Distributionally Robust State Estimation

by
WANG Shixiong
(M.Eng., Northwestern Polytechnical University)
(B.Eng., Northwestern Polytechnical University)

A Thesis Submitted for the Degree of
Doctor of Philosophy

Under the Advice of
Associate Professor YE Zhisheng

Examined by
Associate Professor CHEN Nan
Assistant Professor LI Xiaobo

Department of Industrial Systems Engineering and Management
National University of Singapore

December 2021

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

WANG Shixiong
December 2021

Acknowledgments

First and foremost, my sincere gratitude should be given to my supervisors Associate Professor Zhisheng YE and Professor Andrew LIM. They gave me this valuable opportunity working as a Ph.D. student, motivated and instructed my research, and imparted their precious academic knowledge and experiences to me. Without their considerate help during my Ph.D. candidature, I cannot smoothly start my academic career and finish this thesis. Together with, I want to thank the National University of Singapore for providing me an internationalized and high-level platform where my understandings and philosophies towards personal life and academic research have been becoming complete and mature.

Second, I deeply thank Assistant Professor Xiaobo Li, Associate Professor Vincent Tan, and Associate Professor Shuangchi He for their kind encouragements and help in improving the writing and normalizing the notation system of this thesis.

Next, I would like to thank my academic collaborators during my candidature: Dr. Chongshou Li, Dr. Zhongming Wu, Dr. Huangjie Zhao, Miss Yuxin Che, Mr. Xinke Li, and Mr. Zhirui Chen. Through working with those knowledgeable friends, my academic outlook has been greatly expanded and several interesting articles have been published in leading journals.

In addition, I must gratefully mention my fellow friends, teammates, and colleagues for their kind companion, help, discussions, and encouragements when I was pursuing my Ph.D. degree: Dr. Yue Zhao, Dr. Haowei Wang, Dr. Yuming Huang, Dr. Qiuzhuang Sun, Dr. Xiaoyang Li, Dr. Xun Zhang, Dr. Xingchen Liu, Dr. Binbin Pan, Dr. Xiaoyang Wei, Mr. Jinyang Wang, Mr. Kanxin Hu, Miss Ruixue Gu, Mr. Qianhao Cong, Miss Jingwen Li, Mr. Yang Yang, Mr. Weiliang Liu, Mr. Jianxiang Wang, and many others.

In the end, I need to express my deepest appreciation to my families, Mr. Shouzhong Wang (father), Mrs. Meimei Kang (mother), and Mr. Shilin Wang (elder brother), etc., for their unconditional love and sacrifices. I cannot thank them enough for whatever they have given to me.

Sources

This thesis is adapted from the author's articles listed below.[1] Chapter 2 integrates [1] and [2] (specifically, Section 2.2 is based on [1] and Section 2.3 is based on [2]), whereas Chapter 3 exhibits [3].

[1] **Shixiong Wang**, Zhongming Wu, and Andrew Lim, "Robust State Estimation for Linear Systems Under Distributional Uncertainty", *IEEE Transactions on Signal Processing*, vol. 69, pp. 5963–5978, 2021. DOI:10.1109/TSP.2021.3118540.

[2] **Shixiong Wang** and Zhisheng Ye, "Distributionally Robust State Estimation for Linear Systems Subject to Uncertainty and Outlier", *IEEE Transactions on Signal Processing*, vol. 70, pp. 452-467, 2021. DOI:10.1109/TSP.2021.3136804.

[3] **Shixiong Wang**, "Distributionally Robust State Estimation for Nonlinear Systems", Major Revision at *IEEE Transactions on Signal Processing*. First Submitted on 15 Nov 2021. (Current Status: Major Revision.)

All the source codes are available online at GitHub: `https://github.com/Spratm-Asleaf/DRSE-PhD-Thesis`. One may use them to reproduce the experimental results and verify the claims in this thesis.

The slides for the oral defense, which highlight main points in this thesis, can be accessed at GitHub: `https://github.com/Spratm-Asleaf/DRSE-PhD-Thesis`. One may read it for a quick digest.

For any comments, suggestions, discussions, and queries, please contact the author through E-Mail: `s.wang@u.nus.edu` or `wsx.gugo@gmail.com`.

[1] Articles that have been published during the author's Ph.D. candidature but are not directly related to this thesis include:

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

[5] **S. Wang**, Z. Wu, and A. Lim, "Denoising, outlier/dropout correction, and sensor selection in range-based positioning," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[6] **S. Wang**, C. Li, and A. Lim, "A model for non-stationary time series and its applications in filtering and anomaly detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1-11, 2021.

[7] **S. Wang**, Y. Che, H. Zhao, and A. Lim, "Accurate tracking, collision detection, and optimal scheduling of airport ground support equipment," *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 572-584, 2021.

[8] **S. Wang**, C. Li, and A. Lim, "ROPHS: Determine real-time status of a multi-carriage logistics train at airport," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Abstract

Parameter uncertainties and measurement outliers unavoidably exist in real linear systems. Such uncertainties and outliers make the true joint state-measurement distributions (induced by the true system models) deviate from the nominal ones (induced by the nominal system models) so that the performance of the optimal state estimators designed for the nominal models becomes unsatisfactory or even unacceptable in practice. The challenges are to quantitatively describe the uncertainties in the models and the outliers in the measurements, and then robustify the estimators in a right way. This thesis studies a distributionally robust state estimation framework for linear systems subject to parameter uncertainties and measurement outliers. It utilizes a family of distributions near the nominal one to implicitly describe the uncertainties and outliers, and the robust state estimation in the worst case is made over the least-favorable distribution. The advantages of the presented framework include: 1) it only uses a few scalars to parameterize the method and does not require the structural information of uncertainties; 2) it generalizes several classical filters (e.g., fading Kalman filter, risk-sensitive Kalman filter, relative-entropy Kalman filter, outlier-insensitive Kalman filters) into a unified framework. We show that the distributionally robust state estimation problem can be reformulated into a linear/nonlinear semi-definite program and in some special cases it can be analytically solved.

Likewise, uncertainties unavoidably exist in modeling for nonlinear systems: process dynamics, measurement dynamics, and/or noises statistics might be uncertain. As a result, nominally optimal state estimators become deteriorated or even unsatisfactory, and robust filters insensitive to modeling uncertainties have to be designed. Since uncertainties in nominal models make prior state distributions and likelihood distributions uncertain as well, this thesis proposes a distributionally robust particle filtering framework for nonlinear systems under modeling uncertainties. Specifically, we use worst-case prior state distributions (near the nominal prior state distributions) to generate prior state particles and/or determine their weights. Similarly, worst-case likelihood distributions (near the nominal likelihood distributions) are used to evaluate the worst-case likelihoods of prior state particles given measurements. The "worst-case" scenario is quantified by entropy of distributions, and maximum entropy distributions are found in balls centered at nominal distributions with radii defined by statistical similarity measures such as moments-based similarity, Wasserstein distance, and Kullback-Leibler divergence. We prove that Gaussian approximation filters (e.g., unscented/cubature Kalman filter) are distributionally robust in the sense that they use maximum entropy prior state distributions and maximum entropy likelihood distributions. Moreover, we show that the distributionally robust particle filtering framework provides a likelihood evaluation method for general nonlinear measurement dynamics with non-additive and non-multiplicative measurement noises. At last, measurement outlier treatment strategies in the distributionally robust particle filtering framework is discussed.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Background

Research on state estimation for both linear and nonlinear systems is lastingly active in several academic communities such as target tracking [1, 2], power systems [3], reliability engineering [4], geodesy [5], sensor network [6], control and automation (e.g., robotics [7]), and astronautics [8]. State estimation problems aim to estimate unknown and unobservable system states based on known system dynamics and observable system outputs. From the viewpoint of statistics, state estimation problems are statistical inference problems where hidden (i.e., unobservable) quantities are inferred from observable quantities, and the joint distribution of hidden variables and observable variables is defined by linear/nonlinear system dynamics. Some real-world examples are listed below.

1) In robotics and aeronautics, people might be concerned with obtaining the real-time position and velocity of a moving robot/airplane. The position and velocity may not be directly observable for trackers but they can be inferred out from observable information from radars such as real-time distances, pitch angles, and azimuth angles [2].

2) In reliability engineering, people might be interested in estimating the remaining useful life (RUL) of an industrial plant or product. In this case, the RUL is not directly observable but it can be inferred out from observable information from sensors such as degradation data [4].

3) In supply chain management, estimating (or forecasting) the future demand based on the available information might be of high interest. Under this circumstance, the demands in the (short) future are unknown, but they can be inferred out (not exactly but to some extent) by leveraging some observable market signals (e.g., expected weather condition, early order placement by customers) from both the retailer and the supplier [9, 10].

To be specific, we take a motivating and simple example in robotics to formally explain the state estimation problem; see Figure 1.1 for an illustration.

Suppose the position of a moving robot at the discrete time k is $\mathbf{p}_k$, and at the time $k-1$ is

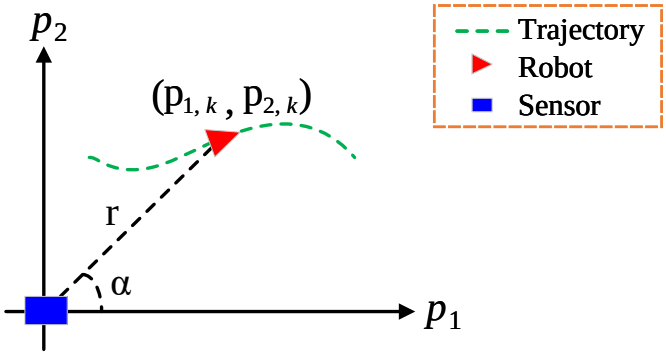


Figure 1.1: A 2-dimensional robot tracking problem. We aim to infer the real-time positions (sometimes and also velocities) of the moving robot based on some observable information from a sensor. The real-time position of the robot at the time k is $\mathbf{p}_k := [\mathrm{p}_{1,k}, \mathrm{p}_{2,k}]^\top$ where $\mathrm{p}_{1,k}$ and $\mathrm{p}_{2,k}$ denote the position in the horizontal coordinate and the vertical coordinate, respectively. At any time k, the exact value of $\mathbf{p}_k$ and the actual trajectory of the robot are unknown to us and the sensor. However, the sensor can capture the noisy value of $\mathbf{p}_k$, or the noisy values of some transforms of $\mathbf{p}_k$. In this example, the sensor is placed at the origin and its position is $[0, 0]^\top$. The distance from the robot to the sensor is termed the range r, while the angle between the line-of-sight and the horizontal line is termed the azimuth α.

$\mathbf{p}_{k-1}$. According to basic kinematics, we have

$$\begin{cases} \mathbf{p}_k &= \mathbf{p}_{k-1} + T\boldsymbol{v}_{k-1} + \dfrac{T^2}{2}\mathbf{a}_{k-1}, \\ \boldsymbol{v}_k &= \boldsymbol{v}_{k-1} + T\mathbf{a}_{k-1}, \end{cases}$$

where T denotes the sampling time between the time instant $k-1$ and the time instant k,[1] $\boldsymbol{v}_{k-1}$ is the average velocity in-between the time instants $k-1$ and k, and $\mathbf{a}_{k-1}$ is the average acceleration during the same time slot; for all k, $\boldsymbol{v}_k$ (resp. $\mathbf{a}_k$) is a 2×1 vector where the first element denotes the velocity (resp. acceleration) in the horizontal axis while the second element the velocity (resp. acceleration) in the vertical axis. However, note that $\mathbf{p}_k$, $\mathbf{p}_{k-1}$, $\boldsymbol{v}_{k-1}$, and $\mathbf{a}_{k-1}$ are all unknown to us, for every k. A compact form can be written as

$$\begin{bmatrix} \mathbf{p}_k \\ \boldsymbol{v}_k \end{bmatrix} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{p}_{k-1} \\ \boldsymbol{v}_{k-1} \end{bmatrix} + \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix} \mathbf{a}_{k-1}.$$

Since we are mainly concerned with inferring $\mathbf{p}_k$ and $\boldsymbol{v}_k$, we may use a random vector $\mathbf{w}_{k-1}$ with assumed-known distribution to model the unknown acceleration $\mathbf{a}_{k-1}$. This gives the

[1] Different sensors may have different data-updating rate. For example, one may capture measurements once per $T = 0.1$ seconds but another may capture measurements once per $T = 0.25$ seconds.

process dynamics equation (also known as the state evolution equation or the state transition equation)

$$\mathbf{x}_k = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \mathbf{x}_{k-1} + \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix} \mathbf{w}_{k-1}, \tag{1.1}$$

where $\mathbf{x}_k := [\mathbf{p}_k^\top, \boldsymbol{v}_k^\top]^\top$ and $\mathbf{x}_k$ is termed the **state** vector; $\mathbf{w}_{k-1}$ is the **process noise** vector.

If the sensor can provide the noisy measurement of $\mathbf{p}_k$ (i.e., the sensor is a positioning device but it has positioning errors), we have

$$\mathbf{y}_k = \mathbf{p}_k + \mathbf{v}_k,$$

where the random vector $\mathbf{y}_k$ is termed the **measurement** vector, and the random vector $\mathbf{v}_k$ is used to model the measurement error of the sensor and is termed the **measurement noise** vector. Hence, the **measurement dynamics** equation (also known as the state measurement equation or the state observation equation) is given as

$$\begin{aligned} \mathbf{y}_k &= \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p}_k \\ \boldsymbol{v}_k \end{bmatrix} + \mathbf{v}_k \\ &= \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x}_k + \mathbf{v}_k. \end{aligned} \tag{1.2}$$

Since both the process dynamics (1.1) and the measurement dynamics (1.2) are of linear forms, the system defined by (1.1) and (1.2) is called a **linear system**. In general, a linear system is compactly given as

$$\begin{cases} \mathbf{x}_k &= \boldsymbol{F}_{k-1}\mathbf{x}_{k-1} + \boldsymbol{G}_{k-1}\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases} \tag{1.3}$$

where $\boldsymbol{F}_{k-1}$, $\boldsymbol{G}_{k-1}$, and $\boldsymbol{H}_k$ are termed the state matrix, the noise-driven matrix, and the observation matrix, respectively. In the contexts of the robot tracking problem above, we specifically have

$$\boldsymbol{F}_{k-1} := \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{G}_{k-1} := \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix}, \quad \text{and} \quad \boldsymbol{H}_k := \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

If, alternatively, the sensor can provide the noisy measurements of the range r_k and the azimuth α_k, we have

$$\begin{cases} \mathrm{r}_k &= \sqrt{\mathrm{p}_{1,k}^2 + \mathrm{p}_{2,k}^2} + \mathrm{v}_{1,k}, \\ \alpha_k &= \arctan\left(\frac{\mathrm{p}_{2,k}}{\mathrm{p}_{1,k}}\right) + \mathrm{v}_{2,k}, \end{cases}$$

where the random variable $\mathrm{v}_{1,k}$ is used to model the ranging error and the random variable $\mathrm{v}_{2,k}$ is used to model the heading error. Let $\mathbf{y}_k := [\mathrm{r}_k, \alpha_k]^\top$ be the measurement vector and $\mathbf{v}_k := [\mathrm{v}_{1,k}, \mathrm{v}_{2,k}]^\top$ the measurement noise vector, the measurement dynamics equation is given as

$$\mathbf{y}_k = \begin{bmatrix} \sqrt{\mathrm{p}_{1,k}^2 + \mathrm{p}_{2,k}^2} \\ \arctan\left(\frac{\mathrm{p}_{2,k}}{\mathrm{p}_{1,k}}\right) \end{bmatrix} + \mathbf{v}_k. \tag{1.4}$$

Since (1.4) is of a nonlinear form, the system defined by the process dynamics (1.1) and the measurement dynamics (1.4) is called a **nonlinear system**. In general, if either (resp. both) the process dynamics equation or (resp. and) the measurement dynamics equation is (resp. are) nonlinear, the system is called a nonlinear system. A generic nonlinear system is compactly given as

$$\begin{cases} \mathbf{x}_k &= \boldsymbol{f}_k(\mathbf{x}_{k-1}, \mathbf{w}_{k-1}), \\ \mathbf{y}_k &= \boldsymbol{h}_k(\mathbf{x}_k, \mathbf{v}_k), \end{cases} \tag{1.5}$$

where $\boldsymbol{f}_k(\cdot,\cdot)$ and $\boldsymbol{h}_k(\cdot,\cdot)$ are termed the **process dynamics** function and the **measurement dynamics** function, respectively. In the contexts of the robot tracking problem above, we specifically have

$$\begin{aligned} \mathbf{x}_k &= \boldsymbol{f}_k(\mathbf{x}_{k-1}, \mathbf{w}_{k-1}) \\ &:= \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \mathbf{x}_{k-1} + \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix} \mathbf{w}_{k-1}, \end{aligned}$$

where $\boldsymbol{f}_k(\cdot,\cdot)$ degenerates to a linear form and

$$\begin{aligned} \mathbf{y}_k &= \boldsymbol{h}_k(\mathbf{x}_k, \mathbf{v}_k) \\ &:= \begin{bmatrix} \sqrt{\mathrm{p}_{1,k}^2 + \mathrm{p}_{2,k}^2} \\ \arctan\left(\frac{\mathrm{p}_{2,k}}{\mathrm{p}_{1,k}}\right) \end{bmatrix} + \mathbf{v}_k, \end{aligned}$$

where $\boldsymbol{h}_k(\cdot,\cdot)$ is of a nonlinear form.

For a linear or nonlinear system, the process dynamics and the measurement dynamics are collectively referred to as the **system dynamics**, which is also known as the **system model**.

As the time proceeds, we can collect the measurements in the past: the measurement set $\mathcal{Y}_k := (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k)$ is available. We aim to find an estimator $\hat{\mathbf{x}}_k$ of $\mathbf{x}_k$, which is a function of the random sequence $\mathcal{Y}_k$, such that the mean square estimation error is minimized;

$$\hat{\mathbf{x}}_k := \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \operatorname{Tr} \mathbb{E}[\boldsymbol{\phi}(\mathcal{Y}_k) - \mathbf{x}_k][\boldsymbol{\phi}(\mathcal{Y}_k) - \mathbf{x}_k]^\top, \tag{1.6}$$

over all Borel-measurable functions $\boldsymbol{\phi}$, where the expectation is taken over the joint distribution of $(\mathbf{x}_k, \mathcal{Y}_k)$.

In this context, the random vector $\boldsymbol{\phi}(\mathcal{Y}_k)$, which is $\sigma(\mathcal{Y}_k)$-measurable,[2] is termed a **state estimator** or a **filter**, and the optimal one is called the optimal state estimator or the optimal filter. When the measurement set $\mathcal{Y}_k$ has one realization $\boldsymbol{Y}_k := (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_k)$, the deterministic value $\hat{\boldsymbol{x}}_k := \boldsymbol{\phi}(\boldsymbol{Y}_k)$ is termed a **state estimate** of $\mathbf{x}_k$, and the optimal one is called the optimal state estimate.

Instead of only providing a single estimate $\hat{\boldsymbol{x}}_k$ for $\mathbf{x}_k$, we may also be interested in finding a probability density function $p_{\mathbf{x}_k|\mathcal{I}_k}(\boldsymbol{x}_k|\mathcal{I}_k)$ conditioned on all the available information $\mathcal{I}_k$ till the time k. Intuitively, the density $p_{\mathbf{x}_k|\mathcal{I}_k}(\boldsymbol{x}_k|\mathcal{I}_k)$ describes the relative likelihood that $\mathbf{x}_k$ takes the specific value $\boldsymbol{x}_k$. This gives the **Bayesian estimation** setting [11, 12].[3] In the state estimation contexts, $\mathcal{I}_k$ explicitly stands for all the past measurements $\mathcal{Y}_k$. Specifically, we aim to find the filtered state distribution (also known as the **posterior state distribution**) of $\mathbf{x}_k$, i.e., $p_{\mathbf{x}_k|\mathcal{Y}_k}(\boldsymbol{x}_k|\mathcal{Y}_k)$, through the Bayes' rule:

$$p_{\mathbf{x}_k|\mathcal{Y}_k}(\boldsymbol{x}_k|\mathcal{Y}_k) \propto p_{\mathbf{y}_k|\mathbf{x}_k=\boldsymbol{x}_k}(\boldsymbol{y}_k|\boldsymbol{x}_k) \cdot p_{\mathbf{x}_k|\mathcal{Y}_{k-1}}(\boldsymbol{x}_k|\mathcal{Y}_{k-1}), \tag{1.7}$$

where $p_{\mathbf{y}_k|\mathbf{x}_k=\boldsymbol{x}_k}(\boldsymbol{y}_k|\boldsymbol{x}_k)$ is termed the conditional measurement distribution given $\boldsymbol{x}_k$ (also known as the likelihood distribution given $\boldsymbol{x}_k$), and $p_{\mathbf{x}_k|\mathcal{Y}_{k-1}}(\boldsymbol{x}_k|\mathcal{Y}_{k-1})$ the predicted state distribution (also known as the **prior state distribution**). When $p_{\mathbf{x}_k|\mathcal{Y}_k}(\boldsymbol{x}_k|\mathcal{Y}_k)$ is available, the optimal estimator that solves (1.6) is the posterior mean [11]:

$$\hat{\mathbf{x}}_k := \int \boldsymbol{x}_k \cdot p_{\mathbf{x}_k|\mathcal{Y}_k}(\boldsymbol{x}_k|\mathcal{Y}_k) \mathrm{d}\boldsymbol{x}_k.$$

Although theoretically attractive, the posterior mean is not always easy to obtain because the posterior density is usually hard to compute [13]; cf. also [14]. Hence, approximation techniques for computing the posterior mean or the posterior density function such as Gaussian filters [13], varitional Bayesian inference [15], and particle filters [16] must be studied. Alternatively, one may also obtain the maximum *a-posteriori* (MAP) estimator of $\mathbf{x}_k$:

$$\hat{\mathbf{x}}_k := \underset{\boldsymbol{x}_k}{\operatorname{argmax}} \; p_{\mathbf{x}_k|\mathcal{Y}_k}(\boldsymbol{x}_k|\mathcal{Y}_k).$$

However, the optimality in the MAP sense is not considered in this thesis, and we only focus on the optimality in the minimum mean square error sense defined in (1.6). This is because the latter is the most popular one in the state estimation community and also in the applied statistics community. If the closed-form expression of $p_{\mathbf{x}_k|\mathcal{Y}_k}(\boldsymbol{x}_k|\mathcal{Y}_k)$ is hard to derive, we may leverage the sequential Monte Carlo method to use samples to approximate involved distributions. This

[2] $\sigma(\mathcal{Y}_k)$ denotes the σ-algebra generated by $\mathcal{Y}_k$.

[3] For differences between Bayesian estimation and Frequentist estimation, see Appendix A.5.

gives the **particle filter** and each sample is termed a **particle** [17, Section 13.3.4].

For the robot tracking problem above, people usually use a random vector $\mathbf{w}_k$ with assumed-known distribution (usually a Gaussian distribution) to model the unknown acceleration $\mathbf{a}_k$ for every k, and $\mathbb{E}\mathbf{w}_k\mathbf{w}_j^\top = \mathbf{0}$ (i.e., uncorrelatedness) for every $k \neq j$ [2]. This Gaussianity and uncorrelatedness assumption is hardly exact in practice because a true robot never maneuvers with a non-smooth trajectory. (Note that under the Gaussianity and uncorrelatedness assumption of acceleration, the expected trajectory of the robot, i.e., the continuous-time stochastic process $\mathbf{p}(t)$ from which the discrete-time process $\{\mathbf{p}_k\}_{k=1,2,...}$ is sampled, is non-smooth.[4]) In addition, the measurement noise $\mathbf{v}_k$ is usually assumed to be Gaussian with mean $\mathbf{0}$ and covariance $\boldsymbol{R}_k$ [18, 19]. However, in reality, the measurement noise $\mathbf{v}_k$ may not exactly follow a Gaussian distribution or the noise statistics are not guaranteed to be exactly the same as $\mathbf{0}$ and $\boldsymbol{R}_k$ [20, 21]. Also, due to clock error, the sensor's true sampling time might be different from the nominal sampling time T so that the nominal state matrix, i.e.,

$$\boldsymbol{F}_k = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}$$

is just an approximation to the true state matrix. Hence, the nominal linear system model (1.1) and (1.2), and the nominal nonlinear system model (1.1) and (1.4) might be inexact, or **uncertain**. More specifically, the uncertainty in modeling $\mathbf{w}_k$ and $\boldsymbol{F}_k$ leads to the uncertainty in the process dynamics (1.1), and the uncertainty in modeling $\mathbf{v}_k$ induces the uncertainty in the measurement dynamics (1.2) and (1.4). Formally, a nominal model $\boldsymbol{O} = \mathcal{M}(\boldsymbol{I})$ is said to be uncertain if it is not guaranteed to be exactly the same as the true governing model $\boldsymbol{O} = \mathcal{M}_0(\boldsymbol{I})$, where $\boldsymbol{O}$ denotes the output and $\boldsymbol{I}$ the input. Other equivalent terms to "uncertain model" that are widely used include "mismatched model", "deviated model", and "perturbed model", etc. Possible cases are as follows.

