_2 (varying from 1 to 1.1).

Fig. 3. The results with different γ_2 values. In (a), RMSE: TMKF = 2.44, KF = 48.75, MKF (1.005) = 39.97, MKF (1.02) = 12.52, MKF (1.05) = 106.43.

know some structural information of modeling uncertainties, the newly proposed distributionally robust estimation framework performs worse than the existing specifically designed structural-information-aware frameworks.

5) The risk-sensitive Kalman-like filter and the fading-memory Kalman-like filter are distributionally robust state estimation solutions under Kullback–Leibler divergence (in general, τ-divergence) ambiguity and moment-based ambiguity, respectively. However, it is not always beneficial to adaptively adjust the risk-sensitive parameter of a risk-sensitive Kalman-like filter and the fading factor of a fading-memory Kalman-like filter.

From Fig. 3, we can see that the proposed algorithm is not robust with respect to the size of the ambiguity set (i.e., γ_2). Unfortunately, the optimal or convincing tuning method for the size of ambiguity sets (e.g., γ_2 in this paper; ρ in [68]; and c in [35], [63]) has yet to be found. We invite scholars in this field to collaborate with the authors on addressing the two issues below in the future.

1) How can γ_2 be tuned in a real system where the true state is unknown?
2) How can we ensure that the state estimator remains tuned over varying conditions? In other words, how do we select a time-varying $\gamma_{2,k}$ where k denotes the discrete time?

Although imperfect, the proposed method is still promising because tuning a scalar γ_2 is easier than tuning structural matrices $\boldsymbol{\Gamma}_{k-1}$ in (49), $\boldsymbol{M}_{k-1}$, $\boldsymbol{E}_{f,k-1}$, and $\boldsymbol{E}_{g,k-1}$ in (51), and $\boldsymbol{F}_{i,k-1}$ and $\boldsymbol{G}_{i,k-1}$ in (52).

Appendix A
Intuitive Explanations for Notations

The notation in this paper is consistent with that in [8] and/or [62]. In this appendix, we provide intuitive explanations for some notations for better readability. If a random vector $\boldsymbol{x} \in L_c^2(dP)$, we have $\mathrm{Tr}[\int \boldsymbol{x}\boldsymbol{x}^T dP] = \int \boldsymbol{x}^T\boldsymbol{x} dP < \infty$; i.e., $\boldsymbol{x}$ has a finite second moment. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are random vectors in the stochastic Hilbert space $L_c^2(dP)$, the inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \mathbb{E}\boldsymbol{x}\boldsymbol{y}^T$ denotes their cross-correlation matrix. Therefore, $\langle \boldsymbol{x}, \boldsymbol{x} \rangle$ denotes the second-order moment matrix (i.e., autocorrelation matrix) of the random vector $\boldsymbol{x}$. When $\boldsymbol{x}$ is centered (viz., zero-mean), $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = \mathbb{E}\boldsymbol{x}\boldsymbol{x}^T$ denotes the covariance matrix.

Appendix B
On Matrix-Type Objective

In the state estimation literature, some people directly work on minimizing a covariance matrix (see, e.g., [8] and [2, Chapter 3], while others work on minimizing its trace (see, e.g., [1], [3]). They give the same solution because the trace operator is monotonically increasing. Since this paper follows the notation convention of [8], we study a matrix-type objective.

Appendix C
Derive (15)

By (1), we have

$$\begin{cases} \boldsymbol{x}_k = \boldsymbol{F}_{k-1}\boldsymbol{x}_{k-1} + \boldsymbol{G}_{k-1}\boldsymbol{w}_{k-1}, \\ \boldsymbol{y}_k = \boldsymbol{H}_k\boldsymbol{F}_{k-1}\boldsymbol{x}_{k-1} + \boldsymbol{H}_k\boldsymbol{G}_{k-1}\boldsymbol{w}_{k-1} + \boldsymbol{v}_k, \end{cases}$$

namely,

$$\begin{aligned} \boldsymbol{z}_k &= \begin{bmatrix} \boldsymbol{x}_k \\ \boldsymbol{y}_k \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{F}_{k-1} \\ \boldsymbol{H}_k\boldsymbol{F}_{k-1} \end{bmatrix} \boldsymbol{x}_{k-1} + \begin{bmatrix} \boldsymbol{G}_{k-1} & \boldsymbol{0} \\ \boldsymbol{H}_k\boldsymbol{G}_{k-1} & \boldsymbol{1} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_{k-1} \\ \boldsymbol{v}_k \end{bmatrix}. \end{aligned}$$