1) **Parameter Uncertainty**. Suppose the nominal model $\boldsymbol{O} = \mathcal{M}(\boldsymbol{I}; \boldsymbol{\beta})$ is parameterized by $\boldsymbol{\beta}$. If the model type is exact and only the parameter $\boldsymbol{\beta}$ is uncertain, the model uncertainty is reflected by "parameter uncertainty". In the state estimation contexts, a possible example is that the true system model is guaranteed to be linear and the noises are guaranteed to be Gaussian, but we do not exactly know the noise statistics.

2) **Type Uncertainty**. In the state estimation contexts, an example might be the case that the

[4] Let t denote the continuous time. We have an Ito process

$$\begin{bmatrix} \frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} \\ \frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p}(t) \\ \boldsymbol{v}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \mathbf{w}(t),$$

where $\mathbf{w}(t)$ denotes a continuous-time white Gaussian process. The discrete-time system (1.1) is sampled from this continuous-time system.

true system model is nonlinear but we might use a linear nominal model. Another example might be the case that the true system model is known to be the one among candidate models. However, at one time instant, we do not exactly know which candidate model is governing the true plant [22, 23]. In this case, one may also call it "mode uncertainty".

3) **Measurement Outlier**. This is a special case of the type uncertainty. We discuss the measurement outlier problem separately because it is extremely important in practice. If outliers unexpectedly exist in measurements, the true measurement distribution deviates from the nominal measurement distribution. In the linear-system state estimation contexts, a possible example is that the nominal measurement noise model is Gaussian, whereas the true measurement noise model is fat-tailed (e.g., Laplacian, Student's t). In other words, the type of noise distribution is uncertain.

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

1.2 Problem Statements and Methodological Motivations

For linear systems, if process and measurement noises are Gaussian and parameters involved are exact, it is well known that the reputed Kalman filter gives the optimal solution in the sense of the following: (a) the linear unbiased minimum variance estimation [18]; (b) the least/minimum mean square error estimation [24–26]; (c) the regularized least square estimation [27]; (d) the Bayesian *a posteriori* mean estimation of state conditioned on measurement process [11, 25] (also recall Sherman's theorem); (e) the orthogonal projection of the state onto the stochastic Hilbert space spanned by the corresponding innovation process (or equivalently, spanned by the measurement process) generated from the linear system models [24, 28]; or (f) the optimal state estimate that reaches the posterior Cramer-Rao lower bound [26]. However, for many real problems in engineering, the nominal linear system models usually suffer from nontrivial and uncertain modeling errors, e.g., uncertain channel characteristics in wireless communication [29, 30], unknown maneuvers in target tracking [31, 32], uncertain attacks/faults in sensor networks [33], unknown noise statistics of sensors [34], and outliers in ultrawideband (UWB) range measurements [35]. Unfortunately, the Kalman filter is sensitive to possible modeling uncertainties and measurement outliers: they may significantly deteriorate the performance of the Kalman filter [20] or even cause divergence [18, 36]. Hence, uncertainty-aware state estimation solutions for linear systems need to be studied. For nonlinear systems, typical treatment frameworks include: 1) linearization methods, e.g., extended Kalman filters [37] and Takagi–Sugeno fuzzy approximation [38], 2) Gaussian approximation methods including Unscented Kalman filters [39], Cubature Kalman filters [40], and Ensemble Kalman filter (EnKF) [41], etc., and 3) approximated Bayesian inference methods such as variational Bayesian inference [23, 42–44] and sequential Monte Carlo [16, 45–47]. Linearization methods and Gaussian approximation methods are doubted for their incapability of capturing severe nonlinearities, while approximated Bayesian inference methods are criticized for their high computational burdens. However, continuous improvement

in computation powers of modern microprocessors/computers is reducing such criticisms on approximated Bayesian inference methods and encouraging signal processing practitioners to implement these methods for higher estimation accuracy. On this basis, sequential Monte Carlo methods (i.e., particle filters) are of more interest because solving functional optimization problems in variational Bayesian inference is theoretically challenging and therefore additional assumptions, e.g. parameterized function representation and mean field approximation [15], are required. Over the years, tremendous efforts have been made to perfect particle filters, especially in designing efficient sampling and resampling techniques [45, 48–50]. However, virtually all of the past literature assume that the process dynamics and measurement dynamics are accurate. This assumption is suspect because uncertainties are unavoidable in modeling; i.e., nominal models designed by scientists/engineers are not guaranteed to be exactly the same as the true governing models. Such uncertainties may be incurred by oscillating but unknown values of elements in circuits (e.g., resistors/inductors influenced by thermal/electromagnetic noises), by uncontrollable factors in model identification (e.g., numerical errors in parameter estimation; mismatched model assumptions), etc. Therefore, uncertainty-aware particle-based state estimation solutions for nonlinear systems have to be studied.

There are two philosophies in statistics, optimization, and also engineering to handle uncertainties. The one is to reduce such uncertainties by, e.g., jointly estimating the true values of the uncertain factors whenever it is possible [22, 51–54], whereas the other is to tolerate the uncertainties by, e.g., designing robust solutions that are insensitive to them [55–59]. The former is referred to as **adaptive methods**, and the latter is termed **robust methods**. Specifically, in the state estimation literature, adaptive methods include unknown-input Kalman filters [53], adaptive Kalman filters [54], etc., while robust methods contain, e.g., robust Kalman filters [27, 29] and distributionally robust state estimators [57, 58]. Since not all uncertain factors can be correctly characterized, quantitatively modeled, and exactly estimated, sometimes and also generally, robust solutions are attractive. Distributionally robust optimization theory,[5] an offspring of robust statistics and optimization theories, is a mainstream framework dealing with modeling uncertainties. It is currently popular in operations research [62, 63], machine learning [56, 64], systems control [65], to name a few. When some statistical information of uncertain factors are known in prior, distributionally robust optimization methods are preferable over classical robust optimization methods which only take into account possible values that the uncertain factors can take; this is because distributional information can be utilized to counteract conservativeness, to some extent [66].

In this thesis, for linear systems, we study a distributionally robust state estimation framework against parameter uncertainties and measurement outliers. It utilizes a family of distributions near the nominal one to implicitly describe the uncertainties and outliers, and the robust state

[5] The term "distributional" means probability-distribution-related. One should differentiate it with another term "distributed" in engineering literature; cf. [60, 61]. For more information of distributionally robust optimization theory, see Appendix A.1.

estimation in the worst case is made over the least-favorable distribution. Simultaneously, for nonlinear systems, distributionally robust optimization theory is leveraged to robustify particle filters; this is because particle filters are Bayesian statistical methods, and therefore, natural to be discussed in "distributional" contexts. Specifically, when a nominal process dynamics is not guaranteed to be exactly the same as the true one, we argue that the associated nominal prior state distribution, which is represented by weighted particles that are generated from this nominal process dynamics, is different from the true prior state distribution as well. Therefore, we propose to find the worst-case distribution near the nominal prior state distribution, and use this worst-case distribution as a surrogate to generate new prior state particles and/or update their weights. On the other hand, when the measurement dynamics is inexact, the likelihoods of the prior state particles cannot be exactly evaluated either. Likewise, we suggest finding worst-case likelihood distributions for prior state particles to evaluate their worst-case likelihoods at given measurements.

1.3 Contributions

The contributions of this thesis can be summarized as follows.

a) For linear systems,

   1) We propose a distributionally robust state estimation framework against both parameter uncertainties and measurement outliers. It uses a family of distributions to describe the parameter uncertainties and measurement outliers, and the robust state estimation is made over the least-favorable distribution. For details, see Sections 2.2 and 2.3.

   2) We show that the proposed framework generalizes several existing estimation methodologies, including the fading Kalman filter, the Student's t Kalman filter, the risk-sensitive Kalman filter, the M-estimation-based Kalman filters, the relative-entropy Kalman filter, the τ-divergence Kalman filter, and the Wasserstein Kalman filter. For details, see Theorem 12 (and Theorem 10).

   3) We show that the proposed distributionally robust state estimation problem can be reformulated into a linear/nonlinear semi-definite program and in some special cases it can be analytically (i.e., efficiently) solved. For details, see Theorems 1, 2, 6, and 7.

   4) Comprehensive comparisons between the newly proposed distributionally robust estimation framework and state-of-the-art frameworks are made. For details, see Sections 2.2.5 and 2.3.4.

b) For nonlinear systems,

   1) We propose a robustification scheme for particle filters. Specifically, in implementing a particle filter, we use worst-case distribution (i.e., maximum entropy distribution) near the nominal prior state distribution to generate new prior state particles and/or update

their weights, and use worst-case distribution (i.e., maximum entropy distribution) near the nominal likelihood distribution to evaluate the worst-case likelihoods of these prior state particles. For details, see Sections 3.1 and 3.3.

2) We derive maximum entropy distributions in balls centered at nominal distributions with radii defined by the Wasserstein distance and the Kullback-Leibler divergence. For details, see Sections 3.2.2 and 3.2.3, especially Theorems 15, 16, 17, and 18.

3) We show that this robustification scheme serves yet a new resampling strategy against particle degeneracy. In detail, maximum entropy distributions tend to have uniform probability for each support point, and therefore, in particle filter, particles tend to have equal weights. For details, see Section 3.3.1, especially Eqs. (3.38) and (3.39).

4) We show that the proposed robustification scheme offers a universal likelihood evaluation method for prior state particles when measurement dynamics is driven by non-additive and non-multiplicative noises. For details, see Section 3.3.2, especially Methods 4 and 5.

5) We illustrate that Gaussian approximation state estimators are distributionally robust. For details, see Section 3.2.1, especially Corollary 4.

6) We provide a measurement outlier identification and treatment method for particle filters. For details, see Section 3.3.3.

1.4 Notations

Random quantities are denoted by *Roman type* symbols while deterministic quantities are denoted by *Italic type* symbols. We use boldface lowercase symbols for vectors and boldface uppercase symbols for matrices. For example, x denotes a deterministic scalar, $\boldsymbol{x}$ a deterministic vector, and $\boldsymbol{X}$ a deterministic matrix; x denotes a random scalar, and $\mathbf{x}$ denotes a random vector. Suppose $\mathbf{x}$ is a continuous random vector on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbf{x}$ takes value on $\mathbb{R}^n$, where Ω denotes the sample space, $\mathcal{F}$ the σ-algebra on Ω, and $\mathbb{P}$ the probability measure on $(\Omega, \mathcal{F})$. The probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ induced by $\mathbf{x}$, also called the distribution or the law of $\mathbf{x}$, is denoted as $\mathbb{P}_{\mathbf{x}}$, where $\mathcal{B}(\mathbb{R}^n)$ denotes the Borel σ-algebra on $\mathbb{R}^n$. Let $F_{\mathbf{x}}(\cdot)$ and $p_{\mathbf{x}}(\cdot)$ denote the cumulative distribution function (CDF) and the probability density function (PDF) of $\mathbf{x}$, respectively; when it is clear in the context, we suppress the subscript $\mathbf{x}$ for simplicity. Let $\mathbb{E}\mathbf{x}$ denote the expectation of the random vector $\mathbf{x}$. Suppose $\mathbf{y}$ is another continuous random vector on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbf{y}$ takes value on $\mathbb{R}^m$. Let $\mathbb{P}_{\mathbf{x}|\mathbf{y}}$ denote the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$, which is specified by $\mathbb{P}_{\mathbf{x}|\mathbf{y}}(B|\boldsymbol{y}), \forall B \in \mathcal{B}(\mathbb{R}^n), \forall \boldsymbol{y} \in \mathbb{R}^m$, where $\boldsymbol{y}$ is a given realization of $\mathbf{y}$. Let $\mathbb{E}(\mathbf{x}|\mathbf{y})$ denote the conditional expectation of $\mathbf{x}$ given $\mathbf{y}$. Note that $\mathbb{E}(\mathbf{x}|\mathbf{y})$ is a $\sigma(\mathbf{y})$-measurable random vector. However, whenever $\mathbf{y} = \boldsymbol{y}$ is specified, $\mathbb{E}(\mathbf{x}|\boldsymbol{y})$ becomes deterministic. The collection of all probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is denoted as $\mathcal{P}(\mathbb{R}^d)$. Supposing $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathbb{R}^d)$, we use $D(\mathbb{P}, \mathbb{Q})$ to define a possible statistical similarity measure (e.g., Wasserstein distance, Kullback–Leibler divergence) between $\mathbb{P}$ and $\mathbb{Q}$. For every

integer-valued discrete time index k, let $\mathcal{Y}_k := (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k)$ denote a collection of random vectors, and $\boldsymbol{Y}_k := (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_k)$ a possible realization of $\mathcal{Y}_k$ (i.e., a trajectory of the stochastic process $\{\mathbf{y}_k\}_{k=1,2,...}$). Let $\delta_{\boldsymbol{x}_0}(\boldsymbol{x})$ be the Dirac delta function: $\delta_{\boldsymbol{x}_0}(\boldsymbol{x}) = \infty$ if $\boldsymbol{x} = \boldsymbol{x}_0$ and 0 otherwise; $\int \delta_{\boldsymbol{x}_0}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$. For a given integer N, we let $[N]$ denote the running index set and $[N] := \{1, 2, 3, ..., N\}$. Let $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ define the d-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let an arbitrary d-dimensional distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ be denoted as $\mathcal{D}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{I}$ and $\boldsymbol{0}$ denote the identity and the null matrices with appropriate dimensions, respectively. We use $\boldsymbol{M}^\top$ to denote the transpose of the matrix $\boldsymbol{M}$, and $\mathrm{Tr}\left[\boldsymbol{M}\right]$ its trace when $\boldsymbol{M}$ is square. Let $\mathbb{S}^d$ denote the set 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 semi-definite (resp. positive definite) matrices in $\mathbb{S}^d$. If $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}^d$, $\boldsymbol{A} \succeq \boldsymbol{B}$ (resp. $\boldsymbol{A} \succ \boldsymbol{B}$) indicates that $\boldsymbol{A} - \boldsymbol{B} \in \mathbb{S}^d_+$ (resp. $\boldsymbol{A} - \boldsymbol{B} \in \mathbb{S}^d_{++}$). If $\boldsymbol{S} \in \mathbb{S}^d_+$, let $\boldsymbol{S}^{1/2}$ be a square root of $\boldsymbol{S}$ (i.e., $\boldsymbol{S}^{1/2}\boldsymbol{S}^{1/2} = \boldsymbol{S}$). To avoid notational clutter, an ellipsis in a bracket means a copy of the content in the immediately previous bracket (e.g., $[\mathcal{E}][\cdots] := [\mathcal{E}][\mathcal{E}]$ when an expression $\mathcal{E}$ is long). We use $\langle \boldsymbol{A}, \boldsymbol{B} \rangle := \mathrm{Tr}\left[\boldsymbol{A}^\top \boldsymbol{B}\right]$ to denote the trace inner product of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. We use $\|\boldsymbol{a}\| := \sqrt{\boldsymbol{a}^\top \boldsymbol{a}}$ (resp. $\|\boldsymbol{a}\|_{\boldsymbol{W}} := \sqrt{\boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{a}}$) to denote the (resp. weighted) Euclidean norm of $\boldsymbol{a}$. In this thesis, all vectors are column vectors by default.

Chapter 2

State Estimation for Linear Systems

2.1 Problem Formulation

Motivated by (1.3), we are concerned with estimating the hidden (i.e., unobservable) state vector $\mathbf{x}_k$ of a linear Markov system [18, 19, 24]

$$\begin{cases} \mathbf{x}_k &= \boldsymbol{F}_{k-1}\mathbf{x}_{k-1} + \boldsymbol{G}_{k-1}\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases} \tag{2.1}$$

where k is the discrete time index; $\mathbf{x}_k \in \mathbb{R}^n$ is the state vector; $\mathbf{y}_k \in \mathbb{R}^m$ is the measurement vector; $\mathbf{w}_{k-1} \in \mathbb{R}^p$, $\mathbf{v}_k \in \mathbb{R}^m$ are the process noise and measurement noise, respectively. For every k, $\mathbf{x}_k$, $\mathbf{y}_k$, $\mathbf{w}_k$, and $\mathbf{v}_k$ are assumed to have finite second moments: this is a standard assumption in applied statistics to guarantee the existence of an estimator; see, e.g., [67, Chap. 4], [68, Chap. 11].

The nominal system (2.1) defines two discrete-time stochastic vector processes $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$, where $k = 1, 2, \cdots$. Therefore, mathematically, we aim to estimate the unobservable stochastic process $\{\mathbf{x}_k\}$ based on an observable stochastic process $\{\mathbf{y}_k\}$. The more general setting is the state inference problems for hidden Markov processes [12].

Let $\mathcal{H}_{\mathcal{Y}_k}$ denote the collection of all possible linear combinations of $\{\mathbf{1}, \mathcal{Y}_k\}$ and $\mathcal{H}'_{\mathcal{Y}_k}$ denote the collection of all second-moment-finite functions of $\mathcal{Y}_k$. Specifically,

$$\mathcal{H}_{\mathcal{Y}_k} := \left\{ \boldsymbol{B}_k\mathbf{1} + \sum_{i=1}^{k} \boldsymbol{A}_i\mathbf{y}_i \middle| \boldsymbol{B}_k, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_k \in \mathbb{R}^{n\times m} \right\}.$$

Meanwhile,

$$\mathcal{H}'_{\mathcal{Y}_k} := \left\{ \boldsymbol{\phi}(\mathbf{y}_1, ..., \mathbf{y}_k) \middle| \begin{array}{l} \boldsymbol{\phi} : \underbrace{\mathbb{R}^m \times \cdots \times \mathbb{R}^m}_{k} \to \mathbb{R}^n \\ \boldsymbol{\phi} \text{ is Borel-measurable} \\ \displaystyle\int_{\mathbb{R}^{m\times k}} [\boldsymbol{\phi}(\boldsymbol{Y}_k)]^\top [\boldsymbol{\phi}(\boldsymbol{Y}_k)] \mathrm{d}\mathbb{P}_{\mathcal{Y}_k}(\boldsymbol{Y}_k) < \infty \end{array} \right\}.$$

Note that $\boldsymbol{\phi}$ might be nonlinear. Intuitively, $\mathcal{H}_{\mathcal{Y}_k}$ defines all linear estimators whereas $\mathcal{H}'_{\mathcal{Y}_k}$ can offer nonlinear estimators. Note also that $\mathcal{H}_{\mathcal{Y}_k} \subset \mathcal{H}'_{\mathcal{Y}_k}$. One can easily verify that both $\mathcal{H}_{\mathcal{Y}_k}$ and $\mathcal{H}'_{\mathcal{Y}_k}$ are convex because they are closed under linear operations. In addition, both $\mathcal{H}_{\mathcal{Y}_k}$ and $\mathcal{H}'_{\mathcal{Y}_k}$ are stochastic Hilbert spaces where the inner product between $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ is defined by [28, 67]

$$\int \boldsymbol{\phi}_1^\top \boldsymbol{\phi}_2 \mathrm{d}\mathbb{P}_{\mathcal{Y}_k}(\boldsymbol{Y}_k), \quad \forall \boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \in \mathcal{H}_{\mathcal{Y}_k},$$

or

$$\int \boldsymbol{\phi}_1^\top \boldsymbol{\phi}_2 \mathrm{d}\mathbb{P}_{\mathcal{Y}_k}(\boldsymbol{Y}_k), \quad \forall \boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \in \mathcal{H}'_{\mathcal{Y}_k}.$$

If the linear system (2.1) satisfies the following three properties [18, 19, 24]:

P1) $\mathbf{x}_0 \sim \mathcal{N}_n(\bar{\boldsymbol{x}}_0, \boldsymbol{M}_0)$; For every k, $\mathbf{w}_k \sim \mathcal{N}_p(\boldsymbol{\mu}_k^{\boldsymbol{w}}, \boldsymbol{Q}_k)$, and $\mathbf{v}_k \sim \mathcal{N}_m(\boldsymbol{\mu}_k^{\boldsymbol{v}}, \boldsymbol{R}_k)$;

P2) For every $j \neq k$, $\mathbf{w}_k$ and $\mathbf{x}_0$ are uncorrelated, so are $\mathbf{v}_k$ and $\mathbf{x}_0$, $\mathbf{w}_k$ and $\mathbf{w}_j$, and $\mathbf{v}_k$ and $\mathbf{v}_j$ (viz., $\forall j \neq k$, $\mathbb{E}\mathbf{w}_k\mathbf{x}_0^\top = \mathbf{0}$, $\mathbb{E}\mathbf{v}_k\mathbf{x}_0^\top = \mathbf{0}$, $\mathbb{E}\mathbf{w}_k\mathbf{w}_j^\top = \mathbf{0}$, and $\mathbb{E}\mathbf{v}_k\mathbf{v}_j^\top = \mathbf{0}$). For every k, j, $\mathbf{v}_k$ and $\mathbf{w}_j$ are uncorrelated (viz., $\forall k, j$, $\mathbb{E}\mathbf{v}_k\mathbf{w}_j^\top = \mathbf{0}$);

P3) For every k, $\boldsymbol{\mu}_k^{\boldsymbol{w}}$, $\boldsymbol{\mu}_k^{\boldsymbol{v}}$, $\boldsymbol{Q}_k$, $\boldsymbol{R}_k$, $\boldsymbol{F}_{k-1}$, $\boldsymbol{G}_{k-1}$, and $\boldsymbol{H}_k$ are exactly known (typically $\boldsymbol{\mu}_k^{\boldsymbol{w}}$ and $\boldsymbol{\mu}_k^{\boldsymbol{v}}$ are zero-valued),

the canonical Kalman filter gives the optimal state estimate for the linear system (2.1) in the sense of minimum mean square error; for technical details, see Appendix A.2. Briefly speaking, supposing at the time k the nominal joint state-measurement distribution defined by the nominal system model (2.1) is $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$, we would like to solve the following optimization problem[1]

$$\min_{\boldsymbol{\phi} \in \mathcal{H}'_{\mathcal{Y}_k}} \operatorname{Tr} \mathbb{E}[\mathbf{x}_k - \boldsymbol{\phi}(\mathcal{Y}_k)][\mathbf{x}_k - \boldsymbol{\phi}(\mathcal{Y}_k)]^\top, \tag{2.2}$$

where the expectation is taken over $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$ and $\boldsymbol{\phi}(\cdot)$ is referred to as an estimator (the optimal one is called the optimal estimator). The optimal estimator of $\mathbf{x}_k$ in this minimum mean square error sense is $\mathbb{E}(\mathbf{x}_k|\mathcal{Y}_k) \in \mathcal{H}'_{\mathcal{Y}_k}$. In particular, if $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$ is jointly Gaussian, $\mathbb{E}(\mathbf{x}_k|\mathcal{Y}_k)$ is of a linear form, i.e., $\mathbb{E}(\mathbf{x}_k|\mathcal{Y}_k) \in \mathcal{H}_{\mathcal{Y}_k}$. Nice properties (e.g., linearity, Gaussianity) of the nominal system (2.1) produce a beautiful solution to (2.2), i.e., the Kalman filter. However, in general, problem (2.2) is not always easy to solve if the involved distribution $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$ is not Gaussian [69].

As elucidated in Introduction 1.1, the nominal model (2.1) might be uncertain in practice. To be specific, the Kalman's basic assumptions P1)-P3) might be violated individually or in batch form: one of P1), P2), and P3), or any two of them, or all of them, may be breached in practical state estimation problems. The solutions are standard for the case when the assumption P2) is breached, for example, correlated/colored Kalman filters [18, 19]. (However, $\mathbb{E}\mathbf{w}_k\mathbf{x}_0^\top = \mathbf{0}$ and $\mathbb{E}\mathbf{v}_k\mathbf{x}_0^\top = \mathbf{0}$ are always required.) Thus, in this thesis, we consider only model

[1] For more information of matrix-type objective function, see Appendix A.3.

uncertainties when P1) and/or P3) are/is violated for the linear system (2.1). Namely, possible parameter uncertainties and measurement outliers will be taken into particular consideration. The parameter uncertainties mean that the value(s) of $\boldsymbol{\mu}_k^{\boldsymbol{w}}$, $\boldsymbol{\mu}_k^{\boldsymbol{v}}$, $\boldsymbol{Q}_k$, $\boldsymbol{R}_k$, $\boldsymbol{F}_{k-1}$, $\boldsymbol{G}_{k-1}$, and/or $\boldsymbol{H}_k$ might be uncertain [70], while the measurement outliers might be due to non-Gaussian, fat-tailed noise distributions of $\mathbf{w}_k$ and/or $\mathbf{v}_k$ [21, 71].