Since $\boldsymbol{w}_{k-1}$ and $\boldsymbol{v}_k$ are mutually independent and Gaussian, the augmented vector $[\boldsymbol{w}_{k-1}^T, \boldsymbol{v}_k^T]^T$ is jointly Gaussian with a mean vector of $[\boldsymbol{0}^T, \boldsymbol{0}^T]^T$ and covariance of

$$\mathbb{E} \begin{bmatrix} \boldsymbol{w}_{k-1} \\ \boldsymbol{v}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_{k-1} \\ \boldsymbol{v}_k \end{bmatrix}^T = \begin{bmatrix} \boldsymbol{Q}_{k-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_k \end{bmatrix}.$$

Therefore, given $\boldsymbol{x}_{k-1}$, $\boldsymbol{z}_k$ is jointly Gaussian with mean of

$$\begin{bmatrix} \boldsymbol{F}_{k-1} \\ \boldsymbol{H}_k\boldsymbol{F}_{k-1} \end{bmatrix} \boldsymbol{x}_{k-1} + \begin{bmatrix} \boldsymbol{G}_{k-1} & \boldsymbol{0} \\ \boldsymbol{H}_k\boldsymbol{G}_{k-1} & \boldsymbol{1} \end{bmatrix} \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix},$$

and covariance of

$$\begin{aligned} & \boldsymbol{M} \begin{bmatrix} \boldsymbol{Q}_{k-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_k \end{bmatrix} \boldsymbol{M}^T \\ &= \boldsymbol{M} \begin{bmatrix} \boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_k^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_k^{\frac{1}{2}} \end{bmatrix}^T \boldsymbol{M}^T \\ &= \boldsymbol{M} \begin{bmatrix} \boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_k^{\frac{1}{2}} \end{bmatrix} \left(\boldsymbol{M} \begin{bmatrix} \boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_k^{\frac{1}{2}} \end{bmatrix} \right)^T \end{aligned}$$

where

$$\boldsymbol{M} := \begin{bmatrix} \boldsymbol{G}_{k-1} & \boldsymbol{0} \\ \boldsymbol{H}_k \boldsymbol{G}_{k-1} & \boldsymbol{1} \end{bmatrix}.$$

In summary,

$$\mathbb{P}_k(\boldsymbol{z}_k \mid \boldsymbol{x}_{k-1}) = \mathcal{N}_{n+m}\left(\begin{bmatrix} \boldsymbol{F}_{k-1} \\ \boldsymbol{H}_k \boldsymbol{F}_{k-1} \end{bmatrix} \boldsymbol{x}_{k-1}, \boldsymbol{\Sigma}_k^\circ \right)$$

where

$$\boldsymbol{\Sigma}_k^\circ = \begin{bmatrix} \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{0} \\ \boldsymbol{H}_k \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{R}_k^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{0} \\ \boldsymbol{H}_k \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{R}_k^{\frac{1}{2}} \end{bmatrix}^T,$$

which is (15).