There is a large body of literature on coping with uncertainties in the parameters and outliers in the measurements, leading to two streams of research. The first stream focuses on parameter uncertainties in state estimation of linear systems. The earliest solutions include the fading (a.k.a. fading-memory) Kalman filter [36, 72], the finite horizon memory filters [73, Section V] especially the UFIR filter [74], the risk-sensitive (a.k.a. exponential-cost) Kalman filter [73, Section IV], [75], the set-valued Kalman filter [76], the $\mathcal{H}_\infty$ filter [73, 77], the adaptive Kalman filter [78–81], and their extensions. Comprehensive reviews and comparisons of these methods can be found in [27, 70, 72, 74, 79]. Later solutions contain the multiple-model methods which handle the case when the system modes are assumed to be multiple [82, 83], and the unknown-input filters designed for systems that have uncertain inputs [53, 84–86]. Later on, robust filters that are insensitive to parameter uncertainties are introduced. They try to minimize/limit the worst-case estimation error and the uncertainties are modeled in different ways. Remarkable frameworks include the Sayed's norm-constrained filter [27], the stochastic-parameter filter [29, 87, 88], the relative-entropy Kalman filter [89], the τ-divergence Kalman filter [90], and the Wasserstein Kalman filter [91]. The second stream of research deals with outlier-insensitive state estimation. The earliest solution is the Gaussian-sum Kalman filter which approximates non-Gaussian noise distribution by a Gaussian sum [20, 92]. In order to lower the computation burden, two categories of methods are introduced afterwards. The first category uses heavy-tailed distributions for the noises which are inherently outlier-aware [21, 93–96]. The second category contains the M-estimation-based[2] Kalman filters [71, 97–99]. They are designed to identify outliers and then take actions to remove/attenuate them, by leveraging various influence functions [55, 100]. A notable extension for M-estimation-based Kalman filtering is introduced in [101], which jointly estimates an unknown-input existing in both the system dynamics and the measurement dynamics. However, two issues exist in literature addressing parameter uncertainties and measurement outliers. First, for state estimation problems under parameter uncertainties, typical robust solutions [27, 29, 87, 88] and adaptive solutions [53, 84–86] require some structural information of uncertainties so that the uncertainties can be gracefully structured and/or parameterized. However, in practice, the information of uncertainties might be scarce, which denies the possibility to implement existing uncertainty-aware filters, e.g., [27, 29, 53]. Second, up to now, there does not exist a robust state estimation method that is able to simultaneously address both parameter uncertainties and measurement outliers, and a unified viewpoint to understand the various existing state estimation methods is lacking. To this end, motivated by the distributionally robust optimization theories, this thesis studies distributionally robust state estimation (DRSE)

[2]For strict definition of "M-estimation-based", see Appendix A.5.

solutions. We will later show that the DRSE framework can fix the two issues aforementioned.

If the underlying system dynamics (i.e., the true model) deviates from the nominal model (2.1), the true joint state-measurement distribution $\mathbb{P}_{\mathbf{x}_k,\mathcal{Y}_k}$ will more or less diverge from the nominal $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$. In this scenario, we aim to find a robust state estimation solution that is insensitive to the deviation. Inspired by the distributionally robust optimization theory, we can write the distributionally robust counterpart of (2.2) as

$$\min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathcal{Y}_k}} \max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x}_k,\mathcal{Y}_k}(\theta)} \operatorname{Tr}\mathbb{E}[\mathbf{x}_k-\boldsymbol{\phi}(\mathcal{Y}_k)][\mathbf{x}_k-\boldsymbol{\phi}(\mathcal{Y}_k)]^\top, \tag{2.3}$$

where the expectation is taken over a possibly true $\mathbb{P}_{\mathbf{x}_k,\mathcal{Y}_k}$ and

$$\mathcal{F}_{\mathbf{x}_k,\mathcal{Y}_k}(\theta) := \left\{\mathbb{P}_{\mathbf{x}_k,\mathcal{Y}_k}\in\mathcal{P}(\mathbb{R}^n\times\mathbb{R}^{m\times k}) \,\middle|\, D(\mathbb{P}_{\mathbf{x}_k,\mathcal{Y}_k},\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k})\le\theta\right\}$$

is the associated ambiguity set constructed around the nominal distribution $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$ with radius of θ. This worst-case optimization problem can be treated as a zero-sum statistical game [102] where the two adversarial players are the statistician who chooses the optimal estimator and nature that chooses the uncertain/hostile distribution (i.e., one tries to lower the cost but the other to improve).

Nevertheless, the state estimation problem is an online (i.e., time-series) problem and the optimal estimator operates along the discrete time in a recursive way as the time proceeds [28]. This is because the measurements $\mathbf{y}_1$, $\mathbf{y}_2$, ..., $\mathbf{y}_k$ arrive in sequence one by one, not in block as $\mathcal{Y}_k$. Hence, at the time k, we expect to handle only one measurement vector $\mathbf{y}_k$ rather than a bulk of measurements $\mathcal{Y}_k$. This also helps to reduce the calculation complexity at each time step. Thus, we instead try to solve a time-incremental [89] (i.e., one-time-step) alternative problem

$$\min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}_k}} \max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)} \operatorname{Tr}\mathbb{E}\left\{[\mathbf{x}_k-\boldsymbol{\phi}(\mathbf{y}_k)][\mathbf{x}_k-\boldsymbol{\phi}(\mathbf{y}_k)]^\top \,\middle|\, \mathcal{Y}_{k-1}\right\}, \tag{2.4}$$

where the expectation is taken over $\mathbb{P}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$ and the ambiguity set

$$\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta) := \left\{\mathbb{P}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}\in\mathcal{P}(\mathbb{R}^n\times\mathbb{R}^m) \,\middle|\, D(\mathbb{P}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}},\bar{\mathbb{P}}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}})\le\theta\right\}$$

contains all possibly true conditional joint state-measurement distribution given the previous measurement sequence $\mathbb{P}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$, and is constructed around the nominal conditional joint state-measurement distribution given the previous measurement sequence $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$. Note that in (2.4), the space of $\boldsymbol{\phi}(\cdot)$ is only defined by $\mathbf{y}_k$ instead of $\mathcal{Y}_k$. In order to solve (2.4), we need to: 1) design proper forms of the associated ambiguity set $\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)$ so that both parameter uncertainties and measurement outliers can be taken into consideration, and 2) find the explicit optimization equivalent(s) of (2.4) so that it can be efficiently solved.

In order to make the problem (2.4) tractable and also follow the typical problem settings of the state estimation literature, we assume that the nominal conditional prior state distribu-

tion $\bar{\mathbb{P}}_{\mathbf{x}_k|\mathcal{Y}_{k-1}}$ and the nominal measurement noise distribution $\bar{\mathbb{P}}_{\mathbf{v}_k}$ are Gaussian. (Therefore, $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$ would be Gaussian as well.) In other words, no matter what the true distributions $\bar{\mathbb{P}}_{\mathbf{x}_k|\mathcal{Y}_{k-1}}$ and $\bar{\mathbb{P}}_{\mathbf{v}_k}$ are, we use Gaussian distributions to approximate them. The Gaussian approximation is popular in the state estimation community to reduce the computational complexity, especially for nonlinear systems. For instance, recall the cubature Kalman filter [40], the unscented Kalman filter [39], the Ensemble Kalman filter [41], etc. Besides, the Gaussian distribution has the following properties, which adapt into our worst-case robust perspective.

1) The Gaussian distribution admits maximum entropy (i.e., maximum degree of indeterminacy) among all distributions with given/fixed first- and second-order moments [103].

2) Concerning a linear measurement system $\mathbf{y} = \boldsymbol{H}\mathbf{x} + \mathbf{v}$, if the state $\mathbf{x}$ is Gaussian, then among all noise distributions with bounded variance for $\mathbf{v}$, the Gaussian minimizes the mutual information between the state $\mathbf{x}$ and the measurement $\mathbf{y}$ (i.e., the dependence between $\mathbf{x}$ and $\mathbf{y}$ is minimized). Namely, the Gaussian noise makes the measurement least informative to estimate the state [69, 104].

3) Concerning the linear measurement system above, if the noise $\mathbf{v}$ is Gaussian, then among all state distributions with bounded variance for $\mathbf{x}$, the Gaussian maximizes the minimum mean square error. Namely, the Gaussian state is most difficult to estimate [69, 105].

The third reason to make the Gaussianity assumption is that the Wasserstein metric and the Kullback–Leibler divergence for Gaussian distributions admit closed-form expressions.

The structure of the ambiguity set $\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)$ significantly matter. In fact, what types of distributions are included in $\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)$ implicitly admits different types of model uncertainties. To be specific, when we handle measurement outliers in the distributionally robust state estimation framework, the ambiguity set $\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)$ must contain fat-tailed distributions for measurements $\mathbf{y}_k$; i.e., $\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)$ cannot be a Gaussian family. In other words, when addressing measurement outliers, any simplification aiming at reducing computational complexity should strictly avoid taking Gaussian assumption. However, when we only handle parameter uncertainties in the linear system model (2.1), we may take Gaussian assumption for prior states and noises, i.e., $\mathbb{P}_{\mathbf{x}_k|\mathcal{Y}_{k-1}}$ and $\mathbb{P}_{\mathbf{v}_k}$, to reduce the computational complexity of the distributionally robust state estimation problem. Additionally, when there exist outliers in measurements, linear estimators, i.e., $\hat{\mathbf{x}}_k = \boldsymbol{b}_k + \boldsymbol{A}_k\mathbf{y}_k$ where $\boldsymbol{b}_k \in \mathbb{R}^n$ and $\boldsymbol{A}_k \in \mathbb{R}^{n\times m}$ are constant, are not admissible. Note that linear estimators are sensitive to measurement outliers: a large error in $\boldsymbol{y}_k$ also leads to a large error in $\hat{\boldsymbol{x}}_k$ because $\boldsymbol{A}_k$ is just a matrix (i.e., a linear operator), which linearly propagates the error contained in $\boldsymbol{y}_k$ to $\hat{\boldsymbol{x}}_k$ (i.e., it cannot attenuate or limit the influence that the error contained in $\boldsymbol{y}_k$ may bring to the state estimate). In contrast, some nonlinear estimators are inherently insensitive to measurement outliers. This point will be explained in detail later; see Theorem 11 and (2.76) for a snapshot. Briefly speaking, there exists a nonlinear function $\boldsymbol{\psi}(\cdot)$, called influence function, to limit the influence that a measurement outlier (i.e., an unexpectedly large value of $\boldsymbol{y}_k$) may bring to the estimator.

2.2 Addressing Parameter Uncertainty

As a special case, we first study the distributionally robust state estimation problem for the linear system (2.1) subject to only parameter uncertainties. In this case, the ambiguity set $\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)$ is not necessarily required to contain heavy-tailed distributions for $\mathbf{y}_k$; i.e., Gaussian assumption is still acceptable. Additionally, linear estimators that are inherently sensitive to measurement outliers are also acceptable. We first give a detailed literature review for this special class of problems.

2.2.1 Supplementary Literature Review

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[3] filters [36, 106, 107], adaptive Kalman-like filters [54, 78, 79], multiple-model Kalman bank filters [82, 83], and finite-horizon-memory Kalman-like filters [73, Section V], [74]. 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 [108, 109], $\mathcal{H}_\infty$ filters [73, 77], set-valued Kalman-like filters [76], risk-sensitive (i.e., exponential-cost) Kalman-like filters [28, 75, 90], guaranteed-cost (i.e., upper-bound [31, 110]) Kalman-like filters [111], 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 [73], or α_k in [31]), which prevents online operations [27]. Extensions to these methods involve making a trade-off between robustness and nominal performance [112, 113] or considering a greater number of general uncertainty types [113–115].

The third-generation methods include unknown-input Kalman-like filters [53, 84–86, 101, 116] and filters for stochastic parametric uncertainties [29, 87, 88]. 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 [29, 87]. 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

[3] For strict definition of "Kalman-like", see Appendix A.5.

G_k in [53] or given Eq. (3) in [29]. Notable extensions include solutions for the case where unknown inputs and measurement outliers exist simultaneously [101], as well as for the case where unknown inputs exist in multiple-model settings [116], etc.

The fourth-generation methods are represented by [27], 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 [27] 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 [27], because they are usually matrices/vectors with many entries to be designed. The extensions of this framework include [117–119], etc.

This thesis studies a new framework that is as general as the third-generation representatives in [29, 53] and the fourth-generation representative in [27]. However, it does not require a filter designer to determine the structure of the modeling uncertainties (e.g., G_k in [53]; $\boldsymbol{F}_{i,k-1}$, $\boldsymbol{G}_{i,k-1}$ in [29]; $\boldsymbol{M}_i$, $\boldsymbol{\Delta}_i$, $\boldsymbol{E}_{f,i}$, and $\boldsymbol{E}_{g,i}$ in [27]), 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 2.2.5.

2.2.2 Distributionally Robust State Estimation

With the distributionally robust estimation model (2.4) on hand, the next steps are 1) to identify the explicit expression of the nominal distribution $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$, 2) to explicitly define a proper form of the ambiguity set $\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)$ around $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$, and 3) to derive tractable reformulation(s) of (2.4) based on $\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)$. We progressively work on the three steps in this subsection.

First, we find the nominal distribution $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$. For notation brevity, let $\mathbf{z}_k := [\mathbf{x}_k^\top, \mathbf{y}_k^\top]^\top$. From (2.1), the nominal distribution conditioned on $\mathbf{x}_{k-1}$ is known as

$$\bar{\mathbb{P}}_{\mathbf{z}_k|\mathbf{x}_{k-1}} = \mathcal{N}_{n+m}\left(\begin{bmatrix} \boldsymbol{F}_{k-1} \\ \boldsymbol{H}_k\boldsymbol{F}_{k-1} \end{bmatrix}\mathbf{x}_{k-1},\ \boldsymbol{\Sigma}_k^\circ\right),^{4} \tag{2.5}$$

[4] This is a random probability measure because $\mathbf{x}_{k-1}$ is random. Nevertheless, whenever $\mathbf{x}_{k-1}$ has a realization, this probability measure becomes deterministic. This random measure is also known as a transition kernel or probability kernel: 1) for every Borel set B on $\mathbb{R}^{n+m}$, $\bar{\mathbb{P}}_{\mathbf{z}_k|\mathbf{x}_{k-1}}(B)$ is a $\sigma(\mathbf{x}_{k-1})$-measurable random variable; 2) for every specified $\boldsymbol{x}_{k-1}$, $\bar{\mathbb{P}}_{\mathbf{z}_k|\mathbf{x}_{k-1}=\boldsymbol{x}_{k-1}}$ is a distribution/law of $\mathbf{z}_k$.

where

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

in which we note that the notation of the square root of a positive semidefinite matrix is $\boldsymbol{Q}^{\frac{1}{2}}(\boldsymbol{Q}^{\frac{1}{2}})^{\top} = \boldsymbol{Q}$. For details of derivation, see Appendix B.1. The extension of $\boldsymbol{\Sigma}_k^{\circ}$ to the case where $\mathbf{w}_{k-1}$ and $\mathbf{v}_k$ are correlated is straightforward. We do not discuss it here. Suppose that the conditional distribution of $\mathbf{x}_{k-1}$ given $\mathcal{Y}_{k-1}$ is

$$\mathbb{P}_{\mathbf{x}_{k-1}|\mathcal{Y}_{k-1}} = \mathcal{D}_n\left(\hat{\mathbf{x}}_{k-1|k-1}, \boldsymbol{P}^*_{k-1|k-1}\right),^{5}$$

where the optimal (robust) estimate of $\mathbf{x}_{k-1}$ is $\hat{\mathbf{x}}_{k-1|k-1}$ and the corresponding estimation error covariance is $\boldsymbol{P}^*_{k-1|k-1}$. Note that the system (2.1) is not guaranteed to be exact so that the distribution $\mathbb{P}_{\mathbf{x}_{k-1}|\mathcal{Y}_{k-1}}$ may not be Gaussian. This is because, for example, if $\boldsymbol{F}_k$ contains a random variable at one entry, even though $\mathbf{x}_k$ and $\mathbf{w}_k$ are white (i.e., mutually independent) Gaussian and $\boldsymbol{G}_k$ is deterministically constant, $\mathbf{x}_{k+1}$ will no longer be Gaussian. However, for simplicity, we may limit our estimation problem within the Gaussian filtering framework [13] (cf. the unscented [39]/cubature [40] Kalman filter for nonlinear system filtering problem). That is, we use a Gaussian distribution $\mathcal{N}_n\left(\hat{\mathbf{x}}_{k-1|k-1}, \boldsymbol{P}^*_{k-1|k-1}\right)$ to approximate $\mathcal{D}_n\left(\hat{\mathbf{x}}_{k-1|k-1}, \boldsymbol{P}^*_{k-1|k-1}\right)$ in the state estimation procedure. By using the nominal system model (2.1), we can obtain the nominal joint state-measurement distribution conditioned on the previous measurements as

$$\bar{\mathbb{P}}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}(B) = \int_{\mathbb{R}^n} \bar{\mathbb{P}}_{\mathbf{z}_k|\mathbf{x}_{k-1}=\boldsymbol{x}_{k-1}}(B) \cdot \mathbb{P}_{\mathbf{x}_{k-1}|\mathcal{Y}_{k-1}}(\mathrm{d}\boldsymbol{x}_{k-1} \mid \mathcal{Y}_{k-1}), \qquad \forall B \in \mathcal{B}(\mathbb{R}^n \times \mathbb{R}^m), \quad (2.6)$$

where $\mathcal{B}(\mathbb{R}^n \times \mathbb{R}^m)$ denotes the Borel σ-algebra on $\mathbb{R}^n \times \mathbb{R}^m$ (n.b., $\mathbf{z}_k$ is a random variable on $\mathbb{R}^n \times \mathbb{R}^m$). Hence, the **time-update step** (i.e., prior estimation step) in the estimation procedure is given as

$$\bar{\mathbb{P}}_{\mathbf{z}_k|\mathcal{Y}_{k-1}} = \mathcal{N}_{n+m}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right),^{6} \quad (2.7)$$

where

$$\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_{x,k} \\ \boldsymbol{\mu}_{y,k} \end{bmatrix} = \begin{bmatrix} \boldsymbol{F}_{k-1} \\ \boldsymbol{H}_k\boldsymbol{F}_{k-1} \end{bmatrix} \hat{\mathbf{x}}_{k-1|k-1} \quad (2.8)$$

---

[5] This is a random probability measure (a.k.a., transition kernel or probability kernel) because $\mathcal{Y}_{k-1}$ is random. However, whenever $\mathcal{Y}_{k-1}$ has a realization, this probability measure becomes deterministic. Strictly speaking, $\hat{\mathbf{x}}_{k-1|k-1}$ and $\boldsymbol{P}^*_{k-1|k-1}$ are random because $\mathcal{Y}_{k-1}$ is random. However, they are non-random in terms of $\mathbf{x}_{k-1}$: when $\mathcal{Y}_{k-1} = \boldsymbol{Y}_{k-1}$ is specified, they become deterministic.

[6] Strictly speaking, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are random because they are conditioned on the random sequence $\mathcal{Y}_{k-1}$. However, whenever we have a realization of $\mathcal{Y}_{k-1}$, e.g., $\boldsymbol{Y}_{k-1}$, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ will become deterministic. Either off or on, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are non-random in terms of $\mathbf{z}_k$, and they are mean and co-variance of the random vector $\mathbf{z}_k$. Hence, we still use Italic fond for $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ to emphasize that they are non-random in terms of $\mathbf{z}_k$.

and

$$\begin{aligned}\boldsymbol{\Sigma}_k = &\begin{bmatrix} \boldsymbol{F}_{k-1} \\ \boldsymbol{H}_k\boldsymbol{F}_{k-1} \end{bmatrix} \boldsymbol{P}^*_{k-1|k-1} \begin{bmatrix} \boldsymbol{F}_{k-1} \\ \boldsymbol{H}_k\boldsymbol{F}_{k-1} \end{bmatrix}^\top + \\ &\begin{bmatrix} \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{0} \\ \boldsymbol{H}_k\boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{R}_k^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{0} \\ \boldsymbol{H}_k\boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}^{\frac{1}{2}} & \boldsymbol{R}_k^{\frac{1}{2}} \end{bmatrix}^\top.\end{aligned} \tag{2.9}$$

Specifically, in (2.8), we let $\boldsymbol{\mu}_{x,k} := \boldsymbol{F}_{k-1}\hat{\mathbf{x}}_{k-1|k-1}$ and $\boldsymbol{\mu}_{y,k} := \boldsymbol{H}_k\boldsymbol{F}_{k-1}\hat{\mathbf{x}}_{k-1|k-1}$, respectively. In state estimation literature, $\boldsymbol{\mu}_{x,k}$ usually writes $\hat{\mathbf{x}}_{k|k-1}$, i.e., the prior state estimate. Meanwhile, the left-top block of $\boldsymbol{\Sigma}_k$ usually writes $\boldsymbol{P}_{k|k-1}$, i.e., the prior state estimation error covariance.

Remark 1. *The time-update step at the time k means the algorithmic step through which the* **prior** *estimate of $\mathbf{x}_k$, i.e., $\hat{\mathbf{x}}_{k|k-1} := \mathbb{E}(\mathbf{x}_k|\mathcal{Y}_{k-1})$, can be obtained* **before** *observing $\mathbf{y}_k$. In contrast, the measurement-update step at the time k means the algorithmic step through which the* **posterior** *estimate of $\mathbf{x}_k$, i.e., $\hat{\mathbf{x}}_{k|k} := \mathbb{E}(\mathbf{x}_k|\mathcal{Y}_k)$, can be obtained* **after** *observing $\mathbf{y}_k$. Details can be found in, e.g., [18, Chapter 5.1].* □

Second, we define the ambiguity set $\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)$ which contains all the possibly true distributions $\mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}$. Using the moment-based ambiguity set proposed in [120], we have

$$\begin{aligned}&\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta_1,\theta_2,\theta_3) \\ &= \left\{ \mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}} \in \mathcal{P}(\mathbb{R}^n\times\mathbb{R}^m) \;\middle|\; \begin{array}{l} \left[\mathbb{E}(\mathbf{z}_k|\mathcal{Y}_{k-1}) - \boldsymbol{\mu}_k\right]^\top \boldsymbol{\Sigma}_k^{-1} \left[\mathbb{E}(\mathbf{z}_k|\mathcal{Y}_{k-1}) - \boldsymbol{\mu}_k\right] \le \theta_3 \\ \mathbb{E}\left[(\mathbf{z}_k-\boldsymbol{\mu}_k)(\mathbf{z}_k-\boldsymbol{\mu}_k)^\top|\mathcal{Y}_{k-1}\right] \preceq \theta_2\boldsymbol{\Sigma}_k \\ \mathbb{E}\left[(\mathbf{z}_k-\boldsymbol{\mu}_k)(\mathbf{z}_k-\boldsymbol{\mu}_k)^\top|\mathcal{Y}_{k-1}\right] \succeq \theta_1\boldsymbol{\Sigma}_k \end{array} \right\}\end{aligned} \tag{2.10}$$

where $\theta_3 \ge 0$ and $\theta_2 \ge 1 \ge \theta_1 \ge 0$. Note that $\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$ is parameterized by three parameters: θ_1, θ_2, and θ_3. Suppose a possibly true distribution has the conditional mean $\boldsymbol{c}_k$ and the conditional co-variance $\boldsymbol{S}_k$, i.e., $\mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}} = \mathcal{D}_{n+m}(\boldsymbol{c}_k, \boldsymbol{S}_k) \in \mathcal{P}(\mathbb{R}^n\times\mathbb{R}^m)$, where $\boldsymbol{c}_k = [\boldsymbol{c}_{x,k}^\top, \boldsymbol{c}_{y,k}^\top]^\top$, $\boldsymbol{c}_{x,k} = \mathbb{E}(\mathbf{x}_k|\mathcal{Y}_{k-1})$, and $\boldsymbol{c}_{y,k} = \mathbb{E}(\mathbf{y}_k|\mathcal{Y}_{k-1})$. Then, $\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$ can be given as

$$\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta_1,\theta_2,\theta_3) = \left\{ \mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}} = \mathcal{D}_{n+m}(\boldsymbol{c}_k,\boldsymbol{S}_k) \;\middle|\; \begin{array}{l} (\boldsymbol{c}_k-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{c}_k-\boldsymbol{\mu}_k) \le \theta_3 \\ \boldsymbol{S}_k + (\boldsymbol{c}_k-\boldsymbol{\mu}_k)(\boldsymbol{c}_k-\boldsymbol{\mu}_k)^\top \preceq \theta_2\boldsymbol{\Sigma}_k \\ \boldsymbol{S}_k + (\boldsymbol{c}_k-\boldsymbol{\mu}_k)(\boldsymbol{c}_k-\boldsymbol{\mu}_k)^\top \succeq \theta_1\boldsymbol{\Sigma}_k \end{array} \right\} \tag{2.11}$$

Eq. (2.11) implies that the possibly true mean $\boldsymbol{c}_k$ lies in a ball centered at $\boldsymbol{\mu}_k$ and the possibly true covariance $\boldsymbol{S}_k$ is linearly bounded by the nominal covariance $\boldsymbol{\Sigma}_k$. This ambiguity set describes the trust level that we have towards the nominal distribution (2.7), and every probability measure inside is parameterized by $\boldsymbol{c}_k$ and $\boldsymbol{S}_k$. The trust level is quantified by θ_3, θ_2, and θ_1. The smaller

θ_3 is and the closer θ_2 and θ_1 are to one, the more trust we have towards the nominal distribution. Note that when $\theta_3 = 0$ and $\theta_2 = \theta_1 = 1$, the ambiguity set contains only the nominal distribution $\bar{\mathbb{P}}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}$ whose mean is $\boldsymbol{\mu}_k$ and covariance is $\boldsymbol{\Sigma}_k$ [cf. (2.7)]. In highlights, the set (2.11) defines a space for distributional model uncertainties (cf. the norm-based model uncertainties in [27]).