Appendix D
Derive (27)

By definition, $\boldsymbol{S}_{xx,k} := \langle \boldsymbol{x}_k - \boldsymbol{c}_{x,k}, \boldsymbol{x}_k - \boldsymbol{c}_{x,k} \rangle = \mathbb{E}(\boldsymbol{x}_k - \boldsymbol{c}_{x,k})(\boldsymbol{x}_k - \boldsymbol{c}_{x,k})^T = \mathbb{E}(\boldsymbol{x}_k)(\boldsymbol{x}_k)^T - \mathbb{E}(\boldsymbol{x}_k)(\boldsymbol{c}_{x,k})^T - \boldsymbol{c}_{x,k}\mathbb{E}(\boldsymbol{x}_k)^T + (\boldsymbol{c}_{x,k})(\boldsymbol{c}_{x,k})^T = \mathbb{E}(\boldsymbol{x}_k)(\boldsymbol{x}_k)^T - (\boldsymbol{c}_{x,k})(\boldsymbol{c}_{x,k})^T = \langle \boldsymbol{x}_k, \boldsymbol{x}_k \rangle - \boldsymbol{c}_{x,k}\boldsymbol{c}_{x,k}^T$. Hence, $\langle \boldsymbol{x}_k, \boldsymbol{x}_k \rangle = \boldsymbol{S}_{xx,k} + \boldsymbol{c}_{x,k}\boldsymbol{c}_{x,k}^T$. Similarly, we have $\langle \boldsymbol{x}_k, \boldsymbol{y}_k \rangle = \boldsymbol{S}_{xy,k} + \boldsymbol{c}_{x,k}\boldsymbol{c}_{y,k}^T$, $\langle \boldsymbol{y}_k, \boldsymbol{x}_k \rangle = \boldsymbol{S}_{yx,k} + \boldsymbol{c}_{y,k}\boldsymbol{c}_{x,k}^T$, and $\langle \boldsymbol{y}_k, \boldsymbol{y}_k \rangle = \boldsymbol{S}_{yy,k} + \boldsymbol{c}_{y,k}\boldsymbol{c}_{y,k}^T$. As a result, we have (53). Applying the trace operator on both sides of (53) gives (27). Note that $\mathrm{Tr}[\boldsymbol{A}_k(\boldsymbol{S}_{yy,k} + \boldsymbol{c}_{y,k}\boldsymbol{c}_{y,k}^T)\boldsymbol{A}_k^T] = \mathrm{Tr}[\boldsymbol{A}_k^T\boldsymbol{A}_k(\boldsymbol{S}_{yy,k} + \boldsymbol{c}_{y,k}\boldsymbol{c}_{y,k}^T)]$.

$$\begin{aligned} & \langle \boldsymbol{x}_k - (\boldsymbol{A}_k\boldsymbol{y}_k + \boldsymbol{b}_k), \boldsymbol{x}_k - (\boldsymbol{A}_k\boldsymbol{y}_k + \boldsymbol{b}_k) \rangle \\ = & \langle \boldsymbol{x}_k, \boldsymbol{x}_k \rangle - \langle \boldsymbol{A}_k\boldsymbol{y}_k + \boldsymbol{b}_k, \boldsymbol{x}_k \rangle - \langle \boldsymbol{x}_k, \boldsymbol{A}_k\boldsymbol{y}_k + \boldsymbol{b}_k \rangle \\ & + \langle \boldsymbol{A}_k\boldsymbol{y}_k + \boldsymbol{b}_k, \boldsymbol{A}_k\boldsymbol{y}_k + \boldsymbol{b}_k \rangle \\ = & \langle \boldsymbol{x}_k, \boldsymbol{x}_k \rangle - \boldsymbol{A}_k\langle \boldsymbol{y}_k, \boldsymbol{x}_k \rangle - \langle \boldsymbol{b}_k, \boldsymbol{x}_k \rangle - \langle \boldsymbol{x}_k, \boldsymbol{y}_k \rangle\boldsymbol{A}_k^T \\ & - \langle \boldsymbol{x}_k, \boldsymbol{b}_k \rangle + \boldsymbol{A}_k\langle \boldsymbol{y}_k, \boldsymbol{y}_k \rangle\boldsymbol{A}_k^T + \boldsymbol{A}_k\langle \boldsymbol{y}_k, \boldsymbol{b}_k \rangle \\ & + \langle \boldsymbol{b}_k, \boldsymbol{y}_k \rangle\boldsymbol{A}_k^T + \langle \boldsymbol{b}_k, \boldsymbol{b}_k \rangle \\ = & \langle \boldsymbol{x}_k, \boldsymbol{x}_k \rangle + \boldsymbol{A}_k\langle \boldsymbol{y}_k, \boldsymbol{y}_k \rangle\boldsymbol{A}_k^T - \langle \boldsymbol{x}_k, \boldsymbol{y}_k \rangle\boldsymbol{A}_k^T - \boldsymbol{A}_k\langle \boldsymbol{y}_k, \boldsymbol{x}_k \rangle \\ & - \langle \boldsymbol{b}_k, \boldsymbol{x}_k \rangle - \langle \boldsymbol{x}_k, \boldsymbol{b}_k \rangle + \boldsymbol{A}_k\langle \boldsymbol{y}_k, \boldsymbol{b}_k \rangle \\ & + \langle \boldsymbol{b}_k, \boldsymbol{y}_k \rangle\boldsymbol{A}_k^T + \langle \boldsymbol{b}_k, \boldsymbol{b}_k \rangle \\ = & \langle \boldsymbol{x}_k, \boldsymbol{x}_k \rangle + \boldsymbol{A}_k\langle \boldsymbol{y}_k, \boldsymbol{y}_k \rangle\boldsymbol{A}_k^T - \langle \boldsymbol{x}_k, \boldsymbol{y}_k \rangle\boldsymbol{A}_k^T - \boldsymbol{A}_k\langle \boldsymbol{y}_k, \boldsymbol{x}_k \rangle \\ & + \langle \boldsymbol{b}_k, \boldsymbol{A}_k\boldsymbol{y}_k - \boldsymbol{x}_k \rangle + \langle \boldsymbol{A}_k\boldsymbol{y}_k - \boldsymbol{x}_k, \boldsymbol{b}_k \rangle + \langle \boldsymbol{b}_k, \boldsymbol{b}_k \rangle \\ = & (\boldsymbol{S}_{xx,k} + \boldsymbol{c}_{x,k}\boldsymbol{c}_{x,k}^T) + \boldsymbol{A}_k(\boldsymbol{S}_{yy,k} + \boldsymbol{c}_{y,k}\boldsymbol{c}_{y,k}^T)\boldsymbol{A}_k^T \\ & - (\boldsymbol{S}_{xy,k} + \boldsymbol{c}_{x,k}\boldsymbol{c}_{y,k}^T)\boldsymbol{A}_k^T - \boldsymbol{A}_k(\boldsymbol{S}_{yx,k} + \boldsymbol{c}_{y,k}\boldsymbol{c}_{x,k}^T) \\ & + 2(\boldsymbol{A}_k\boldsymbol{c}_{y,k} - \boldsymbol{c}_{x,k})\boldsymbol{b}_k^T + \langle \boldsymbol{b}_k, \boldsymbol{b}_k \rangle. \end{aligned} \tag{53}$$

References

[1] B. D. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1979.

[2] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Englewood Cliffs, NJ, USA: Prentice Hall, 2000.

[3] D. Simon, *Optimal State Estimation: Kalman, H_∞, and Nonlinear Approaches*. Hoboken, NJ, USA: Wiley, 2006.

[4] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, pp. 35–45, 1960.

[5] P. Tichavsky, C. H. Muravchik, and A. Nehorai, "Posterior Cramér-Rao bounds for discrete-time nonlinear filtering," *IEEE Trans. Signal Process.*, vol. 46, no. 5, pp. 1386–1396, May 1998.

[6] 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.

[7] Y. Ho and R. Lee, "A Bayesian approach to problems in stochastic estimation and control," *IEEE Trans. Autom. Control*, vol. 9, no. 4, pp. 333–339, Oct. 1964.

[8] B. Hassibi, A. H. Sayed, and T. Kailath, "Linear estimation in Krein spaces. I. Theory," *IEEE Trans. Autom. Control*, vol. 41, no. 1, pp. 18–33, Jan. 1996.