Third, we derive tractable reformulation(s) of (2.4). Recall from (A.5) that under the linear estimation case [i.e., the linear estimator is used no matter whether the nominal $\bar{\mathbb{P}}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}$ and the possibly true $\mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}$ are Gaussian or not], the optimal estimator $\boldsymbol{\phi}(\cdot)$ has an affine form, i.e.,

$$\hat{\mathbf{x}}_{k|k} = \boldsymbol{\phi}(\mathbf{y}_k) = \boldsymbol{A}_k\mathbf{y}_k + \boldsymbol{b}_k, \tag{2.12}$$

where $\boldsymbol{A}_k \in \mathbb{R}^{n\times m}$ and $\boldsymbol{b}_k \in \mathbb{R}^n$, and they are to be determined. We have the following theorem.

Theorem 1. *With the optimal estimator (2.12), the distributionally robust state estimation problem (2.4) subject to (2.11) admits von Neumann's min-max theorem (i.e., saddle point theorem):*

$$\begin{aligned} &\min_{\boldsymbol{A}_k,\boldsymbol{b}_k}\max_{\boldsymbol{c}_k,\boldsymbol{S}_k} \operatorname{Tr}\mathbb{E}\left[\mathbf{x}_k - (\boldsymbol{A}_k\mathbf{y}_k + \boldsymbol{b}_k)\right]\left[\mathbf{x}_k - (\boldsymbol{A}_k\mathbf{y}_k + \boldsymbol{b}_k)\right]^\top \\ = \quad &\\ &\max_{\boldsymbol{c}_k,\boldsymbol{S}_k}\min_{\boldsymbol{A}_k,\boldsymbol{b}_k} \operatorname{Tr}\mathbb{E}\left[\mathbf{x}_k - (\boldsymbol{A}_k\mathbf{y}_k + \boldsymbol{b}_k)\right]\left[\mathbf{x}_k - (\boldsymbol{A}_k\mathbf{y}_k + \boldsymbol{b}_k)\right]^\top, \end{aligned} \tag{2.13}$$

where the expectation is taken over $\mathbb{P}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$. In addition, if $\boldsymbol{\Sigma}_k \succ \mathbf{0}$, this optimization problem is equivalent to a nonlinear positive semidefinite program (NSDP)

$$\max_{\boldsymbol{S}_k} \operatorname{Tr}\left[\boldsymbol{S}_{xx,k} - \boldsymbol{S}_{xy,k}\boldsymbol{S}_{yy,k}^{-1}\boldsymbol{S}_{yx,k}\right], \tag{2.14}$$

subject to

$$\begin{cases} \boldsymbol{S}_k \preceq \theta_2\boldsymbol{\Sigma}_k, \\ \boldsymbol{S}_k \succeq \theta_1\boldsymbol{\Sigma}_k, \\ \boldsymbol{S}_k = \begin{bmatrix} \boldsymbol{S}_{xx,k} & \boldsymbol{S}_{xy,k} \\ \boldsymbol{S}_{yx,k} & \boldsymbol{S}_{yy,k} \end{bmatrix} \succ \mathbf{0}, \\ \boldsymbol{S}_{xx,k} \succ \mathbf{0}, \\ \boldsymbol{S}_{yy,k} \succ \mathbf{0}. \end{cases} \tag{2.15}$$

Proof. See Appendix B.2. □

Remark 2. *When there are no uncertainties in (2.1), the ambiguity set $\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta_1,\theta_2,\theta_3)$ defined in (2.11) contains only the nominal distribution $\bar{\mathbb{P}}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}$. 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 (2.4) to the canonical Kalman filter (A.10). Moreover, the worst-case estimation error covariance (2.14) becomes the nominal estimation error covariance (A.13). □

To further lower the number of parameters of the uncertainty set, motivated by the reputed restricted isometry property [121], we may consider an alternative as

$$\begin{cases} \boldsymbol{S}_k \preceq (1+\theta)\boldsymbol{\Sigma}_k, \\ \boldsymbol{S}_k \succeq (1-\theta)\boldsymbol{\Sigma}_k, \\ \boldsymbol{S}_k = \begin{bmatrix} \boldsymbol{S}_{xx,k} & \boldsymbol{S}_{xy,k} \\ \boldsymbol{S}_{yx,k} & \boldsymbol{S}_{yy,k} \end{bmatrix} \succ \boldsymbol{0}, \\ \boldsymbol{S}_{xx,k} \succ \boldsymbol{0}, \\ \boldsymbol{S}_{yy,k} \succ \boldsymbol{0}, \end{cases} \tag{2.16}$$

in which $0 \leq \theta < 1$. However, (2.16) is not equivalent to (2.15).

Theorem 2. *The NSDP (2.14) subject to (2.15) is analytically solved by*

$$\boldsymbol{S}_k^* = \theta_2 \boldsymbol{\Sigma}_k. \tag{2.17}$$

Proof. See Appendix B.4. □

By comparing with (2.9), Theorem 2 implies that the robust filter is achieved by simultaneously inflating state estimation error covariance in the last time step (i.e., $\boldsymbol{P}_{k-1|k-1}^*$), process error covariance (i.e., $\boldsymbol{Q}_{k-1}$), and observation error covariance (i.e., $\boldsymbol{R}_k$), by θ_2. This is intuitively understandable because

1) if there exist model uncertainties in the process dynamics, we are less confident of state evolution equation so that we should improve the process noise. This corresponds to multiplying $\boldsymbol{Q}_{k-1}$ by θ_2;

2) if there exist model uncertainties in the measurement dynamics, we are less confident of state observation equation so that we should improve the measurement noise. This corresponds to multiplying $\boldsymbol{R}_k$ by θ_2;

3) if there exist model uncertainties in both/either the process dynamics and/or the measurement dynamics, we are less confident of state estimates in the past so that we should improve the state estimation error covariance in the past. This corresponds to multiplying $\boldsymbol{P}_{k-1|k-1}^*$ by θ_2.

Since $\theta_2 \geq 1$, via multiplying by θ_2, we can inflate state estimation error covariance in the last time step, process error covariance, and observation error covariance. Through Theorem 2, we admit the estimate at the last time step, the process evolution equation, and the measurement

equation are all uncertain with the same uncertainty level parameter θ_2. Using the same uncertainty level parameter might be questionable. Thus, later in Section 2.3, a generalized case where the three quantities can be inflated by different parameters is discussed; compare Theorem 7 (and also Theorem 12 and Theorem 10) with Theorem 2.

Corollary 1. *By comparing (2.17) with [106], we can conclude that the traditional fading-memory Kalman-like filter is a distributionally robust state estimation solution under moment-based ambiguity.* □

Corollary 2 (**Measurement-Update Step**). *Suppose that $\boldsymbol{S}_k^*$ solves the optimization problem (2.14) and (2.15). By recalling (2.12), (B.3), (B.7), and (B.8), the distributionally robust estimator in the sense of linear minimum mean square estimation error is given as*

$$\begin{aligned}\hat{\mathbf{x}}_{k|k} &= \boldsymbol{\phi}^*(\mathbf{y}_k) = \boldsymbol{A}_k^*\mathbf{y}_k + \boldsymbol{b}_k^* \\ &= \boldsymbol{\mu}_{x,k} + \boldsymbol{S}_{xy,k}^* \cdot (\boldsymbol{S}_{yy,k}^*)^{-1}(\mathbf{y}_k - \boldsymbol{\mu}_{y,k}) \\ &= \boldsymbol{F}_{k-1}\hat{\mathbf{x}}_{k-1|k-1} + \boldsymbol{S}_{xy,k}^* \cdot (\boldsymbol{S}_{yy,k}^*)^{-1}(\mathbf{y}_k - \boldsymbol{H}_k\boldsymbol{F}_{k-1}\hat{\mathbf{x}}_{k-1|k-1}),\end{aligned} \tag{2.18}$$

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

$$\boldsymbol{P}_{k|k}^* = \boldsymbol{S}_{xx,k}^* - \boldsymbol{S}_{xy,k}^*(\boldsymbol{S}_{yy,k}^*)^{-1}\boldsymbol{S}_{yx,k}^*. \tag{2.19}$$

Note that $\boldsymbol{\mu}_{x,k}$ and $\boldsymbol{\mu}_{y,k}$ in (2.18) are defined in (2.8). Moreover, the least-favorable (i.e., worst-case) conditional distribution of $\mathbf{z}_k$ given $\mathcal{Y}_{k-1}$ is $\mathbb{P}^*_{\mathbf{z}_k|\mathcal{Y}_{k-1}} = \mathcal{D}_{n+m}(\boldsymbol{\mu}_k, \boldsymbol{S}_k^*)$ and the worst-case conditional distribution of $\mathbf{x}_k$ given $\mathcal{Y}_k$ is $\mathbb{P}^*_{\mathbf{x}_k|\mathcal{Y}_k} = \mathcal{D}_n\left(\hat{\mathbf{x}}_{k|k}, \boldsymbol{P}_{k|k}^*\right)$; cf. $\mathbb{P}_{\mathbf{x}_{k-1}|\mathcal{Y}_{k-1}} = \mathcal{D}_n\left(\hat{\mathbf{x}}_{k-1|k-1}, \boldsymbol{P}_{k-1|k-1}^*\right)$ in (2.6). In the Gaussian filter framework, we have approximately $\mathbb{P}^*_{\mathbf{z}_k|\mathcal{Y}_{k-1}} = \mathcal{N}_{n+m}(\boldsymbol{\mu}_k, \boldsymbol{S}_k^*)$ and $\mathbb{P}^*_{\mathbf{x}_k|\mathcal{Y}_k} = \mathcal{N}_n\left(\hat{\mathbf{x}}_{k|k}, \boldsymbol{P}_{k|k}^*\right)$. □

Theorem 2 reveals that $\boldsymbol{S}_{xy,k}^* \cdot (\boldsymbol{S}_{yy,k}^*)^{-1}$ equals $\boldsymbol{\Sigma}_{xy,k} \cdot (\boldsymbol{\Sigma}_{yy,k})^{-1}$ so that (2.18) admits

$$\hat{\mathbf{x}}_{k|k} = \boldsymbol{\mu}_{x,k} + \boldsymbol{\Sigma}_{xy,k} \cdot (\boldsymbol{\Sigma}_{yy,k})^{-1}(\mathbf{y}_k - \boldsymbol{\mu}_{y,k}), \tag{2.20}$$

which is in the same form as the optimal estimation under the nominal distribution, i.e., (A.11). This implies that under the moments-based ambiguity set, the optimal robust state estimate is not directly influenced by the worst-case distribution at the current step.

The overall moments-based distributionally robust state estimator to the linear system (2.1) subject to parameter uncertainty is summarized in Algorithm 2.1.

2.2.3 Computational Complexity

From Remark 2, we know that $\boldsymbol{S}_k^* \equiv \boldsymbol{\Sigma}_k$ gives the canonical Kalman filter. Since the moment-based distributionally robust state estimator is solved by $\boldsymbol{S}_k^* = \theta_2\boldsymbol{\Sigma}_k$, where θ_2 is just a scalar

Algorithm 2.1: Moment-Based Distributionally Robust Estimator for Linear Systems Subject to Parameter Uncertainty

Definition: $\hat{\mathbf{x}}_{k|k}$ as the robust state estimator and $\hat{\boldsymbol{x}}_{k|k}$ the robust state estimate; $\boldsymbol{P}^*_{k|k}$ as the state estimation error covariance.

Initialize: $\hat{\boldsymbol{x}}_{0|0}$, $\boldsymbol{P}^*_{0|0}$, θ.

Remark: In (2.18), $\boldsymbol{c}^*_k$ has already been replaced with $\boldsymbol{\mu}_k$; cf. (B.8). In general, θ_1, θ_2 can be independently initialized without θ. By (2.17), the robust state estimation results only depend on θ_2. Therefore, we do not initialize θ_1. When $\mathbf{y}_k$ has a realization $\boldsymbol{y}_k$, the estimator of $\mathbf{x}_k$, i.e., $\hat{\mathbf{x}}_{k|k}$, gives an estimate $\hat{\boldsymbol{x}}_{k|k}$ to $\mathbf{x}_k$.

Input : measurement $\boldsymbol{y}_k$, $k = 1, 2, 3, ...$

$\theta_2 \leftarrow 1 + \theta$. // *See (2.16)*

while *true* **do**

  // *Time-Update Step, i.e., Prior Estimation*

    Use (2.8) and (2.9) to obtain $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$;

  // *Obtain the Worst-Case Scenario*

    Solve (2.14) and (2.15) with (2.17) to obtain $\boldsymbol{S}^*_k$;

  // *Measurement-Update Step, i.e., Posterior Estimation*

    Use (2.18) and (2.19) to obtain $\hat{\boldsymbol{x}}_{k|k}$ and $\boldsymbol{P}^*_{k|k}$;

  // *Next Time Step*

    $k \leftarrow k + 1$;

end

Output : $\hat{\boldsymbol{x}}_{k|k}$

[cf. (2.17)], it has the same order of computational complexity as the canonical Kalman filter. Specifically, at each time instant k, the computational complexity is $\mathcal{O}(n^3)$ because for a state estimation problem, we usually have $n \geq m$ and $n \geq p$; see Remark 3. This means that the moment-based distributionally robust state estimator is computationally as efficient as the canonical Kalman filter.

Remark 3. *We use $\mathcal{O}(t)$ to denote that the number of operations of an algorithmic process is t. First, note that the computational complexity of the matrix multiplication for two matrices $\boldsymbol{M}_{n\times m}$ and $\boldsymbol{M}_{m\times p}$ is $\mathcal{O}(nmp)$ using the definition of matrix multiplication, and of the matrix inverse for a matrix $\boldsymbol{M}_{m\times m}$ is $\mathcal{O}(m^3)$ using the Gauss–Jordan elimination method. (The results can be improved by advanced algorithms, e.g., the Strassen algorithm.) Therefore,*

1) in the time-update step, the computational complexity of (2.8) is $\mathcal{O}[(n+m)\times n+(n+m)]$, and of

(2.9) is $\mathcal{O}[(n+m)\times n\times n+(n+m)\times n\times(n+m)+(n+m)\times(p+m)\times(n+m)+2\times(n+m)\times(n+m)]$*;*

2) in the step of obtaining the worst-case scenario, the computational complexity of (2.17) is $\mathcal{O}[2\times(n+m)\times(n+m)]$*;*

3) in the measurement-update step, the computational complexity of (2.18) is $\mathcal{O}[n^2+m^3+nmm+nm+n]$*, and of (2.19) is* $\mathcal{O}[n^2+m^3+nmm+nmn+n^2]$*.*

Let $r:=\max\{n,m,p\}$*. As a result, the computational complexity of Algorithm 2.1 is asymptotically* $\mathcal{O}(r^3)$*. Since for a usual state estimation problem,* $n\geq p$ *and* $n\geq m$*, the computational complexity of Algorithm 2.1 is* $\mathcal{O}(n^3)$*.* □

2.2.4 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 [90] 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.

Kullback–Leibler Divergence

Suppose $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{Q}_{\mathbf{x}}$ have the same support $\mathcal{S}$. If $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{Q}_{\mathbf{x}}$ are absolutely continuous with respect to the Lebesgue measure and $\mathbb{P}_{\mathbf{x}}$ is absolutely continuous with respect to $\mathbb{Q}_{\mathbf{x}}$, then the Kullback–Leibler divergence (KL-Divergence) of $\mathbb{P}_{\mathbf{x}}$ from $\mathbb{Q}_{\mathbf{x}}$ is defined as

$$\int_{\mathcal{S}}\ln\left(\frac{d\mathbb{P}_{\mathbf{x}}}{d\mathbb{Q}_{\mathbf{x}}}\right)d\mathbb{P}_{\mathbf{x}}=\int_{\mathcal{S}}\ln\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)p(\boldsymbol{x})d\boldsymbol{x}, \tag{2.21}$$

where $d\mathbb{P}_{\mathbf{x}}/d\mathbb{Q}_{\mathbf{x}}$ denotes the Radon-Nikodym derivative.

In this case, the ambiguity set is as (A.2). See also [89]. When we consider the distributionally robust estimation problem (2.4), (A.2) is specified into

$$\mathcal{F}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}(\theta)=\left\{\mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}\in\mathcal{P}(\mathbb{R}^n\times\mathbb{R}^m)\;\middle|\;\mathrm{KL}\left(\mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}\|\bar{\mathbb{P}}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}\right)\leq\theta\right\}. \tag{2.22}$$

In general, if we use the τ-divergence, $\mathrm{KL}(\mathbb{P}\|\bar{\mathbb{P}})\leq\theta$ should be replaced with $D_\tau(\mathbb{P}\|\bar{\mathbb{P}})\leq\theta$, where $D_\tau(\mathbb{P}\|\bar{\mathbb{P}})$ denotes the τ-divergence [90]. Supposing $\mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}$ is also Gaussian, (2.22) can be explicitly expressed as

$$\mathrm{KL}(\mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}\|\bar{\mathbb{P}}_{\mathbf{z}_k|\mathcal{Y}_{k-1}})=\frac{1}{2}\left[\|\boldsymbol{c}_k-\boldsymbol{\mu}_k\|^2_{\boldsymbol{\Sigma}_k^{-1}}+\mathrm{Tr}\left[\boldsymbol{\Sigma}_k^{-1}\boldsymbol{S}_k-\boldsymbol{I}\right]-\ln\det\left(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{S}_k\right)\right]\leq\theta. \tag{2.23}$$

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

$$\mathbb{P}^*_{\mathbf{z}_k|\mathcal{Y}_{k-1}} = \mathcal{N}_{n+m}\left(\boldsymbol{\mu}_k, \boldsymbol{S}^*_k\right), \tag{2.24}$$

where

$$\boldsymbol{S}^*_k = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{xx,k} & \boldsymbol{\Sigma}_{xy,k} \\ \boldsymbol{\Sigma}_{yx,k} & \boldsymbol{\Sigma}_{yy,k} \end{bmatrix}, \tag{2.25}$$

and $\tilde{\boldsymbol{\Sigma}}_{xx,k}$ is determined by the boundary condition $\mathrm{KL}(\mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}} \| \bar{\mathbb{P}}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}) = \theta$ [89]. In the τ-divergence case, the forms of the corresponding $\mathbb{P}^*_{\mathbf{z}_k|\mathcal{Y}_{k-1}}$ and $\boldsymbol{S}^*_k$ are the same as those in (2.24) and (2.25), respectively, but $\tilde{\boldsymbol{\Sigma}}_{xx,k}$ is determined instead from the boundary condition $D_\tau(\mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}} \| \bar{\mathbb{P}}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}) = \theta$ [90].

Eq. (2.25) admits that the distributionally robust state estimation under the Kullback–Leibler divergence (in general, the τ-divergence) can be written as

$$\hat{\mathbf{x}}_{k|k} = \boldsymbol{\mu}_{x,k} + \boldsymbol{\Sigma}_{xy,k} \cdot (\boldsymbol{\Sigma}_{yy,k})^{-1}(\mathbf{y}_k - \boldsymbol{\mu}_{y,k}), \tag{2.26}$$

which is of the same form as the optimal estimation under the nominal distribution, i.e., (A.11). This means that under the Kullback–Leibler divergence or the τ-divergence, the worst-case conditional distribution of $\mathbf{z}_k$ given $\mathcal{Y}_{k-1}$ at the current time step does not directly influence the optimal robust estimation at the same time step. This phenomenon keeps the same as that under the moment-based ambiguity; cf. (2.20). 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 [89, 90].

Wasserstein Distance

The origin of the Wasserstein distance (i.e., Kantorovich-Rubinshtein metric) was inspired by the optimal transport theory [122]; see also [123]. It is currently of most interests in operations research [62] and machine learning [56, 124]. For any two distributions $\mathbb{P}_{\mathbf{a}}$ and $\mathbb{Q}_{\mathbf{b}}$ supported on the same set, the Wasserstein distance is defined as [62, 122]

$$\mathrm{W}(\mathbb{P}_{\mathbf{a}}, \mathbb{Q}_{\mathbf{b}}) := \inf_{\Pi_{\mathbf{a},\mathbf{b}}} \int \|\boldsymbol{a} - \boldsymbol{b}\| \Pi_{\mathbf{a},\mathbf{b}}(\mathrm{d}\boldsymbol{a}, \mathrm{d}\boldsymbol{b}) \tag{2.27}$$

where $\mathbf{a}$ and $\mathbf{b}$ are random vectors associated with $\mathbb{P}_{\mathbf{a}}$ and $\mathbb{Q}_{\mathbf{b}}$, respectively; $\Pi_{\mathbf{a},\mathbf{b}}$ is any possible joint distribution of $(\mathbf{a}, \mathbf{b})$ whose marginals are $\mathbb{P}_{\mathbf{a}}$ and $\mathbb{Q}_{\mathbf{b}}$; $\|\cdot\|$ denotes any possible vector norm.

In this case, the ambiguity set is as (A.3). See also [91]. When we consider the distributionally

robust estimation problem (2.4), (A.3) is specified into

$$\mathcal{F}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}(\theta)=\left\{\mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}\in\mathcal{P}(\mathbb{R}^n\times\mathbb{R}^m)\;\middle|\;\mathrm{W}\left(\mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}},\bar{\mathbb{P}}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}\right)\leq\theta\right\}. \tag{2.28}$$

If we suppose $\mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}$ is also Gaussian, (2.28) can be explicitly expressed as

$$\mathrm{W}\left(\mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}},\bar{\mathbb{P}}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}\right)=\sqrt{\|\boldsymbol{c}_k-\boldsymbol{\mu}_k\|^2+\mathrm{Tr}\left[\boldsymbol{S}_k+\boldsymbol{\Sigma}_k-2\left(\boldsymbol{\Sigma}_k^{\frac{1}{2}}\boldsymbol{S}_k\boldsymbol{\Sigma}_k^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]}\leq\theta. \tag{2.29}$$

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

$$\mathbb{P}^*_{\mathbf{z}_k|\mathcal{Y}_{k-1}}=\mathcal{N}_{n+m}\left(\boldsymbol{\mu}_k,\boldsymbol{S}_k^*\right), \tag{2.30}$$

where $\boldsymbol{S}_k^*$ solves (2.14) subject to (2.29) [91].

Eq. (2.30) suggests that the distributionally robust state estimation under the Wasserstein ambiguity set is (2.18), which is generally not guaranteed to have the same form as (2.20) and (2.26). This means that, under the Wasserstein distance, the worst-case conditional distribution of $\mathbf{z}_k$ given $\mathcal{Y}_{k-1}$ at the current time step directly influences the optimal robust estimate at the same time step.

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 $\mathbf{z}_k$ given $\mathcal{Y}_{k-1}$ is Gaussian. Otherwise, there is no explicit equivalence between (2.22) and (2.23) and between (2.28) and (2.29). 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 $\mathbf{x}_{k-1}$ and $\mathbf{w}_{k-1}$ are white (i.e., mutually independent) Gaussian and $\boldsymbol{G}_{k-1}$ is deterministically constant, $\mathbf{x}_k$ 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., (2.14) *s.t.* (2.29)], the nonlinearity of our NSDP under the moment ambiguity set [i.e., (2.14) *s.t.* (2.15)] 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 $\mathbf{z}_k$ given $\mathcal{Y}_{k-1}$ at the current time step (i.e., k) **directly influences** the optimal robust estimate at the same time step, while under the Kullback–Leibler (in general, the τ-divergence) ambiguity set and the moment ambiguity set [cf. (2.20)], it **does not directly influence** the optimal robust estimate 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-based distributionally robust estimator do nothing to robustify the state estimator. 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 $\boldsymbol{P}_{k|k}^{*}$ at the time step k. Therefore, according to (2.9), they have different $\boldsymbol{\Sigma}_{k+1}$ values, which leads to different state estimates at the time step $k+1$.

2.2.5 Comparisons with Existing Frameworks

Regarding modeling uncertainties in (2.1), 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 [53, 85]), in the sense of maximum likelihood (see [86]), or by leveraging an auxiliary term (see [84]).