[9] 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.

[10] A. Pourkabirian and M. H. Anisi, "Robust channel estimation in multiuser downlink 5G systems under channel uncertainties," *IEEE Trans. Mobile Comput.*, to be published, doi: 10.1109/TMC.2021.3084398.

[11] Y. Liang, D. Zhou, L. Zhang, and Q. Pan, "Adaptive filtering for stochastic systems with generalized disturbance inputs," *IEEE Signal Process. Lett.*, vol. 15, pp. 645–648, 2008, doi: 10.1109/LSP.2008.2002707.

[12] L. Han, Z. Ren, and D. S. Bernstein, "Maneuvering target tracking using retrospective-cost input estimation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 5, pp. 2495–2503, Oct. 2016.

[13] L. Hu, Z. Wang, Q.-L. Han, and X. Liu, "State estimation under false data injection attacks: Security analysis and system protection," *Automatica*, vol. 87, pp. 176–183, 2018.

[14] K. Dedecius and O. Tichỳ, "Collaborative sequential state estimation under unknown heterogeneous noise covariance matrices," *IEEE Trans. Signal Process.*, vol. 68, pp. 5365–5378, 2020, doi: 10.1109/TSP.2020.3023823.

[15] S. Wang, Z. Wu, and A. Lim, "Denoising, outlier/dropout correction, and sensor selection in range-based positioning," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021, Art no. 1007613, doi: 10.1109/TIM.2021.3078537.

[16] G. Chen, *Approximate Kalman Filtering*, vol. 2. Singapore: World Scientific, 1993.

[17] C. Masreliez and R. Martin, "Robust Bayesian estimation for the linear model and robustifying the Kalman filter," *IEEE Trans. Autom. Control*, vol. 22, no. 3, pp. 361–371, Jun. 1977.

[18] G. Agamennoni, J. I. Nieto, and E. M. Nebot, "Approximate inference in state-space models with heavy-tailed noise," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5024–5037, Oct. 2012.

[19] Y. Huang, Y. Zhang, Y. Zhao, and J. A. Chambers, "A novel robust Gaussian-student's t mixture distribution based Kalman filter," *IEEE Trans. Signal Process.*, vol. 67, no. 13, pp. 3606–3620, Jul. 2019.

[20] T. J. Tarn and J. Zaborszky, "A practical nondiverging filter," *AIAA J.*, vol. 8, no. 6, pp. 1127–1133, 1970.

[21] G. Gawrys and V. Vandelinde, "On the steady-state error of the fading memory filter," *IEEE Trans. Autom. Control*, vol. 21, no. 4, pp. 624–625, Aug. 1976.

[22] Q. Xia, M. Rao, Y. Ying, and X. Shen, "Adaptive fading Kalman filter with an application," *Automatica*, vol. 30, no. 8, pp. 1333–1338, 1994.

[23] R. Mehra, "On the identification of variances and adaptive Kalman filtering," *IEEE Trans. Autom. Control*, vol. 15, no. 2, pp. 175–184, Apr. 1970.

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

[25] A. Mohamed and K. Schwarz, "Adaptive Kalman filtering for INS/GPS," *J. Geodesy*, vol. 73, no. 4, pp. 193–203, 1999.

[26] E. Mazor, A. Averbuch, Y. Bar-Shalom, and J. Dayan, "Interacting multiple model methods in target tracking: A survey," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 34, no. 1, pp. 103–123, Jan. 1998.

[27] Y. Ma, S. Zhao, and B. Huang, "Multiple-model state estimation based on variational Bayesian inference," *IEEE Trans. Autom. Control*, vol. 64, no. 4, pp. 1679–1685, Apr. 2019.

[28] B. Hassibi, A. H. Sayed, and T. Kailath, "Linear estimation in Krein spaces. II. Applications," *IEEE Trans. Autom. Control*, vol. 41, no. 1, pp. 34–49, Jan. 1996.

[29] Y. S. Shmaliy, F. Lehmann, S. Zhao, and C. K. Ahn, "Comparing robustness of the Kalman, H_∞, and UFIR filters," *IEEE Trans. Signal Process.*, vol. 66, no. 13, pp. 3447–3458, Jul. 2018.

[30] S. Kosanam and D. J. Simon, "Kalman filtering with uncertain noise covariances," in *Intelligent Systems and Control*. ACTA Press, 2004, pp. 375–379.

[31] M. Salvoldi and D. Choukroun, "Process noise covariance design in Kalman filtering via bounds optimization," *IEEE Trans. Autom. Control*, vol. 64, no. 2, pp. 834–840, Feb. 2019.

[32] X. Shen and L. Deng, "Game theory approach to discrete H_∞ filter design," *IEEE Trans. Signal Process.*, vol. 45, no. 4, pp. 1092–1095, Apr. 1997.

[33] D. Bertsekas and I. Rhodes, "Recursive state estimation for a set-membership description of uncertainty," *IEEE Trans. Autom. Control*, vol. AC-16, no. 2, pp. 117–128, Apr. 1971.

[34] J. Speyer, J. Deyst, and D. Jacobson, "Optimization of stochastic linear systems with additive measurement and process noise using exponential performance criteria," *IEEE Trans. Autom. Control*, vol. AC-19, no. 4, pp. 358–366, Aug. 1974.

[35] M. Zorzi, "Robust Kalman filtering under model perturbations," *IEEE Trans. Autom. Control*, vol. 62, no. 6, pp. 2902–2907, Jun. 2017.

[36] Y. Qin, Y. Liang, Y. Yang, Q. Pan, and F. Yang, "Minimum upper-bound filter of Markovian jump linear systems with generalized unknown disturbances," *Automatica*, vol. 73, pp. 56–63, 2016.

[37] L. Xie, Y. C. Soh, and C. E. De Souza, "Robust Kalman filtering for uncertain discrete-time systems," *IEEE Trans. Autom. Control*, vol. 39, no. 6, pp. 1310–1314, Jun. 1994.

[38] W. M. Haddad, D. S. Bernstein, and D. Mustafa, "Mixed-norm H_2/H_∞ regulation and estimation: The discrete-time case," *Syst. Control Lett.*, vol. 16, no. 4, pp. 235–247, 1991.

[39] Y. Hung and F. Yang, "Robust H_∞ filtering with error variance constraints for discrete time-varying systems with uncertainty," *Automatica*, vol. 39, no. 7, pp. 1185–1194, 2003.

[40] C. E. de Souza, K. A. Barbosa, and A. T. Neto, "Robust H_∞ filtering for discrete-time linear systems with uncertain time-varying parameters," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2110–2118, Jun. 2006.

[41] X.-H. Chang, J. H. Park, and Z. Tang, "New approach to H_∞ filtering for discrete-time systems with polytopic uncertainties," *Signal Process.*, vol. 113, pp. 147–158, 2015.

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

[43] J. George, "A robust estimator for stochastic systems under unknown persistent excitation," *Automatica*, vol. 63, pp. 156–161, 2016.

[44] S. Z. Yong, M. Zhu, and E. Frazzoli, "A unified filter for simultaneous input and state estimation of linear discrete-time stochastic systems," *Automatica*, vol. 63, pp. 321–329, 2016.

[45] S. Wang, C. Li, and A. Lim, "Optimal joint estimation and identification theorem to linear Gaussian system with unknown inputs," *Signal Process.*, vol. 161, pp. 268–288, 2019.

[46] V. Stojanovic, S. He, and B. Zhang, "State and parameter joint estimation of linear stochastic systems in presence of faults and non-Gaussian noises," *Int. J. Robust Nonlinear Control*, vol. 30, no. 16, pp. 6683–6700, 2020.

[47] P. Cheng, M. Chen, V. Stojanovic, and S. He, "Asynchronous fault detection filtering for piecewise homogenous Markov jump linear systems via a dual hidden Markov model," *Mech. Syst. Signal Process.*, vol. 151, 2021, Art. no. 107353.