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 [31, 111]) 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 [53], we study the linear system

$$\begin{cases} \mathbf{x}_k &= \boldsymbol{F}_{k-1}\mathbf{x}_{k-1} + \boldsymbol{\Gamma}_{k-1}\mathbf{d}_{k-1} + \boldsymbol{G}_{k-1}\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases} \tag{2.31}$$

where $\mathbf{d}_{k-1} \in \mathbb{R}^q$ is the unknown input used to describe the modeling uncertainties. Note that the unknown-input $\mathbf{d}_k$ may also exist in the measurement dynamics [85, 86, 101, 116]. 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 [27], we are concerned with the linear system

$$\begin{cases} \mathbf{x}_k &= (\boldsymbol{F}_{k-1} + \delta\boldsymbol{F}_{k-1})\mathbf{x}_{k-1} + (\boldsymbol{G}_{k-1} + \delta\boldsymbol{G}_{k-1})\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases} \tag{2.32}$$

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:

$$\begin{bmatrix} \delta \boldsymbol{F}_{k-1} & \delta \boldsymbol{G}_{k-1} \end{bmatrix} = \boldsymbol{M}_{k-1} \Delta_{k-1} \begin{bmatrix} \boldsymbol{E}_{f,k-1} & \boldsymbol{E}_{g,k-1} \end{bmatrix}, \tag{2.33}$$

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 [29], in which the focused linear system is the same as (2.32), but $\delta \boldsymbol{F}_{k-1}$ and $\delta \boldsymbol{G}_{k-1}$ are modeled as

$$\begin{cases} \delta \boldsymbol{F}_{k-1} &= \sum_{i=1}^{l} \boldsymbol{F}_{i,k-1} \cdot \zeta_{i,k-1} \\ \delta \boldsymbol{G}_{k-1} &= \sum_{i=1}^{l} \boldsymbol{G}_{i,k-1} \cdot \zeta_{i,k-1}, \end{cases} \tag{2.34}$$

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 section where the modeling uncertainties are described by a family of distributions; see (A.1), (2.4), (2.11), and (2.15).

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 θ_1 and θ_2 in (2.15) or only one scale θ in (2.16)] rather than subtly designed matrices [e.g., $\boldsymbol{\Gamma}_{k-1}$ in (2.31); $\boldsymbol{M}_{k-1}$, $\boldsymbol{E}_{f,k-1}$, and $\boldsymbol{E}_{g,k-1}$ in (2.33); and $\boldsymbol{F}_{i,k-1}$ and $\boldsymbol{G}_{i,k-1}$ in (2.34)] are required to describe the modeling uncertainties. This means that when ONLY the nominal model (2.1) 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 (2.31), (2.33), and (2.34) 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 (2.31), (2.33), or (2.34), the specifically designed frameworks might outperform our new distributional framework. The claims in this subsection will be validated in the experiments.

2.2.6 Experiments

This subsection compares the state estimation performance of the existing filters with our newly proposed filter for linear systems subject to parameter uncertainty. 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 in this section 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.

We continue studying the classical instance discussed in [27, 89, 91], 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 (n.b., in [27], α was fixed as 0.099). In this state estimation problem, the nominal system matrix is

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

Candidate Filters

According to Subsection 2.2.1 and Subsection 2.2.5, 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 [79] (cf. [54]).
4. **Fading**: the adaptive fading-memory Kalman-like filter [107].
5. $\boldsymbol{H}_\infty$: the H_∞ filter [73].
6. **UB**: the upper-bound Kalman-like filter [31].
7. **UI**: the unknown-input Kalman-like filter [53].
8. **SPU**: the filter for stochastic parametric uncertainties [29].
9. **SNKF**: Sayed's norm-constrained Kalman-like filter [27].

10. τ**-KF**: the τ-divergence Kalman-like filter [90].
11. **WKF**: the Wasserstein Kalman-like filter [91].
12. **MKF**: the moment-based distributionally robust state estimator introduced in this section.

The twelve methods above are representatives of the five filter generations beginning with the canonical Kalman filter in the 1960s and ending with the filter proposed in this section. We do not consider the set-valued Kalman-like filter [76], the guaranteed-cost Kalman-like filter [111], and the traditional risk-sensitive Kalman-like filter [75] because in [27], they have been substantially studied and compared. Note that the τ-divergence Kalman-like filter [90] generalizes the Kullback–Leibler Kalman-like filter [89] (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 [89, 90].

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 $\hat{\boldsymbol{x}}_{0|0} = [0, 0]^\top$ and its corresponding state estimation error covariance $\boldsymbol{P}^*_{0|0}$ is set as $\mathrm{diag}\{1, 1\}$, where $\mathrm{diag}\{\cdot\}$ denotes a diagonal matrix [27, 89, 91]. 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 [73]) 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 [27], we set $\boldsymbol{M}_{k-1} = [0.0198, 0]^\top$, $\boldsymbol{E}_{f,k-1} = [0, \alpha/0.0198]$, and $\boldsymbol{E}_{g,k-1} = [0, 0]$ in (2.33), 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 [53], we set $\boldsymbol{\Gamma}_k = [1, 0]^\top$ in (2.31) 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 [53].

In the filter for stochastic parametric uncertainties [29], we have $l = 1$ in (2.34),

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

and $\boldsymbol{G}_{1,k-1} = \boldsymbol{0}$. Note that $\zeta_{i,k-1}$ is assumed to have unit variance in [29]. 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 to $\mathrm{diag}\{1,1\}$).

In the τ-divergence Kalman-like filter [90], we let $\tau = 0$ (therefore, the τ-divergence Kalman-like filter specifies the Kullback–Leibler Kalman-like filter [89]) and the radius of the ambiguity set be 1.5×10^{-4}.

In the Wasserstein Kalman-like filter [91], the radius of the ambiguity set is set to 0.1.

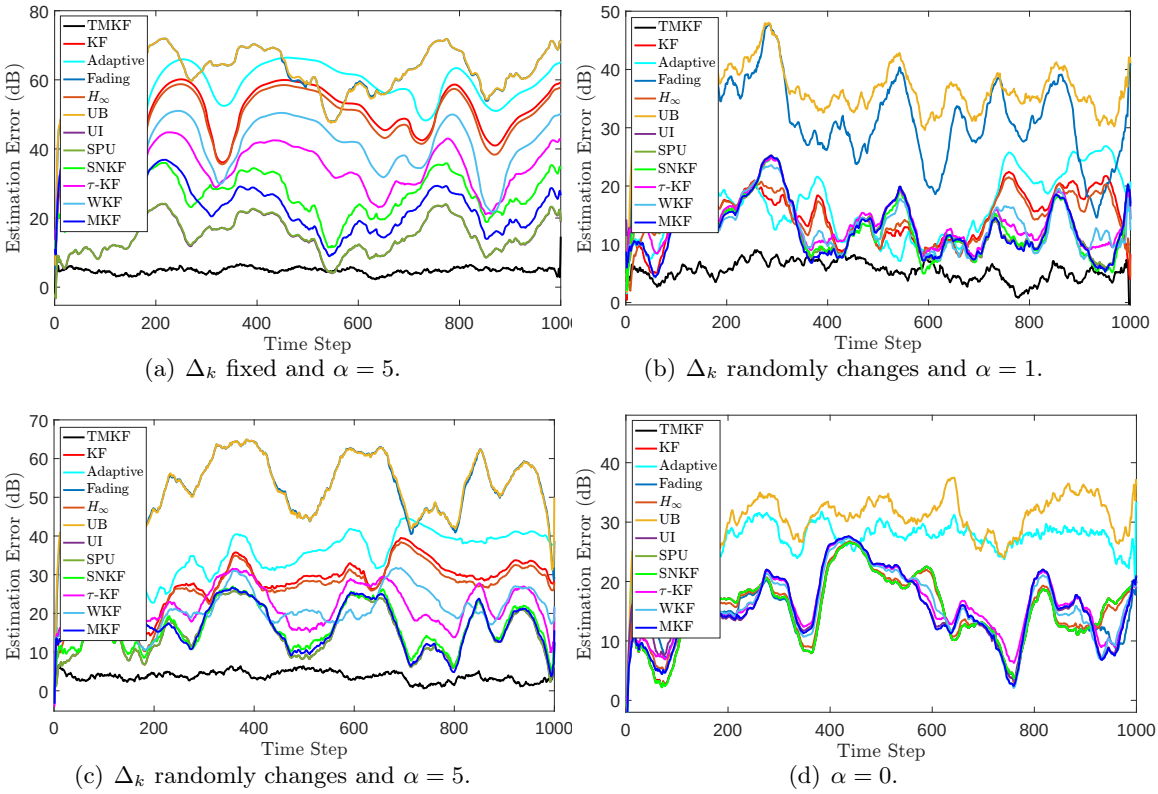
In our moment-based distributionally robust filter, $\theta = 0.02$, and therefore, $\theta_2 = 1.02$ (see Algorithm 2.1).

Suppose each simulation episode runs $T = 1000$ discrete-time steps. The estimation error at each time 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. $x_{2,k}$) 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 [27] and [91] 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. 2.1 (a) and Table 2.1.



(a) Δ_k fixed and $\alpha = 5$. (b) Δ_k randomly changes and $\alpha = 1$.

(c) Δ_k randomly changes and $\alpha = 5$. (d) $\alpha = 0$.

Figure 2.1: Results with prior known structural information (for $\mathcal{H}_\infty$, the prior parametric information is known, i.e., $\gamma = 102$). In (a), SPU and UI coincide.

- 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. 2.1 (b) and Table 2.2.
- 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. 2.1 (c) and Table 2.3.
- Fourth, we let $\alpha = 0$; i.e., there are no modeling uncertainties. The results are shown in Fig. 2.1 (d) and Table 2.4.

Note that the UB filter [31], which is in essence a kind of fading-memory Kalman-like filter (cf. [106]) is inappropriate for the instance discussed in this section because Assumption (19) of [31] 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. To significantly distinguish the different plots in figures, we are using relatively large values for α (i.e., 1 and 5) rather than a small value of 0.099 as in [27].

From Fig. 2.1 and Tables 2.1–2.4, the conclusions below can be outlined.

Table 2.1: 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: 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 2.2: 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 2.1 for table notes.

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.

Table 2.3: 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 2.1 for table notes.

Table 2.4: 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 2.1 for table notes.

5) The MKF is essentially the traditional fading-memory Kalman-like filter with a fixed fading factor θ_2. However, it outperforms the adaptive-factor fading-memory Kalman-like filters in [107] and [31]. 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) [89, 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 semi-definite program (it is well known that solving a semi-definite 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 other robust filtering frameworks (see Table 2.4). 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.

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 $\begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix}$, 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 (2.33), $\boldsymbol{\Gamma}_k = [0, 1]^\top$ in (2.31), and $\boldsymbol{F}_{1,k-1} = \begin{bmatrix} 0 & 0 \\ 3 & 0 \end{bmatrix}$ in (2.34). 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 $\mathcal{H}_\infty$ filter, we do not select a large enough γ (see [73]) in advance to guarantee the existence of the $\mathcal{H}_\infty$ filter. Alternatively, we arbitrarily select $\gamma = 25$ (rather than minimally required 102). As we can expect, the 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. 2.2 and Table 2.5.

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. 2.1, 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

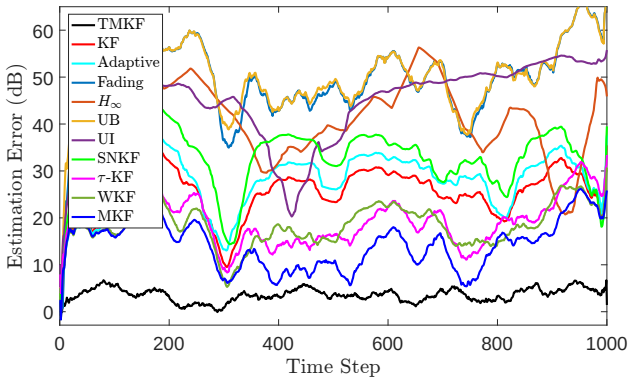


Figure 2.2: Results without prior structural/parametric information. In this case, only the distributionally robust estimators can outperform the canonical Kalman filter. Filters that are aware of structural/parametric information (e.g., UI, SNKF, and $\mathcal{H}_\infty$) perform poorly. Moreover, SPU even fails to work (and therefore is not plotted).

Table 2.5: 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

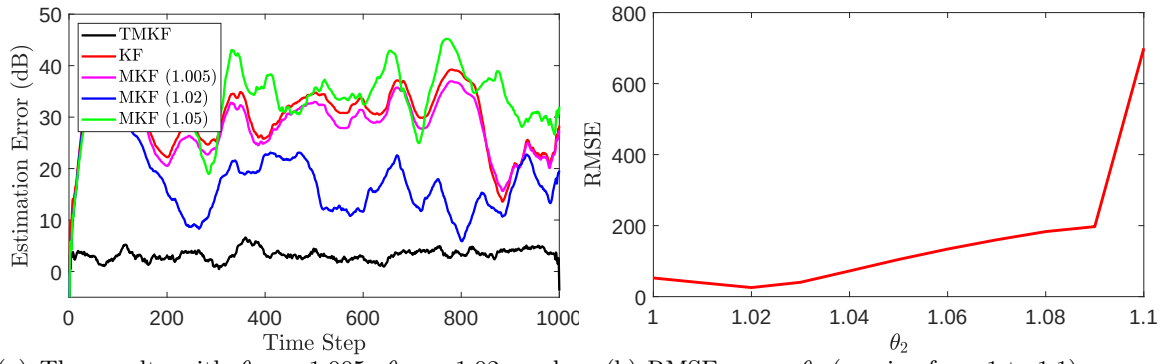
See Table 2.1 for table notes.

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.

Suggestions on Tuning the Size of the Ambiguity Set

The size of the ambiguity set (2.11) is controlled by three scalars, namely, θ_1, θ_2, and θ_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 (2.11), we must have $\theta_3 \geq 0$ and $\theta_2 \geq 1 \geq \theta_1 \geq 0$. Note that when $\theta_3 = 0$ and $\theta_2 = \theta_1 = 1$, the ambiguity set (2.11) 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 $\theta_3 \equiv 0$ [see (B.8)], is irrelevant to θ_1 [see (2.17)], and only depends on θ_2. Therefore, θ_1 can be any value in

$[0, 1]$, and we only investigate how to tune θ_2. The caption of Fig. 2.3 lists the RMSEs of the candidate filters.



(a) The results with $\theta_2 = 1.005$, $\theta_2 = 1.02$, and $\theta_2 = 1.05$.

(b) RMSE versus θ_2 (varying from 1 to 1.1).

Figure 2.3: Results with different θ_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.

From Fig. 2.3, it is evident that θ_2 can be neither too large nor too small to obtain satisfactory estimation performance. The robust state estimator with a too small value of θ_2 has insufficient robustness (i.e., insufficient ability against uncertainties), while that with a too large value of θ_2 is too conservative to produce satisfactory estimation performance. Unfortunately, the optimal tuning method for θ_2 is unknown (unless θ_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 author can only suggest that readers try appropriate values for their specific problems. Nevertheless, we believe that tuning a scalar θ_2 is significantly easier than tuning structural matrices $\boldsymbol{\Gamma}_{k-1}$ in (2.31), $\boldsymbol{M}_{k-1}$, $\boldsymbol{E}_{f,k-1}$ and $\boldsymbol{E}_{g,k-1}$ in (2.33), and $\boldsymbol{F}_{i,k-1}$ and $\boldsymbol{G}_{i,k-1}$ in (2.34).

A possible tuning method of θ_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 θ_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.

2.2.7 Section Conclusions

In this section, the distributionally robust state estimation method for linear Markov systems subject to parameter uncertainties is proposed. We integrate the existing Kullback–Leibler-divergence robust state estimation method [89], the τ-divergence robust state estimation method

[90], the Wasserstein-distance robust state estimation method [91], 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 (A.1)] can be described by several means, such as the τ-divergence, the Kullback–Leibler divergence (2.22), the Wasserstein distance (2.28), and the proposed moment-based ambiguity set (2.11). The detailed comparisons among those different ambiguity sets can be revisited in Subsection 2.2.4. The newly proposed moment-based filter in this section 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 Theorem 2). 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 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. 2.3, we can see that the proposed algorithm is not robust with respect to the size of the ambiguity set (i.e., θ_2). Unfortunately, the optimal or convincing tuning method for the size of ambiguity sets (e.g., θ_2 in this section; ρ in [91]; and c in [89, 90]) has yet to be found. We invite scholars in this field to collaborate with the author on addressing the two issues below in the future.

1) How can θ_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 $\theta_{2,k}$ where k denotes the discrete time?

Although imperfect, the proposed method is still promising because tuning a scalar θ_2 is easier than tuning structural matrices $\boldsymbol{\Gamma}_{k-1}$ in (2.31), $\boldsymbol{M}_{k-1}$, $\boldsymbol{E}_{f,k-1}$, and $\boldsymbol{E}_{g,k-1}$ in (2.33), and $\boldsymbol{F}_{i,k-1}$ and $\boldsymbol{G}_{i,k-1}$ in (2.34).

2.3 Addressing Parameter Uncertainty And Measurement Outlier

In this section, we study the distributionally robust state estimation problem for the linear system (2.1) subject to both parameter uncertainties and measurement outliers. In this case, the ambiguity set $\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)$ is necessarily required to contain fat-tailed distributions for $\mathbf{y}_k$. In addition, linear estimators that are inherently sensitive to measurement outliers are not acceptable.

From (2.4), we are inspired to **first** study a distributionally robust Bayesian estimation problem

$$\min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}}} \max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \mathbb{E}[\mathbf{x}-\boldsymbol{\phi}(\mathbf{y})][\mathbf{x}-\boldsymbol{\phi}(\mathbf{y})]^\top \tag{2.35}$$

subject to the nominal prior state distribution $\bar{\mathbb{P}}_{\mathbf{x}}$, the nominal conditional measurement distribution $\bar{\mathbb{P}}_{\mathbf{y}|\mathbf{x}}$, a properly constructed ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$ that contains fat-tailed distributions for $\mathbf{y}$, and the linear measurement equation

$$\mathbf{y} = \boldsymbol{H}\mathbf{x} + \mathbf{v}, \tag{2.36}$$

where $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{v}$ have finite second moments with appropriate dimensions and distributions. In (2.35), the expectation is taken over $\mathbb{P}_{\mathbf{x},\mathbf{y}}$. The subscript k (i.e., discrete time index) is dropped to avoid notational clutter. **Then**, by identifying the joint distribution of $(\mathbf{x}_k, \mathbf{y}_k)$ conditional on $\mathcal{Y}_{k-1}$, we can solve (2.4).

2.3.1 Distributionally Robust Bayesian Estimation

With linear measurement relation (2.36), the joint state-measurement distribution $\mathbb{P}_{\mathbf{x},\mathbf{y}}$ can be determined by (specifically, linearly shifted from) $\mathbb{P}_{\mathbf{x},\mathbf{v}}$ which has marginals $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{v}}$. In such a situation, it is reasonable and common to assume that the state $\mathbf{x}$ is independent of the measurement noise $\mathbf{v}$. As a result, we have $p_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y}) = p_{\mathbf{x},\mathbf{v}}(\boldsymbol{x},\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}) = p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}|\boldsymbol{x})p_{\mathbf{x}}(\boldsymbol{x}) = p_{\mathbf{v}|\mathbf{x}}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}|\boldsymbol{x})p_{\mathbf{x}}(\boldsymbol{x}) = p_{\mathbf{v}}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})p_{\mathbf{x}}(\boldsymbol{x})$, where $p_{\mathbf{x},\mathbf{y}}(\boldsymbol{x},\boldsymbol{y})$ and $p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y})$ are the density associated with $\mathbb{P}_{\mathbf{x},\mathbf{y}}$ and $\mathbb{P}_{\mathbf{x}|\mathbf{y}}$, respectively. Therefore,

$$p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}|\boldsymbol{x}) = p_{\mathbf{v}}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x}). \tag{2.37}$$

To solve the min-max problem (2.35), we are required to identify the least-favorable distribution from the ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$. However, it depends on the specific choice of the estimator $\boldsymbol{\phi}(\cdot)$.

Therefore, motivated by the strong min-max property (i.e., the saddle point property) in (2.13), we can alternatively try to solve the max-min problem of (2.35) first, i.e.,

$$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}}} \operatorname{Tr}\mathbb{E}[\mathbf{x}-\boldsymbol{\phi}(\mathbf{y})][\mathbf{x}-\boldsymbol{\phi}(\mathbf{y})]^\top \tag{2.38}$$

and then try to prove the strong min-max property between (2.35) and (2.38).[7] By the weak min-max property, it is unconditionally true that

$$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}}} \operatorname{Tr}\mathbb{E}[\mathbf{x}-\boldsymbol{\phi}(\mathbf{y})][\mathbf{x}-\boldsymbol{\phi}(\mathbf{y})]^\top \leq \min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}}} \max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr}\mathbb{E}[\mathbf{x}-\boldsymbol{\phi}(\mathbf{y})][\mathbf{x}-\boldsymbol{\phi}(\mathbf{y})]^\top.$$

The equality stands only when the strong min-max property holds which is not generally guaranteed. The max-min problem is easier to solve because for every $\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$, we can find the associated optimal estimator. We first study the optimal estimator for the nominal case.

Theorem 3. *Suppose $\mathbf{x}\sim\mathcal{N}_n(\bar{\boldsymbol{x}},\boldsymbol{M})$ nominally, $\mathbf{x}$ is independent of $\mathbf{v}$, all involved densities exist, and all involved integration and differentiation are interchangeable (i.e., densities are twice continuously differentiable). Let $\mathbf{s}:=\mathbf{y}-\boldsymbol{H}\bar{\boldsymbol{x}}$ denote the innovation vector, $\boldsymbol{S}$ the associated covariance, and $\mathbf{u}:=\boldsymbol{S}^{-1/2}\mathbf{s}$ the diagonalized and normalized innovation. Then for the nominal joint state-measurement distribution $\bar{\mathbb{P}}_{\mathbf{x},\mathbf{y}}$, the optimal estimator $\hat{\mathbf{x}}$ of $\mathbf{x}$, i.e., $\mathbb{E}(\mathbf{x}|\mathbf{y})$, is*

$$\hat{\mathbf{x}}=\bar{\boldsymbol{x}}+\boldsymbol{M}\boldsymbol{H}^\top\boldsymbol{S}^{-1/2}\left[-\frac{\mathrm{d}\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}\right]_{\boldsymbol{\mu}=\mathbf{u}}, \tag{2.39}$$

and the estimation error covariance $\mathbb{E}(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^\top$, evaluated over $\mathbb{P}_{\mathbf{x},\mathbf{y}}$, is

$$\boldsymbol{P}=\boldsymbol{M}-\boldsymbol{M}\boldsymbol{H}^\top\boldsymbol{S}^{-1/2}\mathbb{E}\left\{\left[-\frac{\mathrm{d}^2\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top}\right]_{\boldsymbol{\mu}=\mathbf{u}}\right\}\boldsymbol{S}^{-1/2}\boldsymbol{H}\boldsymbol{M}, \tag{2.40}$$

where

$$p_{\mathbf{u}}(\boldsymbol{\mu})=p_{\mathbf{y}}(\boldsymbol{S}^{1/2}\boldsymbol{\mu}+\boldsymbol{H}\bar{\boldsymbol{x}})\cdot\det(\boldsymbol{S}^{1/2}) \tag{2.41}$$

is the density of $\mathbf{u}$, $p_{\mathbf{y}}(\cdot)$ is the density of $\mathbf{y}$, and $\det(\cdot)$ denotes the determinant of a matrix. In (2.40), the inner expectation is taken over $\mathbb{P}_{\mathbf{u}}$. Note that both $-\frac{\mathrm{d}\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}$ and $-\frac{\mathrm{d}^2\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top}$ are functions of $\boldsymbol{\mu}$. $[-\frac{\mathrm{d}\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}]_{\boldsymbol{\mu}=\mathbf{u}}$ means that $\boldsymbol{\mu}$ is replaced with $\mathbf{u}$ in $-\frac{\mathrm{d}\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}$, and $[-\frac{\mathrm{d}^2\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top}]_{\boldsymbol{\mu}=\mathbf{u}}$ means that $\boldsymbol{\mu}$ is replaced with $\mathbf{u}$ in $-\frac{\mathrm{d}^2\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top}$.

Proof. See Appendix B.5. □

Theorem 3 reveals the benefit of the Gaussianity assumption of $\bar{\mathbb{P}}_{\mathbf{x}}=\mathcal{N}_n(\bar{\boldsymbol{x}},\boldsymbol{M})$. Specifically, without the Gaussianity assumption, we cannot have the closed-form expression of $\hat{\mathbf{x}}$ as in (2.39).

[7] The intuition is that the objective of (2.35) is positive-definite quadratic (thus convex) in $\boldsymbol{\phi}$ and linear (thus concave) in $\mathbb{P}$. Hence, we expect the strong min-max property.

We use an example below to give further intuitions for Theorem 3.

Example 1. *Suppose* $\mathbf{v}$ *follows a Gaussian distribution:* $\mathbb{P}_{\mathbf{v}} = \mathcal{N}_m(\mathbf{0}, \boldsymbol{R})$*. Then, the innovation* $\mathbf{s} := \mathbf{y} - \boldsymbol{H}\bar{\boldsymbol{x}} = \boldsymbol{H}(\mathbf{x} - \bar{\boldsymbol{x}}) + \mathbf{v}$ *is also Gaussian with mean of* $\mathbf{0}$ *and covariance* $\boldsymbol{S} = \boldsymbol{H}\boldsymbol{M}\boldsymbol{H}^\top + \boldsymbol{R}$*. Likewise, the normalized innovation* $\mathbf{u} := \boldsymbol{S}^{-1/2}\mathbf{s}$ *is Gaussian with mean of* $\mathbf{0}$ *and covariance of* $\boldsymbol{I}$*. Namely, the density of* $\mathbf{u}$ *is*

$$p_{\mathbf{u}}(\boldsymbol{\mu}) = \frac{1}{\sqrt{(2\pi)^m}} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\mu}\right).$$

As a result, we have

$$-\frac{\mathrm{d} \ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}} = \frac{1}{2}\frac{\mathrm{d}\boldsymbol{\mu}^\top \boldsymbol{\mu}}{\mathrm{d}\boldsymbol{\mu}} = \boldsymbol{\mu} = \boldsymbol{S}^{-1/2}\left[\boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}\right],$$

and the optimal **estimator** *is given as*

$$\begin{aligned}
\hat{\mathbf{x}} &= \bar{\boldsymbol{x}} + \boldsymbol{M}\boldsymbol{H}^\top \boldsymbol{S}^{-1/2}\left[-\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\mu}} \ln p_{\mathbf{u}}(\boldsymbol{\mu})\right]_{\boldsymbol{\mu}=\mathbf{u}} \\
&= \bar{\boldsymbol{x}} + \boldsymbol{M}\boldsymbol{H}^\top \boldsymbol{S}^{-1}\left[\mathbf{y} - \boldsymbol{H}\bar{\boldsymbol{x}}\right] \\
&= \bar{\boldsymbol{x}} + \boldsymbol{M}\boldsymbol{H}^\top (\boldsymbol{H}\boldsymbol{M}\boldsymbol{H}^\top + \boldsymbol{R})^{-1}\left[\mathbf{y} - \boldsymbol{H}\bar{\boldsymbol{x}}\right].
\end{aligned}$$

Likewise,

$$-\frac{\mathrm{d}^2 \ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top} = \boldsymbol{I},$$

and therefore,

$$\begin{aligned}
\boldsymbol{P} &= \boldsymbol{M} - \boldsymbol{M}\boldsymbol{H}^\top \boldsymbol{S}^{-1/2}\mathbb{E}\left\{\left[-\tfrac{\mathrm{d}^2}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top} \ln p_{\mathbf{u}}(\boldsymbol{\mu})\right]_{\boldsymbol{\mu}=\mathbf{u}}\right\} \boldsymbol{S}^{-1/2}\boldsymbol{H}\boldsymbol{M} \\
&= \boldsymbol{M} - \boldsymbol{M}\boldsymbol{H}^\top \boldsymbol{S}^{-1/2}\boldsymbol{I}\boldsymbol{S}^{-1/2}\boldsymbol{H}\boldsymbol{M} \\
&= \boldsymbol{M} - \boldsymbol{M}\boldsymbol{H}^\top \boldsymbol{S}^{-1}\boldsymbol{H}\boldsymbol{M} \\
&= \boldsymbol{M} - \boldsymbol{M}\boldsymbol{H}^\top (\boldsymbol{H}\boldsymbol{M}\boldsymbol{H}^\top + \boldsymbol{R})^{-1}\boldsymbol{H}\boldsymbol{M}.
\end{aligned}$$

We end up with the standard Kalman formulas. □

Example 1 shows that Theorem 3 is a generalization for the Gaussian-distributed measurement noise $\mathbf{v}$. In Theorem 3, the **true** measurement noise $\mathbf{v}$ is no longer required to be Gaussian. Namely, fat-tailed noise distributions, e.g., t-distribution, Laplacian distribution, can be considered to model the measurement noises. The closed-form state estimator can be obtained by deriving the distribution of the normalized innovation $\mathbf{u}$. However, the **nominal** distribution of $\mathbf{v}$ is stilled assumed to be Gaussian.

Since $\mathbf{u} := \boldsymbol{S}^{-1/2}(\mathbf{y} - \boldsymbol{H}\bar{\boldsymbol{x}}) = \boldsymbol{S}^{-1/2}[\boldsymbol{H}(\mathbf{x} - \bar{\boldsymbol{x}}) + \mathbf{v}]$, we have $\mathbb{E}\mathbf{u} = \mathbf{0}$ and $\mathbb{E}\mathbf{u}\mathbf{u}^\top = \boldsymbol{I}$. Thus, if $\mathbf{x}$

and $\mathbf{v}$ were all normally distributed, $\boldsymbol{p}_{\mathbf{u}}(\boldsymbol{\mu})$ would be a standard Gaussian density because the independence between $\mathbf{x}$ and $\mathbf{v}$ has already been assumed. Specifically, for every $i, j \in [m]$ [8] and $i \neq j$, $\mathbb{E}\mathrm{u}_i = \mathbb{E}\mathrm{u}_j = 0$, $\mathbb{E}\mathrm{u}_i^2 = \mathbb{E}\mathrm{u}_j^2 = 1$, and $\mathbb{E}\mathrm{u}_i\mathrm{u}_j = 0$. Therefore, we have

$$\mathbb{E}\left[-\frac{\mathrm{d}^2 \ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top}\bigg|_{\boldsymbol{\mu}=\mathbf{u}}\right] = \boldsymbol{I} \cdot \mathbb{E}\left[-\frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2}\bigg|_{\mu=\mathrm{u}}\right], \tag{2.42}$$

where the left expectation is taken over $\mathbb{P}_{\mathbf{u}}$ and the right expectation is taken over $\mathbb{P}_{\mathrm{u}}$. Note that the entry-wise $p_{\mathrm{u}}(\mu)$ is different from the joint $p_{\mathbf{u}}(\boldsymbol{\mu})$ and we have $p_{\mathbf{u}}(\boldsymbol{\mu}) = \prod_{i=1}^m p_{\mathrm{u}_i}(\mu_i)$.

Definition 1. *For simplicity in notation, in the following, we use* [9]

$$\mathbb{E}\left[-\frac{\mathrm{d}^2 \ln p(\mu)}{\mathrm{d}\mu^2}\right]$$

to implicitly stand for its full form

$$\mathbb{E}\left[-\frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2}\bigg|_{\mu=\mathrm{u}}\right].$$

Let $g(\mu) := -\frac{\mathrm{d}^2 \ln p(\mu)}{\mathrm{d}\mu^2}$. *One should always keep it in mind that*

$$\begin{aligned}\mathbb{E}\left[-\frac{\mathrm{d}^2 \ln p(\mu)}{\mathrm{d}\mu^2}\right] &:= \mathbb{E}\left[-\frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2}\bigg|_{\mu=\mathrm{u}}\right] \\ &= \mathbb{E}g(\mathrm{u}) \\ &= \int g(\mu)p(\mu)\mathrm{d}\mu \\ &= \int \left[-\frac{\mathrm{d}^2 \ln p(\mu)}{\mathrm{d}\mu^2}\right] p(\mu)\mathrm{d}\mu.\end{aligned}$$

Also, $p(\mu)$ *would implicitly stand for its full form* $p_{\mathrm{u}}(\mu)$ *because there is no risk of confusion.* □

For $p_{\mathrm{u}}(\mu)$, we identify that $-\frac{\mathrm{d}}{\mathrm{d}\mu}\ln p(\mu)$ is the score function[10] of the distribution $p_{\mathrm{u}}(\mu)$ and $\mathbb{E}[-\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln p(\mu)]$ the associated Fisher information. Eq. (2.42) is attractive since it allows us to only study a univariate problem rather than a multivariate one. This motivated us to study the normalized and diagnalized innovation $\mathbf{u}$ instead of the original (i.e., non-normalized) innovation $\mathbf{s}$. Hence, (2.40) can be simplified to

$$\boldsymbol{P} = \boldsymbol{M} - \boldsymbol{M}\boldsymbol{H}^\top \boldsymbol{S}^{-1}\boldsymbol{H}\boldsymbol{M} \cdot \mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln p(\mu)\right]. \tag{2.43}$$

[8] Note that the normalized innovation $\mathbf{u}$ is a m-length random vector because $\mathbf{y} \in \mathbb{R}^m$.

[9] In the applied statistics community, people may also use $\mathbb{E}[-\frac{\mathrm{d}^2 \ln p(\mathrm{u})}{\mathrm{du}^2}]$. To avoid possible confusion, in this thesis, we do not adopt this notation.

[10] Namely, the maximum likelihood estimator of the mean of $p_{\mathrm{u}}(\mu)$; see also Appendix A.4.

By the results in Theorem 3, we can find the explicit and tractable reformulation of the max-min problem (2.38).

Corollary 3. *Suppose the possibly true distribution of* $\mathbf{x}$ *is* $\mathcal{N}_n(\boldsymbol{c_x}, \boldsymbol{\Sigma_x})$ *[cf. the nominal* $\mathcal{N}_n(\bar{\boldsymbol{x}}, \boldsymbol{M})$ *in Theorem 3]. The max-min problem (2.38) can be reformulated to*

$$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \boldsymbol{P}, \tag{2.44}$$

where

$$\boldsymbol{P} = \boldsymbol{\Sigma_x} - \boldsymbol{\Sigma_x} \boldsymbol{H}^\top \boldsymbol{S}^{-1} \boldsymbol{H} \boldsymbol{\Sigma_x} \cdot \mathbb{E}\left[-\frac{d^2}{d\mu^2} \ln p(\mu) \right], \tag{2.45}$$

and $\boldsymbol{S}$ *is the covariance matrix of the innovation vector* $\mathbf{s} := \mathbf{y} - \boldsymbol{H}\boldsymbol{c}_x$*. In this case, the normalized and diagonalized innovation is defined as*

$$\mathbf{u} := \boldsymbol{S}^{-1/2}(\mathbf{y} - \boldsymbol{H}\boldsymbol{c_x})$$

and the corresponding optimal **estimator** *is*

$$\hat{\mathbf{x}} = \boldsymbol{c}_x + \boldsymbol{\Sigma}_x \boldsymbol{H}^\top \boldsymbol{S}^{-1/2} \left[-\frac{d}{d\boldsymbol{\mu}} \ln p_{\mathbf{u}}(\boldsymbol{\mu}) \right]_{\boldsymbol{\mu}=\mathbf{u}}.$$

Proof. This is immediate from Theorem 3 and (2.43). Since

$$\mathbb{E}\left[-\frac{d}{d\boldsymbol{\mu}} \ln p_{\mathbf{u}}(\boldsymbol{\mu}) \Big|_{\boldsymbol{\mu}=\mathbf{u}} \right] = -\int [p(\boldsymbol{\mu})]^{-1} \frac{dp(\boldsymbol{\mu})}{d\boldsymbol{\mu}} p(\boldsymbol{\mu}) d\boldsymbol{\mu} = -\int \frac{dp(\boldsymbol{\mu})}{d\boldsymbol{\mu}} d\boldsymbol{\mu} = \mathbf{0},$$

we have $\mathbb{E}\hat{\mathbf{x}} = \boldsymbol{c}_x = \mathbb{E}\mathbf{x}$, implying that $\hat{\mathbf{x}}$ is an unbiased estimate so that the minimum mean square error matrix coincides with the minimum error covariance matrix. □

To explicitly solve (2.44), we need to define the ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$. The prior state distribution is Gaussian as argued. We can construct the ambiguity set for $\mathbb{P}_{\mathbf{x}}$ as

$$\mathcal{F}_{\mathbf{x}}(\theta) = \left\{ \mathbb{P}_{\mathbf{x}} = \mathcal{N}_n(\boldsymbol{c_x}, \boldsymbol{\Sigma_x}) \;\middle|\; D(\mathbb{P}_{\mathbf{x}}, \bar{\mathbb{P}}_{\mathbf{x}}) \le \theta \right\}.$$

where $\bar{\mathbb{P}}_{\mathbf{x}}$ is the nominal Gaussian distribution of $\mathbf{x}$ [i.e., $\mathcal{N}_n(\bar{\boldsymbol{x}}, \boldsymbol{M})$ in Theorem 3], $D(\cdot,\cdot)$ is a statistical metric (e.g., Wasserstein metric) or divergence (e.g., Kullback–Leibler divergence), and $\theta \in \mathbb{R}_+$ is the radius to control the scale and conservativeness of the set. The larger the θ, the more conservative the robust estimation is. Specially, the ambiguity set for $\mathbb{P}_{\mathbf{x}}$ could be one of the follows.

1) **Kullback–Leibler divergence** (KL divergence).

$$\mathcal{F}_{\mathbf{x}}(\theta_x) = \left\{ \mathbb{P}_{\mathbf{x}} = \mathcal{N}_n(\boldsymbol{c_x}, \boldsymbol{\Sigma_x}) \;\middle|\; \mathrm{KL}(\mathbb{P}_{\mathbf{x}} \| \bar{\mathbb{P}}_{\mathbf{x}}) \le \theta_{\boldsymbol{x}} \right\}, \tag{2.46}$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the KL divergence and under Gaussianity assumption, $\mathrm{KL}(\mathbb{P}_{\mathbf{x}}\|\bar{\mathbb{P}}_{\mathbf{x}}) = \frac{1}{2}[\|\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|^2_{\boldsymbol{M}^{-1}} + \mathrm{Tr}\left[\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{I}\right] - \ln\det\left(\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)]$ [89]. Note that the explicit expression for any two multivariate distributions does not always exist. Only for Gaussians, the above equality holds. Extensions and generalizations for the KL divergence include the τ-divergence [90], the ϕ-divergence (a.k.a. f-divergence) [125], etc. They all contain the KL divergence as a special case.

2) **Wasserstein distance**.

$$\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}}) = \left\{ \mathbb{P}_{\mathbf{x}} = \mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \;\middle|\; \mathrm{W}(\mathbb{P}_{\mathbf{x}}, \bar{\mathbb{P}}_{\mathbf{x}}) \le \theta_{\boldsymbol{x}} \right\}, \tag{2.47}$$

where $\mathrm{W}(\cdot,\cdot)$ denotes the Wasserstein metric and under Gaussianity assumption, the type-2 Wasserstein distance is given as $\mathrm{W}(\mathbb{P}_{\mathbf{x}}, \bar{\mathbb{P}}_{\mathbf{x}}) = \sqrt{\|\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|^2 + \mathrm{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2(\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}})^{\frac{1}{2}}]}$ [69, 91]. Note also that the explicit expression for any two multivariate distributions does not always exist. Only for Gaussians, the above equality holds.

3) **Moment-based set** [120].

$$\begin{aligned}
\mathcal{F}_{\mathbf{x}}(\theta_{1,\boldsymbol{x}}, \theta_{2,\boldsymbol{x}}, \theta_{3,\boldsymbol{x}}) &= \left\{ \mathbb{P}_{\mathbf{x}} = \mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \;\middle|\; \begin{array}{l} [\mathbb{E}\mathbf{x} - \bar{\boldsymbol{x}}]^\top \boldsymbol{M}^{-1} [\mathbb{E}\mathbf{x} - \bar{\boldsymbol{x}}] \le \theta_{3,\boldsymbol{x}} \\ \mathbb{E}(\mathbf{x} - \bar{\boldsymbol{x}})(\mathbf{x} - \bar{\boldsymbol{x}})^\top \preceq \theta_{2,\boldsymbol{x}} \boldsymbol{M} \\ \mathbb{E}(\mathbf{x} - \bar{\boldsymbol{x}})(\mathbf{x} - \bar{\boldsymbol{x}})^\top \succeq \theta_{1,\boldsymbol{x}} \boldsymbol{M} \end{array} \right\} \\
&= \left\{ \mathbb{P}_{\mathbf{x}} = \mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \;\middle|\; \begin{array}{l} [\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}}]^\top \boldsymbol{M}^{-1} [\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}}] \le \theta_{3,\boldsymbol{x}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} + (\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}})(\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}})^\top \preceq \theta_{2,\boldsymbol{x}} \boldsymbol{M} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} + (\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}})(\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}})^\top \succeq \theta_{1,\boldsymbol{x}} \boldsymbol{M} \end{array} \right\}.
\end{aligned} \tag{2.48}$$

As we can see, in general, we need to use three parameters to define a moment-based ambiguity set, $\theta_{3,\boldsymbol{x}} \ge 0$ and $\theta_{2,\boldsymbol{x}} \ge 1 \ge \theta_{1,\boldsymbol{x}} \ge 0$.

Next, we define the ambiguity set for $\mathbb{P}_{\mathbf{y}|\mathbf{x}}$, i.e., the measurement distribution conditioned on prior state. Since the nominal $\bar{\mathbb{P}}_{\mathbf{y}|\mathbf{x}}$ is given by nominal $\bar{\mathbb{P}}_{\mathbf{v}}$,[11] we need to define the ambiguity set for $\mathbb{P}_{\mathbf{v}}$. Suppose the nominal distribution of the measurement noise $\mathbf{v}$, by the Gaussianity assumption, is $\bar{\mathbb{P}}_{\mathbf{v}} := \mathcal{N}_m(\mathbf{0}, \boldsymbol{R})$. The ambiguity set for $\mathbb{P}_{\mathbf{v}}$ can be one of the follows.

1) **Kullback–Leibler divergence** (KL divergence).

$$\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}}) = \left\{ \mathbb{P}_{\mathbf{v}} = \mathcal{N}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \;\middle|\; \mathrm{KL}(\mathbb{P}_{\mathbf{v}}\|\bar{\mathbb{P}}_{\mathbf{v}}) \le \theta_{\boldsymbol{v}} \right\}. \tag{2.49}$$

[11] Recall from (2.37) that $p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}|\boldsymbol{x}) = p_{\mathbf{v}}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})$.

2) **Wasserstein distance**.

$$\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}}) = \left\{ \mathbb{P}_{\mathbf{v}} = \mathcal{N}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \,\middle|\, \mathrm{W}(\mathbb{P}_{\mathbf{v}}, \bar{\mathbb{P}}_{\mathbf{v}}) \leq \theta_{\boldsymbol{v}} \right\}. \tag{2.50}$$

3) **Moment-based set**.

$$\mathcal{F}_{\mathbf{v}}(\theta_{1,\boldsymbol{v}}, \theta_{2,\boldsymbol{v}}, \theta_{3,\boldsymbol{v}}) = \left\{ \mathbb{P}_{\mathbf{v}} = \mathcal{N}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \,\middle|\, \begin{array}{l} [\boldsymbol{c}_{\boldsymbol{v}} - \mathbf{0}]^\top \boldsymbol{R}^{-1} [\boldsymbol{c}_{\boldsymbol{v}} - \mathbf{0}] \leq \theta_{3,\boldsymbol{v}} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} + (\boldsymbol{c}_{\boldsymbol{v}} - \mathbf{0})(\boldsymbol{c}_{\boldsymbol{v}} - \mathbf{0})^\top \preceq \theta_{2,\boldsymbol{v}} \boldsymbol{R} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} + (\boldsymbol{c}_{\boldsymbol{v}} - \mathbf{0})(\boldsymbol{c}_{\boldsymbol{v}} - \mathbf{0})^\top \succeq \theta_{1,\boldsymbol{v}} \boldsymbol{R} \end{array} \right\}. \tag{2.51}$$

The explicit expressions for $\mathrm{KL}(\cdot\|\cdot)$ and $\mathrm{W}(\cdot,\cdot)$ are similar to those for $\mathbb{P}_{\mathbf{x}}$ in (2.46) and (2.47), respectively.

Given the nominal Gaussian distributions of the prior state $\mathbf{x}$ and the measurement noise $\mathbf{v}$, the marginal distribution of the measurement $\mathbf{y}$ (or equivalently, the innovations $\mathbf{s}$ and $\mathbf{u}$) is also Gaussian, so is the joint state-measurement distribution. However, when outliers appear in the measurement $\mathbf{y}$, they appear in the normalized innovation $\mathbf{u}$ (and u) as well. That means the possibly true distribution $\mathbb{P}_{\mathrm{u}}$ is likely to deviate from the nominal standard Gaussian distribution and, simultaneously, has fat tails. Let $F_{\mathrm{u}}(\mu)$ denote the cumulative distribution function induced from $\mathbb{P}_{\mathrm{u}}$, and $\Phi(\mu)$ the cumulative distribution function of the standard Gaussian distribution. Motivated by the M-estimation theory for outlier attenuation/rejection [126], we can construct the ambiguity set for $\mathbb{P}_{\mathrm{u}}$ as one of the follows.

1) **ϵ-contamination set**.

$$\mathcal{F}_{\mathrm{u}}(\epsilon) = \left\{ \mathbb{P}_{\mathrm{u}} \in \mathcal{P}(\mathbb{R}) \,\middle|\, \begin{array}{l} F_{\mathrm{u}}(\mu) = \mathbb{P}_{\mathrm{u}}(\mathrm{u} \leq \mu) \\ \sup\limits_{\mu \in \mathbb{R}} \|F_{\mathrm{u}}(\mu) - \Phi(\mu)\| \leq \epsilon \\ F_{\mathrm{u}}(\mu) = (1-\epsilon)\Phi(\mu) + \epsilon H(\mu) \\ H(\mu) = 1 - H(-\mu), \quad H(\mu) \text{ is a distribution function on } \mathbb{R} \end{array} \right\}. \tag{2.52}$$

Note that $\sup_{\mu\in\mathbb{R}} \|F_{\mathrm{u}}(\mu) - \Phi(\mu)\| = \sup_{\mu\in\mathbb{R}} \|(1-\epsilon)\Phi(\mu) + \epsilon H(\mu) - \Phi(\mu)\| = \epsilon \cdot \sup_{\mu\in\mathbb{R}} \|H(\mu) - \Phi(\mu)\| \leq \epsilon$ (i.e., in this case the statistical metric is particularized into the infinity distance of two functions). Suppose the random scalar z is an indicator and uniformly distributed in the interval $[0,1]$. $F_{\mathrm{u}}(\mu) = \int_{-\infty}^{\mu} \int_0^1 p_{\mathrm{u},\mathrm{z}}(\tau, z) \mathrm{d}z \mathrm{d}\tau = \Phi(\mu)\mathbb{I}(z \geq \epsilon) + H(\mu)\mathbb{I}(z \leq \epsilon) = (1-\epsilon)\Phi(\mu) + \epsilon H(\mu)$ where $\mathbb{I}(\cdot)$ is the indicator function. Therefore, in (2.52), $F_{\mathrm{u}}(\mu) = (1-\epsilon)\Phi(\mu) + \epsilon H(\mu)$ means that with probability $1-\epsilon$ the measurement innovation u (equivalently, a measurement $\boldsymbol{y}$) is from a nominal Gaussian distribution, and with probability ϵ it is from a contamination (fat-tailed) distribution $H(\mu)$ (i.e., outlier). $H(\mu) = 1 - H(-\mu)$ means that the density

associated with $H(\mu)$ is symmetric about $\mu = 0$.

2) **ϵ-normal set**.

$$\mathcal{F}_{\mathrm{u}}(\epsilon) = \left\{ \mathbb{P}_{\mathrm{u}} \in \mathcal{P}(\mathbb{R}) \left| \begin{array}{l} F_{\mathrm{u}}(\mu) = \mathbb{P}_{\mathrm{u}}(\mathrm{u} \leq \mu) \\ \sup\limits_{\mu \in \mathbb{R}} \| F_{\mathrm{u}}(\mu) - \Phi(\mu) \| \leq \epsilon \\ F_{\mathrm{u}}(\mu) = 1 - F_{\mathrm{u}}(-\mu) \end{array} \right. \right\}. \tag{2.53}$$

Clearly, the ϵ-normal set is larger and more general than the ϵ-contamination set for the same radius ϵ. However, we usually prefer the ϵ-contamination set because: 1) it has clearer physical meaning than that of the ϵ-normal set; 2) in view of properties of real measurement data, the least-favorable distribution in (2.52) is more reasonable than that in the ϵ-normal set; and 3) the distributionally robust state estimator over the ϵ-contamination set is much easier to design. Other possible choice for the structure of $\mathcal{F}_{\mathrm{u}}(\epsilon)$ includes the p-value set [97] which is also a subset of (2.53), etc.

Recall that the distribution of the normalized innovation $\mathbf{u}$ is uniquely determined given the distributions of the prior state $\mathbf{x}$ and the measurement noise $\mathbf{v}$ because $\mathbf{u} := \boldsymbol{S}^{-1/2}[\boldsymbol{H}(\mathbf{x} - \boldsymbol{c}_x) + \mathbf{v}]$. Thus, when we admit the ϵ-contamination/normal deviation from the nominal Gaussian distribution for $\mathbf{u}$, we implicitly admit that the deviation is from the distribution(s) of $\mathbf{x}$ or $\mathbf{v}$ or both. Since the ϵ-contamination/normal deviation studied here accounts for measurement outliers, we argue that it is related to $\mathbf{v}$ and regardless of $\mathbf{x}$. However, for technical simplicity in problem solving, we work on $\mathbf{u}$ instead of $\mathbf{v}$ although directly on $\mathbf{v}$ might be intuitively more understandable. Therefore, we would first design the distributions of $\mathbf{x}$ and $\mathbf{u}$, and the value of $\boldsymbol{S}$. Then, we can obtain the distribution of $\mathbf{v}$ through $\mathbf{v} = \boldsymbol{S}^{1/2}\mathbf{u} - \boldsymbol{H}(\mathbf{x} - \boldsymbol{c}_{\boldsymbol{x}})$. Specifically, $\boldsymbol{S}$ controls the covariance of $\mathbf{v}$, while $\mathbf{u}$ controls the type of $\mathbf{v}$. In summary, we have Highlight 1.

Highlight 1. *The measurement noise* $\mathbf{v}$ *suffers from two kinds of distributional uncertainties:*

1) deviations imposed on mean and covariance [see (2.49), (2.50), and (2.51)];

2) deviations existing as outliers [see (2.52) and (2.53)].