[48] U. Shaked, L. Xie, and Y. C. Soh, "New approaches to robust minimum variance filter design," *IEEE Trans. Signal Process.*, vol. 49, no. 11, pp. 2620–2629, Nov. 2001.

[49] W. Liu and P. Shi, "Convergence of optimal linear estimator with multiplicative and time-correlated additive measurement noises," *IEEE Trans. Autom. Control*, vol. 64, no. 5, pp. 2190–2197, May 2019.

[50] A. Subramanian and A. H. Sayed, "A robust minimum-variance filter for time varying uncertain discrete-time systems," in *Proc. Amer. Control Conf.*, 2003, vol. 3, pp. 1885–1889.

[51] H. Xu and S. Mannor, "A Kalman filter design based on the performance/robustness tradeoff," *IEEE Trans. Autom. Control*, vol. 54, no. 5, pp. 1171–1175, May 2009.

[52] J. Y. Ishihara, M. H. Terra, and J. P. Cerri, "Optimal robust filtering for systems subject to uncertainties," *Automatica*, vol. 52, pp. 111–117, 2015.

[53] D. A. Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions*. Hoboken, NJ, USA: Wiley, 1954.

[54] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 1964.

[55] D. Bertsimas, M. Sim, and M. Zhang, "Adaptive distributionally robust optimization," *Manage. Sci.*, vol. 65, no. 2, pp. 604–618, 2019.

[56] M. Staib and S. Jegelka, "Distributionally robust optimization and generalization in Kernel methods," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 9134–9144.

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

[58] 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.

[59] A. Ben-Tal, D. D. Hertog, A. D. 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] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," *Informs Tut. Operations Res.*, pp. 130–166, 2019.

[61] A. Cichocki and S.-I. Amari, "Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, 2010.

[62] J.-C. Bertein, R. Ceschi, and J.-C. Bertein, *Discrete Stochastic Processes and Optimal Filtering*. Hoboken, NJ, USA: Wiley, 2007.

[63] B. C. Levy and R. Nikoukhah, "Robust state space filtering under incremental model perturbations subject to a relative entropy tolerance," *IEEE Trans. Autom. Control*, vol. 58, no. 3, pp. 682–695, Mar. 2013.

[64] 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.

[65] 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. (Cat. No. 00EX373)*, 2000, pp. 153–158.

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

[67] G. Li and Y. Gu, "Restricted isometry property of Gaussian random projection for finite set of subspaces," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1705–1720, Apr. 2018.

[68] S. S. Abadeh, V. A. Nguyen, D. Kuhn, and P. M. M. Esfahani, "Wasserstein distributionally robust Kalman filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8474–8483.

Shixiong Wang (Graduate Student Member, IEEE) received the B.Eng. degree in detection, guidance, and control technology and the M.Eng. degree in systems and control engineering from the School of Electronics and Information, Northwestern Polytechnical University, China, in 2016 and 2018, respectively.

He is currently working toward the Ph.D. degree with the Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore.

His research interests include statistics and optimization theories with applications in signal processing (especially optimal estimation theory), and control technology.

Zhongming Wu received the Ph.D. degree in management science and engineering from the School of Economics and Management, Southeast University, in 2019.

He is an Associate Professor with the School of Management Science and Engineering, Nanjing University of Information Science and Technology.

His research interests include optimization theories and applications.

Andrew Lim received the B.Sc. and Ph.D. degrees in computer and information sciences from the University of Minnesota, Minneapolis, MN, USA, in 1987 and 1992, respectively.

He is currently a Project PI with the School of Computing and Artificial Intelligence, Southwest Jiaotong University, China, and the Chief Scientist with Red Jasper Holdings, Singapore. He held Professorships with the National University of Singapore, Hong Kong University of Science and Technology, and City University of Hong Kong.

His research interests include Big Data analytics, digital twins, digital transformation, demand generation, and supply management problems in the domains of healthcare, logistics, and transportation.