However, the first one does not imply the second, and vice versa. They independently discredit the nominal Gaussian assumption of $\mathbf{v}$*. The Item 1) modifies the first two moments of* $\mathbf{v}$*, while the Item 2) modifies the tails of* $\mathbf{v}$ *to be fat.*[12] □

Intuitively, Item 1) accounts for parameter uncertainty of $\mathbb{P}_{\mathbf{v}}$, while Item 2) for type uncertainty of $\mathbb{P}_{\mathbf{v}}$; recall Introduction 1.1 for the concepts of different types of model uncertainties. Since (2.49), (2.50), and (2.51) are mainly used to define the ambiguity sets for the outlier-unrelated

[12]One may also recall that $\mathbf{y} = \boldsymbol{S}^{1/2}\mathbf{u} + \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}}$. Hence, the type of the distribution of $\mathbf{y}$ (i.e., fat-tailed or not) is only determined by the type of the distribution of $\mathbf{u}$. Other distributions (e.g., the distribution of $\mathbf{x}$), if involved, only influence the parameter(s) of the distribution of $\mathbf{y}$, regardless of its type.

part of $\mathbb{P}_{\mathbf{v}}$ (i.e., mean and covariance of $\mathbf{v}$), taking Gaussianity assumption does not lead to discrepancy. This is because any methods that are suitable to define the ambiguity sets for the mean and covariance of $\mathbf{v}$ are acceptable. Hence, using (2.49), (2.50), and (2.51) is a practical choice. Note that (2.49), (2.50), and (2.51) do not imply that the worst-case distribution of $\mathbf{v}$ is Gaussian; instead, the worst-case distribution of $\mathbf{v}$ is determined by the worst-case distributions of $\mathbf{x}$ and $\mathbf{u}$, and the worst-case value of $\boldsymbol{S}$ through $\mathbf{v} = \boldsymbol{S}^{1/2}\mathbf{u} - \boldsymbol{H}(\mathbf{x} - \boldsymbol{c}_{\boldsymbol{x}})$; see Theorem 8 and especially its proof. Eqs. (2.49), (2.50), and (2.51) only mean that the outlier-unrelated part of $\mathbb{P}_{\mathbf{v}}$ is Gaussian; the three formulas are used to determine the worst-case value of $\boldsymbol{S}$. Instead, the outlier-related part of $\mathbb{P}_{\mathbf{v}}$ is defined by the worst-case distribution of $\mathbf{u}$, which is non-Gaussian.

As a consequence, the max-min problem (2.38) or (2.44) is equivalent to

$$\max_{\mathbb{P}_{\mathbf{x}} \in \mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}}),\ \mathbb{P}_{\mathbf{v}} \in \mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}}),\ \mathbb{P}_{\mathrm{u}} \in \mathcal{F}_{\mathrm{u}}(\epsilon)} \operatorname{Tr} \boldsymbol{P}, \tag{2.54}$$

where $\boldsymbol{P}$ is defined in (2.45); $\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})$, $\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})$, and $\mathcal{F}_{\mathrm{u}}(\epsilon)$ can be any types of ambiguity sets available above.

Theorem 4. *Consider the problem (2.54). The following statements are true.*

1) Reformulations for $\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})$.

a) In (2.46), $\boldsymbol{c}_{\boldsymbol{x}} = \bar{\boldsymbol{x}}$ so that $\mathrm{KL}(\mathbb{P}_{\mathbf{x}} \| \bar{\mathbb{P}}_{\mathbf{x}}) = \frac{1}{2}\left[\operatorname{Tr}\left[\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{I}\right] - \ln\det\left(\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)\right]$.

b) In (2.47), $\boldsymbol{c}_{\boldsymbol{x}} = \bar{\boldsymbol{x}}$ so that $\mathrm{W}(\mathbb{P}_{\mathbf{x}}, \bar{\mathbb{P}}_{\mathbf{x}}) = \sqrt{\operatorname{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2(\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}})^{\frac{1}{2}}]}$.

c) In (2.48), $\boldsymbol{c}_{\boldsymbol{x}} = \bar{\boldsymbol{x}}$ so that $\boldsymbol{\Sigma}_{\boldsymbol{x}} \preceq \theta_{2,\boldsymbol{x}}\boldsymbol{M}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \theta_{1,\boldsymbol{x}}\boldsymbol{M}$.

2) Reformulations for $\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})$.

a) In (2.49), $\boldsymbol{c}_{\boldsymbol{v}} = \mathbf{0}$ so that $\mathrm{KL}(\mathbb{P}_{\mathbf{v}} \| \bar{\mathbb{P}}_{\mathbf{v}}) = \frac{1}{2}\left[\operatorname{Tr}\left[\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{v}} - \boldsymbol{I}\right] - \ln\det\left(\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{v}}\right)\right]$.

b) In (2.50), $\boldsymbol{c}_{\boldsymbol{v}} = \mathbf{0}$ so that $\mathrm{W}(\mathbb{P}_{\mathbf{v}}, \bar{\mathbb{P}}_{\mathbf{v}}) = \sqrt{\operatorname{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{v}} + \boldsymbol{R} - 2(\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{R}^{\frac{1}{2}})^{\frac{1}{2}}]}$.

c) In (2.51), $\boldsymbol{c}_{\boldsymbol{v}} = \mathbf{0}$ so that $\boldsymbol{\Sigma}_{\boldsymbol{v}} \preceq \theta_{2,\boldsymbol{v}}\boldsymbol{R}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}} \succeq \theta_{1,\boldsymbol{v}}\boldsymbol{R}$.

3) The problem (2.54) is equivalent to

$$\max_{\mathbb{P}_{\mathbf{x}} \in \mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})} \max_{\mathbb{P}_{\mathbf{v}} \in \mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})} \max_{\mathbb{P}_{\mathrm{u}} \in \mathcal{F}_{\mathrm{u}}(\epsilon)} \operatorname{Tr} \boldsymbol{P}. \tag{2.55}$$

The order of the three maximizations does not matter.

Proof. The estimation error covariance $\boldsymbol{P}$ in (2.54) does not depend on $\boldsymbol{c}_{\boldsymbol{x}}$ and $\boldsymbol{c}_{\boldsymbol{v}}$. In order to maximize $\boldsymbol{P}$, the larger the feasible sets of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$, the better. This leads to $\boldsymbol{c}_{\boldsymbol{x}} = \bar{\boldsymbol{x}}$ and $\boldsymbol{c}_{\boldsymbol{v}} = \mathbf{0}$. Namely, the distributional uncertainty budgets $\theta_{\boldsymbol{x}}$ and $\theta_{\boldsymbol{v}}$ are completely assigned to describe deviations of covariances of $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{v}}$, respectively, regardless of $\boldsymbol{c}_{\boldsymbol{x}}$ or $\boldsymbol{c}_{\boldsymbol{v}}$. This proves the first two claims 1) and 2). The claim 3) is standard in the optimization community. □

Since the original (i.e., non-normalized) innovation $\mathbf{s} = \mathbf{y} - \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}} = \boldsymbol{H}(\mathbf{x} - \boldsymbol{c}_{\boldsymbol{x}}) + \mathbf{v}$, and $\mathbf{x}$ and $\mathbf{v}$ are Gaussian and independent, the nominal value of $\boldsymbol{S}$ can be obtained as $\boldsymbol{S} = \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}}$. Let $i_\mu \in \mathbb{R}_+$ denote the Fisher information of $p_{\mathrm{u}}(\mu)$; $i_\mu := \mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln p(\mu)\right] \geq 0$; recall Definition 1. Comparing with (2.45), $\boldsymbol{P}$ in (2.55) can be written as

$$\boldsymbol{P} = \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_\mu.$$

In view of the first two claims 1) and 2) in Theorem 4, we identify that $\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})$ is parameterized by $\boldsymbol{\Sigma}_{\boldsymbol{x}} \in \mathbb{S}_+^n$ and $\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})$ is parameterized by $\boldsymbol{\Sigma}_{\boldsymbol{v}} \in \mathbb{S}_{++}^m$ (due to non-singularity of $\boldsymbol{S}$). Hence, (2.55) can be equivalently given as

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \max_{i_\mu} \operatorname{Tr} \boldsymbol{P} \tag{2.56}$$

where the feasible sets of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ are defined in $\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})$ and $\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})$, respectively. As a result, we can solve the reformulated max-min problem (2.56) independently and sequentially, i.e., solving the innermost first and the outermost last.[13]

The following two lemmas solve the innermost sub-problem over i_μ.

Lemma 1. *The functional optimization over the ϵ-contamination ambiguity set*

$$\min_{p(\mu)} \mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln p(\mu)\Big|_{\mu=\mathrm{u}}\right]$$

$$s.t. \quad \begin{cases} p(\mu) = \dfrac{\mathrm{d}F_{\mathrm{u}}(\mu)}{\mathrm{d}\mu} \\ \sup\limits_{\mu\in\mathbb{R}} \|F_{\mathrm{u}}(\mu) - \Phi(\mu)\| \leq \epsilon \\ F_{\mathrm{u}}(\mu) = (1-\epsilon)\Phi(\mu) + \epsilon H(\mu) \\ H(\mu) = 1 - H(-\mu), \qquad\qquad H(\mu) \text{ is a distribution function on } \mathbb{R} \end{cases}$$

is solved by the following least-favorable distribution

$$p(\mu) = \begin{cases} (1-\epsilon)\frac{1}{\sqrt{2\pi}}e^{K\mu+\frac{1}{2}K^2}, & \mu \leq -K \\ (1-\epsilon)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\mu^2}, & |\mu| \leq K \\ (1-\epsilon)\frac{1}{\sqrt{2\pi}}e^{-K\mu+\frac{1}{2}K^2}, & \mu \geq K, \end{cases} \tag{2.57}$$

where the constant $K \in \mathbb{R}_+$ is implicitly defined by ϵ through $\int_{-K}^{K} p(\mu)\mathrm{d}t + \frac{2p(K)}{K} = 1$.

[13] Because in this order, the problem is easy to solve.

Furthermore,

$$\min \mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln p(\mu)\right] = (1-\epsilon)[1-2\Phi(-K)].$$

Proof. See Appendix B.6. □

Lemma 2. *Given* $0 \leq \epsilon \lesssim 0.0303$*, the functional optimization over the* ϵ*-normal ambiguity set*

$$\min_{p(\mu)} \mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln p(\mu)\Big|_{\mu=\mathrm{u}}\right]$$

$$s.t. \quad \begin{cases} p(\mu) = \dfrac{\mathrm{d}F_{\mathrm{u}}(\mu)}{\mathrm{d}\mu} \\ \sup\limits_{\mu\in\mathbb{R}} \|F_{\mathrm{u}}(\mu) - \Phi(\mu)\| \leq \epsilon \\ F_{\mathrm{u}}(\mu) = 1 - F_{\mathrm{u}}(-\mu), \end{cases}$$

is solved by the following least-favorable distribution

$$p(\mu) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a^2} \cdot \cos^{-2}(\frac{1}{2}ca) \cdot \cos^2(\frac{1}{2}c\mu), & 0 \leq \mu \leq a \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\mu^2}, & a \leq \mu \leq b \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}b^2} \cdot e^{-b\mu+b^2}, & \mu \geq b \end{cases} \tag{2.58}$$

and $p(\mu) = p(-\mu)$*, where* a*,* b*, and* c *are implicitly defined by* ϵ *as*

1) $c\tan(\frac{1}{2}ca) = a \quad (0 \leq ca < \pi)$,

2) $\int_0^a p(\mu)\mathrm{d}\mu = \int_0^a \mathrm{d}\Phi(\mu) - \epsilon$,

3) $\int_b^\infty p(\mu)\mathrm{d}\mu = \int_b^\infty \mathrm{d}\Phi(\mu) + \epsilon$.

Furthermore, $\min \mathbb{E}\left[-\dfrac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln p(\mu)\right] = \dfrac{c^2 a}{\cos^2(\frac{1}{2}ca)} p(a) + 2\Phi(b) - 2\Phi(a)$.

Proof. See Appendix B.7. □

Lemma 1 reveals that the least-favorable distribution under the ϵ-contamination distributional uncertainty is Gaussian in the middle (i.e., when $|\mu| \leq K$) and is Laplacian in the tails (i.e., when $\mu \geq K$ and $\mu \leq -K$), while Lemma 2 reveals that the least-favorable distribution under the ϵ-normal distributional uncertainty is $\cos^2(\cdot)$-like in the middle (i.e., when $-a \leq \mu \leq a$), is Gaussian in the transitions (i.e., when $a \leq \mu \leq b$ and $-b \leq \mu \leq -a$), and is Laplacian in the tails (i.e., when $\mu \geq b$ and $\mu \leq -b$). The Laplacian (a.k.a. exponential) tails (i.e., fat tails) explain outliers in measurements. We call them least-favorable distributions because they have smallest Fisher information (i.e., largest asymptotic variance to estimate the mean) among

distributions in $\mathcal{F}_{\mathrm{u}}(\epsilon)$. Although both are theoretically sound, we usually prefer the results in Lemma 1 because they coincide well with our intuitions from practice that the main part of measurements are normally distributed and only a small part of them are outliers. However, the results in Lemma 2 become suitable when quantization noises are non-negligible (e.g., when low-bit sampler is adopted), because quantization noise is close, but not equal, to zero.

Remark 4. *In Lemma 2, we require that* $\epsilon \lesssim 0.0303$ *(n.b., for three real numbers,* $x \lesssim z$ *means that* $x \leq y$ *and* $y \approx z$*). This is a necessary condition to obtain the least-favorable distribution in (2.58). Otherwise, the least-favorable distribution is of a different form; see [55, p. 85 ff.]. Usually, we do not prefer the solution when* $\epsilon \gtrsim 0.0303$ *because the associated M-estimator has significantly larger asymptotic variance; see [55, Exhibit 4.6]. Theoretically, only when the true proportion of outliers is approximately smaller than* 0.0303 *can we use the solution in Lemma 2. However, in practice, the solution in Lemma 2 might not be sensitive to the true proportion of outliers: no matter what the true proportion of outliers (of course, as long as less than* 0.5*) in the true measurements, keeping* $\epsilon \equiv 0.0303$ *in our algorithm might not cause disasters. This observation is also true for the solution in Lemma 1. This point will be illustrated in the experiments in Subsection 2.3.5.* □

After solving the innermost sub-problem of (2.56), we then study the outer sub-problems.

Theorem 5. *The problem (2.56) is equivalent to*

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \ \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top}(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_{\mu}^{\min}\right], \tag{2.59}$$

where

$$i_{\mu}^{\min} := \min i_{\mu} := \min \mathbb{E}\left[-\frac{d^2}{d\mu^2}\ln p(\mu)\right]$$

is a constant defined in Lemma 1 or Lemma 2, whichever is adopted. Besides, $0 \leq i_{\mu}^{\min} \leq 1$.

Proof. Note that

$$\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top}(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \mathbf{0}$$

because $\boldsymbol{\Sigma}_{\boldsymbol{x}} \in \mathbb{S}_{+}^{n}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}} \in \mathbb{S}_{++}^{m}$. Hence, the non-negative and minimal i_{μ} maximizes $\boldsymbol{P}$. In addition, since the standard Gaussian is contained in the ϵ-contamination set and the ϵ-normal set, $i_{\mu}^{\min}$ is upper bounded by the Fisher information of the standard Gaussian which is one. □

Since we have three alternatives for $\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})$, three for $\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})$, and two for $\mathcal{F}_{\mathrm{u}}(\epsilon)$, in principle, we need to solve the problem (2.54) eighteen times. As demonstrations and without loss of generality, we suppose $\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})$ and $\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})$ have the same type of distributional uncertainty and study the distributionally robust Bayesian estimation (DRBE) under Wasserstein ambiguity and moment-based ambiguity, respectively.

Under Wasserstein ambiguities of $\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})$ and $\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})$, the problem (2.59) can be explicitly

written as

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top}(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_{\mu}^{\min}\right], \tag{2.60}$$

subject to

$$\begin{cases} \sqrt{\operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2\left(\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]} \leq \theta_{\boldsymbol{x}} \\ \sqrt{\operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{v}} + \boldsymbol{R} - 2\left(\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{R}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]} \leq \theta_{\boldsymbol{v}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \mathbf{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \mathbf{0}. \end{cases} \tag{2.61}$$

This problem is difficult to solve as: 1) the objective is nonlinear, 2) the feasible set (2.61) is non-convex because the function $\sqrt{\cdot}$ is concave and the constraint $\sqrt{\operatorname{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2(\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}})^{\frac{1}{2}}]} \leq \theta_{\boldsymbol{x}}$ is non-convex, so is the constraint for $\boldsymbol{\Sigma}_{\boldsymbol{v}}$. However, we can still reformulate it into a linear semi-definite program (SDP) using some algebraic tricks. Solving a linear SDP is basic, although still challenging, in the optimization community.

Theorem 6. *Suppose $\boldsymbol{R} \succ \mathbf{0}$. The problem (2.60) subject to (2.61) is solvable and can be reformulated into a linear SDP*

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}, \boldsymbol{V}_{\boldsymbol{x}}, \boldsymbol{V}_{\boldsymbol{v}}, \boldsymbol{U}} \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} - i_{\mu}^{\min} \cdot \boldsymbol{U}\right], \tag{2.62}$$

subject to

$$\begin{cases} \begin{bmatrix} \boldsymbol{U} & \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top} \\ \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top} + \boldsymbol{\Sigma}_{\boldsymbol{v}} \end{bmatrix} \succeq \mathbf{0} \\ \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2\boldsymbol{V}_{\boldsymbol{x}}\right] \leq \theta_{\boldsymbol{x}}^2 \\ \begin{bmatrix} \boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}} & \boldsymbol{V}_{\boldsymbol{x}} \\ \boldsymbol{V}_{\boldsymbol{x}} & \boldsymbol{I} \end{bmatrix} \succeq \mathbf{0} \\ \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{v}} + \boldsymbol{R} - 2\boldsymbol{V}_{\boldsymbol{v}}\right] \leq \theta_{\boldsymbol{v}}^2 \\ \begin{bmatrix} \boldsymbol{R}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{R}^{\frac{1}{2}} & \boldsymbol{V}_{\boldsymbol{v}} \\ \boldsymbol{V}_{\boldsymbol{v}} & \boldsymbol{I} \end{bmatrix} \succeq \mathbf{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \mathbf{0}, \boldsymbol{V}_{\boldsymbol{x}} \succeq \mathbf{0}, \boldsymbol{V}_{\boldsymbol{v}} \succeq \mathbf{0}, \boldsymbol{U} \succeq \mathbf{0}. \end{cases} \tag{2.63}$$

Proof. See Appendix B.8. □

Under moment-based ambiguities of $\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})$ and $\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})$, the problem (2.59) can be explicitly written as

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top}(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_{\mu}^{\min}\right], \tag{2.64}$$

subject to

$$\begin{cases} \boldsymbol{\Sigma}_{\boldsymbol{x}} \preceq \theta_{2,\boldsymbol{x}}\boldsymbol{M} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \theta_{1,\boldsymbol{x}}\boldsymbol{M} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \preceq \theta_{2,\boldsymbol{v}}\boldsymbol{R} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succeq \theta_{1,\boldsymbol{v}}\boldsymbol{R} \succ \mathbf{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \mathbf{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \mathbf{0}. \end{cases} \tag{2.65}$$

This problem is relatively easier to solve than (2.60) because the feasible set (2.65) consists of linear constraints, implying convexity and compactness. Note that $\boldsymbol{R} \succ \mathbf{0}$ indicates $\boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \mathbf{0}$. Therefore, it is solvable (i.e., the optimal solutions exist and are finite).

Theorem 7. *The problem (2.64) subject to (2.65) is analytically solved by* $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \theta_{2,\boldsymbol{x}}\boldsymbol{M}$ *and* $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \theta_{2,\boldsymbol{v}}\boldsymbol{R}$.

Proof. See Appendix B.9. □

Comparing with conclusions in Theorem 2, Theorem 7 allows the measurement noise covariance, and the process noise covariance and the estimation error covariance in the last time step to be inflated with different levels. Specifically, the measurement noise covariance is improved by $\theta_{2,\boldsymbol{v}}$, while the process noise covariance and the estimation error covariance in the last time step are improved by $\theta_{2,\boldsymbol{x}}$; cf. (2.75).

It is also possible to jointly use the Wasserstein metric and the moment-based set, e.g., the Wasserstein metric for $\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})$ and the moment-based set for $\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})$. The derivations are straightforward and we do not cover the details.

As we can see, the max-min problem under the moment-based distributional uncertainties admits attractive closed-form solutions which indicates high computational efficiency, especially for large scale estimation problems when n and m are (extremely) large. As for the problem under the Wasserstein metric, it requires solving a SDP which, although linear and solvable, is still computationally challenging. From the viewpoint of modeling, using the Wasserstein metric (2.61) [which is equivalent to (B.12) in Appendix B.8] or the moment-based set (2.65) just means that the shapes of the feasible sets are different. Since both (B.12) and (2.65) are convex and compact, for every $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ in (B.12), there exists $\theta_1 \in \mathbb{R}_+, \theta_2 \in \mathbb{R}_+$ for (2.65) such that $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ are contained in (2.65). Conversely, for every $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ in (2.65), there exists

$\theta \in \mathbb{R}_+$ for (B.12) such that $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ are contained in (B.12). Therefore, in practice, we are not entangled in which type of ambiguity set we should choose. We use the one under which the problem is easy to solve. It is this reason that we do not study the problem under the KL divergence ambiguity in this thesis. Because nonlinear functions, i.e., $\ln(\cdot)$, $\det(\cdot)$, in (2.46) and (2.49) render the max-min problem being a general nonlinear SDP (without linear reformulations) and difficult to solve. However, it is still convex and therefore solvable, since the constraints $\frac{1}{2}\left[\mathrm{Tr}\left[\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{I}\right] - \ln\det\left(\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)\right] \leq \theta_{\boldsymbol{x}}$ and $\frac{1}{2}\left[\mathrm{Tr}\left[\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{v}} - \boldsymbol{I}\right] - \ln\det\left(\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{v}}\right)\right] \leq \theta_{\boldsymbol{v}}$ are convex. The convexity of the constraints is straightforward to show as: 1) $\mathrm{Tr}\left[\cdot\right]$ is linear and convex; 2) both $\ln\left(\cdot\right)$ and $\det\left(\cdot\right)$ are concave; 3) $\ln\left(\cdot\right)$ is monotonically increasing. **Nevertheless**, note that different ambiguity sets do give different robust state estimates. Therefore, in practice, if computation powers allow, we should try all possible ambiguity sets to obtain better performance.

The theorem below summarizes the solution to the max-min problem (2.38).

Theorem 8. *Suppose the nominal distribution of* $\mathbf{x}$ *is* $\bar{\mathbb{P}}_{\mathbf{x}} = \mathcal{N}_n(\bar{\boldsymbol{x}}, \boldsymbol{M})$ *and of* $\mathbf{v}$ *is* $\bar{\mathbb{P}}_{\mathbf{v}} = \mathcal{N}_m(\mathbf{0}, \boldsymbol{R}), \boldsymbol{R} \succ \mathbf{0}$. *With Gaussianity assumptions for elements in the ambiguity sets* $\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})$ *and* $\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})$, *the max-min problem (2.38) is solved by*

1. *Optimal Estimator:*

$$\hat{\mathbf{x}} = \bar{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{x}}^* \boldsymbol{H}^\top \boldsymbol{S}^{*-1/2} \cdot \boldsymbol{\psi}[\boldsymbol{S}^{*-1/2}(\mathbf{y} - \boldsymbol{H}\bar{\boldsymbol{x}})], \tag{2.66}$$

where $\boldsymbol{S}^* := \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}^*\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}}^*$, $\boldsymbol{\psi}(\mu)$ *is entry-wise identical and for each entry*

$$\psi(\mu) = \begin{cases} -K, & \mu \leq -K \\ \mu, & |\mu| \leq K \\ K, & \mu \geq K, \end{cases} \tag{2.67}$$

if the ϵ*-contamination ambiguity set is used, or*

$$\psi(\mu) = -\psi(-\mu) = \begin{cases} c\tan(\frac{1}{2}c\mu), & 0 \leq \mu \leq a \\ \mu, & a \leq \mu \leq b \\ b, & \mu \geq b, \end{cases} \tag{2.68}$$

if the ϵ*-normal ambiguity set is used;* $\boldsymbol{\Sigma}_{\boldsymbol{x}}^*$ *and* $\boldsymbol{\Sigma}_{\boldsymbol{v}}^*$ *are the optimal solution of (2.62) if the Wasserstein metric is used, or of (2.64) if the moment-based set is used.*

2. *Worst-Case Estimation Error Covariance:*

$$\boldsymbol{P}^* = \boldsymbol{\Sigma}_{\boldsymbol{x}}^* - \boldsymbol{\Sigma}_{\boldsymbol{x}}^*\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}^*\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}}^*)^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}^* \cdot i_\mu^{\min}, \tag{2.69}$$

where

$$i_\mu^{\min} = (1-\epsilon)[1-2\Phi(-K)] \tag{2.70}$$

if the ϵ-contamination ambiguity set is used, or

$$i_\mu^{\min} = \frac{c^2 a}{\cos^2(\frac{1}{2}ca)} p(a) + 2\Phi(b) - 2\Phi(a) \tag{2.71}$$

if the ϵ-normal ambiguity set is used. For parameters K, a, b, and c, see Lemmas 1 and 2.

3. *Least-Favorable Distributions:*

   i) $\mathbb{P}_{\mathbf{x}}^* = \mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}}^*, \boldsymbol{\Sigma}_{\boldsymbol{x}}^*)$, *where* $\boldsymbol{c}_{\boldsymbol{x}}^* = \bar{\boldsymbol{x}}$.

   ii) $\mathbb{P}_{\mathrm{u}}^*$ *is defined in (2.57) if the ϵ-contamination ambiguity set is used, or in (2.58) if the ϵ-normal ambiguity set is used.*

   iii) $\mathbb{P}_{\mathbf{v}}^*$ *is determined by the convolution of* $\mathbb{P}_{\mathrm{u}}^*$ *and* $\mathbb{P}_{\mathbf{x}}^*$ *through* $\mathbf{v}^* = \boldsymbol{S}^{*\frac{1}{2}}\mathbf{u}^* - \boldsymbol{H}(\mathbf{x}^* - \bar{\boldsymbol{x}})$, *where* $\boldsymbol{S}^* := \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}^*\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}}^*$. *Here,* $\mathbf{v}^*$ *denotes the random vector associated with* $\mathbb{P}_{\mathbf{v}}^*$. *Notations keep similar to* $\mathbf{u}^*$ *and* $\mathbf{x}^*$.

Proof. See Appendix B.10. □

As we can see from (2.66), there exists a nonlinear function $\boldsymbol{\psi}(\cdot)$ in the estimator. It is used to limit the influence that an outlier may bring to the estimator. Whenever $\boldsymbol{y}$ is large, the value of $\boldsymbol{\psi}(\cdot)$ is limted to $\pm K$ in (2.67) or $\pm b$ in (2.68). Hence, we term $\boldsymbol{\psi}(\cdot)$ the **influence function**; cf. Appendix A.4. In this sense, the state estimator $\hat{\mathbf{x}}$ in (2.66) is robust against measurement outliers.

At last, we solve the min-max distributionally robust Bayesian estimation problem (2.35).

Theorem 9. *Under Gaussianity assumptions for nominal distributions of* $\mathbf{x}$ *and* $\mathbf{v}$, *the distributionally robust Bayesian estimation problem (2.35) admits the strong min-max property (i.e., the saddle point property)*

$$\min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}}} \max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} V(\boldsymbol{\phi},\mathbb{P}) = \max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}}} V(\boldsymbol{\phi},\mathbb{P}),$$

where $V(\boldsymbol{\phi},\mathbb{P}) := \mathrm{Tr}\,\mathbb{E}[\mathbf{x}-\boldsymbol{\phi}(\mathbf{y})][\mathbf{x}-\boldsymbol{\phi}(\mathbf{y})]^\top$. *Hence, the solutions to the max-min problem (2.38) also solve the min-max problem (2.35).*

Proof. See Appendix B.11. □

So far we have solved the distributionally robust Bayesian estimation problem subject to parameter uncertainties and measurement outliers. As a closing note, we mention that if we were sure that there are no outliers in measurements, we would have another modeling trick

to address the distributionally robust Bayesian estimation problem. The theorem below is an outlier-free supplement to Theorem 8.

Theorem 10. *If there are no outliers in measurements, we can directly model* $\mathbb{P}_{\mathbf{x},\mathbf{y}}$ *(or equivalently* $\mathbb{P}_{\mathbf{x},\mathbf{v}}$*) as a joint Gaussian distribution. In this special case, the ambiguity set admits* $D(\mathbb{P}_{\mathbf{x},\mathbf{y}}, \bar{\mathbb{P}}_{\mathbf{x},\mathbf{y}}) \leq \theta$*, parameterized by just one scalar* θ*.* $D(\cdot,\cdot)$ *can be any possible statistical metric or divergence (e.g., Wasserstein metric, KL divergence, moment-based set). If* $\mathbf{x}$ *and* $\mathbf{y}$ *are jointly Gaussian, the optimal estimate of* $\mathbf{x}$ *given* $\mathbf{y}$*, i.e.,* $\mathbb{E}(\mathbf{x}|\mathbf{y})$*, has an affine form. Suppose the worst-case distribution is*

$$\mathbb{P}^*_{\mathbf{x},\mathbf{y}} = \mathcal{N}_{n+m}\left(\begin{bmatrix} \boldsymbol{c}^*_{\boldsymbol{x}} \\ \boldsymbol{c}^*_{\boldsymbol{y}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}^*_{\boldsymbol{xx}} & \boldsymbol{\Sigma}^*_{\boldsymbol{xy}} \\ \boldsymbol{\Sigma}^*_{\boldsymbol{yx}} & \boldsymbol{\Sigma}^*_{\boldsymbol{yy}} \end{bmatrix}\right).$$

We have the distributionally robust estimator as $\hat{\mathbf{x}} = \boldsymbol{c}^*_{\boldsymbol{x}} + \boldsymbol{\Sigma}^*_{\boldsymbol{xy}}\boldsymbol{\Sigma}^{*-1}_{\boldsymbol{yy}}(\mathbf{y} - \boldsymbol{c}^*_{\boldsymbol{y}})$ *and the worst-case estimation error covariance as* $\boldsymbol{P}^* = \boldsymbol{\Sigma}^*_{\boldsymbol{xx}} - \boldsymbol{\Sigma}^*_{\boldsymbol{xy}}\boldsymbol{\Sigma}^{*-1}_{\boldsymbol{yy}}\boldsymbol{\Sigma}^*_{\boldsymbol{yx}}$*. Note that* $\mathbb{P}^*_{\mathbf{x},\mathbf{y}}$ *can be obtained in analogy to Theorem 6 if the Wasserstein metric is used, or to Theorem 7 if the moment-based set is used.*

Proof. This special case has been discussed in Section 2.2. □

When outliers exist in measurements, we can no longer assume that $\mathbf{x}$ and $\mathbf{y}$ (or equivalently $\mathbf{x}$ and $\mathbf{v}$) are jointly Gaussian. We have to separately discuss the ambiguity sets for $\mathbb{P}_{\mathbf{x}}$, $\mathbb{P}_{\mathbf{v}}$, and $\mathbb{P}_{\mathrm{u}}$, respectively. Even when there are no outliers in measurements, separately designing uncertainty sets for $\mathbf{x}$ and $\mathbf{y}$ (or equivalently $\mathbf{x}$ and $\mathbf{v}$) offers us more flexibility if we have different uncertainty levels towards them, because jointly modeling admits the same uncertainty levels.

2.3.2 Distributionally Robust State Estimation

With the results of distributionally robust Bayesian estimation developed in Section 2.3.1, this section solves the state estimation problem (2.4) at the time k. We just need to identify the nominal conditional prior distribution of the state given the past measurements, i.e., $\bar{\mathbb{P}}_{\mathbf{x}_k|\mathcal{Y}_{k-1}}$. In our Gaussian approximation framework, $\bar{\mathbb{P}}_{\mathbf{x}_k|\mathcal{Y}_{k-1}}$ is assumed to be Gaussian.

By (2.1), the nominal conditional prior distribution of the state $\mathbf{x}_k$ given the last state $\mathbf{x}_{k-1}$ is

$$\bar{\mathbb{P}}_{\mathbf{x}_k|\mathbf{x}_{k-1}} = \mathcal{N}_n\left(\boldsymbol{F}_{k-1}\mathbf{x}_{k-1},\ \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}\boldsymbol{G}^\top_{k-1}\right).$$

At the time $k-1$, suppose the distributionally robust posterior state estimate is $\mathbb{E}(\mathbf{x}_{k-1}|\mathcal{Y}_{k-1}) := \hat{\mathbf{x}}_{k-1|k-1}$ and the associated estimation error covariance is $\boldsymbol{P}^*_{k-1|k-1}$; the conditional distribution of $\mathbf{x}_{k-1}$ given $\mathcal{Y}_{k-1}$ is $\mathbb{P}_{\mathbf{x}_{k-1}|\mathcal{Y}_{k-1}} = \mathcal{N}_n\left(\hat{\mathbf{x}}_{k-1|k-1},\ \boldsymbol{P}^*_{k-1|k-1}\right)$. Therefore, the nominal conditional

prior distribution of the state $\mathbf{x}_k$ given $\mathcal{Y}_{k-1}$ is

$$\bar{\mathbb{P}}_{\mathbf{x}_k|\mathcal{Y}_{k-1}}(B) = \int_{\mathbb{R}^n} \bar{\mathbb{P}}_{\mathbf{x}_k|\mathbf{x}_{k-1}=\boldsymbol{x}_{k-1}}(B) \cdot \mathbb{P}_{\mathbf{x}_{k-1}|\mathcal{Y}_{k-1}}(\mathrm{d}\boldsymbol{x}_{k-1} \mid \mathcal{Y}_{k-1}), \qquad \forall B \in \mathcal{B}(\mathbb{R}^n), \tag{2.72}$$

where $\mathcal{B}(\mathbb{R}^n)$ denotes the Boreal σ-algebra on $\mathbb{R}^n$ (n.b., $\mathbf{x}_k$ is a random vector on $\mathbb{R}^n$).

Therefore,

$$\bar{\mathbb{P}}_{\mathbf{x}_k|\mathcal{Y}_{k-1}} = \mathcal{N}_n\left(\hat{\mathbf{x}}_{k|k-1}, \boldsymbol{M}_{k|k-1}\right), \tag{2.73}$$

where

$$\hat{\mathbf{x}}_{k|k-1} = \boldsymbol{F}_{k-1}\hat{\mathbf{x}}_{k-1|k-1} \tag{2.74}$$

and

$$\boldsymbol{M}_{k|k-1} = \boldsymbol{F}_{k-1}\boldsymbol{P}^*_{k-1|k-1}\boldsymbol{F}^\top_{k-1} + \boldsymbol{G}_{k-1}\boldsymbol{Q}_{k-1}\boldsymbol{G}^\top_{k-1}. \tag{2.75}$$

The nominal distribution of the measurement noise $\mathbf{v}_k$ is $\bar{\mathbb{P}}_{\mathbf{v}_k|\mathcal{Y}_{k-1}} = \bar{\mathbb{P}}_{\mathbf{v}_k} = \mathcal{N}_m(\mathbf{0}, \boldsymbol{R}_k)$ because $\mathbf{v}_k$ is independent of $\mathcal{Y}_{k-1}$.

Now it is sufficient to invoke the results in Theorem 8 to obtain the distributionally robust state estimate $\hat{\boldsymbol{x}}_{k|k}$ at time k given $\boldsymbol{y}_k$.

Theorem 11. *Suppose the radii of the ambiguity sets are $\epsilon \geq 0$, $\theta_{\boldsymbol{x},k} \geq 0$, $\theta_{2,\boldsymbol{x},k} \geq 1 \geq \theta_{1,\boldsymbol{x},k} \geq 0$, $\theta_{\boldsymbol{v},k} \geq 0$, $\theta_{2,\boldsymbol{v},k} \geq 1 \geq \theta_{1,\boldsymbol{v},k} \geq 0$. At the time k, with the nominal Gaussian prior conditional distribution of the state $\bar{\mathbb{P}}_{\mathbf{x}_k|\mathcal{Y}_{k-1}} \sim \mathcal{N}_n\left(\hat{\mathbf{x}}_{k|k-1}, \boldsymbol{M}_{k|k-1}\right)$ and the nominal Gaussian distribution of the measurement noise $\bar{\mathbb{P}}_{\mathbf{v}_k} = \mathcal{N}_m(\mathbf{0}, \boldsymbol{R}_k)$, the distributionally robust state estimator $\hat{\mathbf{x}}_{k|k}$ given $\mathbf{y}_k$ is as follows.*

1. *Optimal Estimator.*

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \boldsymbol{\Sigma}^*_{\boldsymbol{x},k}\boldsymbol{H}^\top_k \boldsymbol{S}^{*-1/2}_k \cdot \boldsymbol{\psi}[\boldsymbol{S}^{*-1/2}_k \mathbf{s}_k], \tag{2.76}$$

   where $\mathbf{s}_k := \mathbf{y}_k - \boldsymbol{H}_k\hat{\mathbf{x}}_{k|k-1}$, $\hat{\mathbf{x}}_{k|k-1} = \boldsymbol{F}_{k-1}\hat{\mathbf{x}}_{k-1|k-1}$, and $\boldsymbol{S}^*_k := \boldsymbol{H}_k\boldsymbol{\Sigma}^*_{\boldsymbol{x},k}\boldsymbol{H}^\top_k + \boldsymbol{\Sigma}^*_{\boldsymbol{v},k}$; $\boldsymbol{\psi}(\cdot)$, $\boldsymbol{\Sigma}^*_{\boldsymbol{x},k}$, and $\boldsymbol{\Sigma}^*_{\boldsymbol{v},k}$ are defined in Theorem 8.

2. *Worst-Case Estimation Error Covariance.*

$$\boldsymbol{P}^*_{k|k} = \boldsymbol{\Sigma}^*_{\boldsymbol{x},k} - \boldsymbol{\Sigma}^*_{\boldsymbol{x},k}\boldsymbol{H}^\top_k \boldsymbol{S}^{*-1}_k \boldsymbol{H}_k\boldsymbol{\Sigma}^*_{\boldsymbol{x},k} \cdot i^{\min}_\mu, \tag{2.77}$$

   where $i^{\min}_\mu$ is defined in Theorem 8.

Proof. Compare with Theorem 8. □

As we can see from (2.76), there exists an influence function $\boldsymbol{\psi}(\cdot)$ to limit the influence that a large-valued outlier may bring to the estimator.

The distributionally robust estimator to the linear Markov system (2.1) is summarized in Algorithm 2.2.

Algorithm 2.2: Distributionally Robust Estimator for Linear Systems Subject to Parameter Uncertainty and Measurement Outlier

---

Definition: $\hat{\mathbf{x}}_{k|k}$ as the distributionally robust state estimator and $\hat{\boldsymbol{x}}_{k|k}$ the robust state estimate when $\boldsymbol{y}_k$ is specified; $\boldsymbol{P}^*_{k|k}$ as the worst-case state estimation error covariance.

Initialize: $\hat{\boldsymbol{x}}_{0|0}$, $\boldsymbol{P}^*_{0|0}$, ϵ, all involved θ as instructed in Theorem 11 (i.e., $\theta_{\boldsymbol{x}}$ and $\theta_{\boldsymbol{v}}$ if we use the Wasserstein ambiguity sets, or $\theta_{2,\boldsymbol{x}}$ and $\theta_{2,\boldsymbol{v}}$ if we use the moment-based ambiguity sets).

Remark: According to Theorem 7, $\theta_{1,\boldsymbol{x}}$ and $\theta_{1,\boldsymbol{v}}$ are irrelevant to this algorithm, and therefore, not initialized. When $\mathbf{y}_k$ has a realization $\boldsymbol{y}_k$, the estimator of $\mathbf{x}_k$, i.e., $\hat{\mathbf{x}}_{k|k}$, gives an estimate $\hat{\boldsymbol{x}}_{k|k}$ to $\mathbf{x}_k$.

Input : $\boldsymbol{y}_k$, $k = 1, 2, 3, ...$

while *true* **do**
- *// Time-Update Step, i.e., Prior Estimation*
- Use (2.74) and (2.75) to obtain $\hat{\boldsymbol{x}}_{k|k-1}$ and $\boldsymbol{M}_{k|k-1}$
- *// Obtain the Nominal Distributions*
- Use (2.73) to obtain $\bar{\mathbb{P}}_{\mathbf{x}_k|\mathcal{Y}_{k-1}=\boldsymbol{Y}_{k-1}}$
- $\bar{\mathbb{P}}_{\mathbf{v}_k} \leftarrow \mathcal{N}_m(\mathbf{0}, \boldsymbol{R}_k)$
- *// Obtain the Worst-Case Scenario*
- Use (2.70) or (2.71) to obtain $i^{\min}_{\mu}$
- Use (2.62) or (2.64) to obtain $\boldsymbol{\Sigma}^*_{\boldsymbol{x},k}$ and $\boldsymbol{\Sigma}^*_{\boldsymbol{v},k}$
- *// Measurement-Update Step, i.e., Posterior Estimation*
- Use (2.76) and (2.77) to obtain $\hat{\boldsymbol{x}}_{k|k}$ and $\boldsymbol{P}^*_{k|k}$
- *// Next Time Step*
- $k \leftarrow k + 1$

end

Output : $\hat{\boldsymbol{x}}_{k|k}$

The theorem below reveals relations among the proposed distributionally robust estimator and the existing estimators.

Theorem 12. *Concerning the distributionally robust state estimator in Algorithm 2.2, the follows are true.*

1) If we set $\epsilon = 0$, $\theta_{\boldsymbol{x}} = \theta_{\boldsymbol{v}} = 0$, $\theta_{1,\boldsymbol{x}} = \theta_{2,\boldsymbol{x}} = 1$, $\theta_{1,\boldsymbol{v}} = \theta_{2,\boldsymbol{v}} = 1$, we obtain the canonical Kalman filter.

2) Under moment-based ambiguities, if we set $\epsilon = 0$, $\theta_{2,\boldsymbol{x}} = \theta_{2,\boldsymbol{v}}$, we obtain the fading Kalman filter [36, 72].

3) The Student's t Kalman filter in [93, Eq. (13)] amounts to a distributionally robust filter because it is a fading Kalman filter whose fading factor is adaptively changeable.

4) Under moment-based ambiguities, if we set $\epsilon = 0$, $\theta_{1,\boldsymbol{x}} = \theta_{2,\boldsymbol{x}} = 1$, we obtain the robust Kalman filter in [127, Eq. (32)] that has an adaptive $\theta_{2,\boldsymbol{v}}$.

5) Under ϵ-contamination ambiguity, if we set $\theta_{\boldsymbol{x}} = \theta_{\boldsymbol{v}} = 0$, $\theta_{1,\boldsymbol{x}} = \theta_{2,\boldsymbol{x}} = 1$, $\theta_{1,\boldsymbol{v}} = \theta_{2,\boldsymbol{v}} = 1$, we obtain the M-estimation-based Kalman filter [97, Thm. 3].

6) When there are no outliers and the special case discussed in Theorem 10 is considered, if we use the Wasserstein metric, we obtain the Wasserstein Kalman filter [91].

7) When there are no outliers and the special case discussed in Theorem 10 is considered, if we use the KL divergence, we obtain the relative-entropy Kalman filter [89].

8) When there are no outliers in measurements and the special case discussed in Theorem 10 is considered, if we use the τ-divergence, we obtain the τ-divergence Kalman filter [90].

9) The relative-entropy Kalman filter and the τ-divergence Kalman filter are risk-sensitive Kalman filters [89, 90].

Proof. In the case 1), all the ambiguity sets only contain nominal distributions. Hence, we have $\boldsymbol{\Sigma}^*_{\boldsymbol{x},k} = \boldsymbol{M}_{k|k-1}$, $\boldsymbol{\Sigma}^*_{\boldsymbol{v},k} = \boldsymbol{R}_k$, $\boldsymbol{\psi}(\mu) = \mu$, and $i^{\min}_{\mu} = 1$, leading to the canonical Kalman filter. In the case 2), if we assume $\theta = \theta_{2,\boldsymbol{x}} = \theta_{2,\boldsymbol{v}}$, we have $\boldsymbol{\Sigma}^*_{\boldsymbol{x},k} = \theta\boldsymbol{M}_{k|k-1}$, $\boldsymbol{\Sigma}^*_{\boldsymbol{v},k} = \theta\boldsymbol{R}_k$, $\boldsymbol{\psi}(\mu) = \mu$, and $i^{\min}_{\mu} = 1$, leading to $\boldsymbol{P}^*_{k|k} = \theta \cdot \boldsymbol{P}_{k|k}$ where $\boldsymbol{P}_{k|k} := \boldsymbol{M}_{k|k-1} - \boldsymbol{M}_{k|k-1}\boldsymbol{H}_k^\top(\boldsymbol{H}_k\boldsymbol{M}_{k|k-1}\boldsymbol{H}_k^\top + \boldsymbol{R}_k)^{-1}\boldsymbol{H}_k\boldsymbol{M}_{k|k-1}$. By comparing with [72], we obtain the fading Kalman filter. For other cases, compare with the given references. □

2.3.3 Computational Complexity

As we can see in Algorithm 2.2, the most computationally intensive step is to solve (2.62) to obtain the worst-case scenario (i.e., $\boldsymbol{\Sigma}^*_{\boldsymbol{x},k}$ and $\boldsymbol{\Sigma}^*_{\boldsymbol{v},k}$) under the Wasserstein ambiguity sets. Problem (2.62) is a SDP which is numerically challenging to solve. Instead, if we use the moment-based ambiguity sets, we need to solve (2.64) to obtain the worst-case scenario. However, (2.64) can be analytically solved by Theorem 7. As a result, all the steps in Algorithm 2.2 have closed-form solutions, implying that the computational complexity is no longer an issue. This is the reason why we adopted the moment-based ambiguity sets throughout experiments.

Let $r := \max\{n, p, m\}$ where n is the dimension of the state vector $\mathbf{x}_k$, p of the process noise vector $\mathbf{w}_k$, and m of the measurement vector $\mathbf{y}_k$. Since for a usual state estimation problem $n \geq p$ and $n \geq m$, it is well-known that the (asymptotic) computational complexity of the canonical Kalman filter at each time step is $\mathcal{O}(r^3) = \mathcal{O}(n^3)$; cf. Subsection 2.2.3. This is because all computational operations at each time step of the Kalman filtering are just matrix

addition/subtraction, matrix multiplicity, and matrix inverse; matrix multiplicity and matrix inverse operations admit cubic order of computational complexity in terms of the dimensions of the involved matrices. Therefore, likewise, the computational complexity of the proposed method is also $\mathcal{O}(n^3)$ at each time step, given that the moment-based ambiguity sets are used.

2.3.4 Comparisons with Existing Frameworks

Comparisons with existing frameworks addressing parameter uncertainties have been made in Subsection 2.2.5. In this subsection, we only discuss existing frameworks addressing measurement outliers.

When we unexpectedly see outliers in a nominal outlier-free population, we usually have two philosophies. The first one is that we no longer believe the nominal population is outlier-free. Instead, we take into account the outliers directly in modeling and correct the nominal distribution into an outlier-aware one. Typical solutions include: 1) direct modeling, e.g., t-distribution, Laplacian distribution; 2) indirect modeling, e.g., Bayesian methods (e.g., if the variance of a Gaussian distribution follows an inverse Gamma distribution, then the samples from this variance-variant Gaussian distribution would follow a t-distribution). The second one is that we still believe the population is outlier-free and treat seen outliers as aggressors to be cleared/modified. Typical solutions are reported, in particular, by Frequentists, e.g., the jackknife method.

The two philosophies can also be understood by leveraging the influence curve (a.k.a. influence function; see Appendix A.4) [55, 100, 128]. Two kinds of influence curves are well-studied:

1) infinite-rejection-point influence curves, including all the monotonic influence curves (e.g., Huber's [126]) and some re-descending influence curves that have infinite rejection points (e.g., maximum-correntropy-criterion [99, 129]).

2) finite-rejection-point influence curves, including some re-descending influence curves that have finite rejection points (e.g., Hampel's [55, 100], Tukey's Biweight [55], Andrew's Sine [55], IGG [130]).

When we use infinite-rejection-point influence curves, we implicitly accept outliers to be unstudied samples and correct the nominal distribution to be heavy-tailed. For example, the influence curve of an M-estimator at a t-distribution is a kind of re-descending influence curve but it has infinite rejection-point [21, Fig. 1]. Contrarily, when we adopt finite-rejection-point influence curves, we actually admit finite support of the nominal distribution and any sample outside of this support would be treated as intruders and trashed.

Most of the existing state estimation frameworks under measurement outliers belong to one of the two philosophies mentioned above. Note that in Bayesians, different from Frequentists, influence curves are imposed on innovation vectors (i.e., difference between true measurement and predicted measurement; cf. Theorem 8) rather than directly on measurement vectors; see,

e.g., [97]. Below lists and discusses some typically existing outlier-insensitive state estimation frameworks.

The earliest outlier-treatment method is the Gaussian-sum filter [20, 92], which uses heavy-tailed distributions for measurements, and the non-Gaussian heavy-tailed distributions are approximated by Gaussian sums. The demerit of this method is that it is computationally intensive and, thus, inefficient.

A remedy methodology to the Gaussian-sum filter is typically the t-distribution Kalman filter [93–95], which no longer uses a Gaussian sum to approximate the non-Gaussian measurement noise. Instead, it directly uses heavy-tailed non-Gaussian distributions such as the t-distribution, which explicitly explain the outliers. An indirect modeling trick is the Bayesian framework that assumes the noise statistics matrix (i.e., $\boldsymbol{R}$) is not exact and follows an inverse Wishart distribution so that the measurements $\mathbf{y}$ from the linear observation $\mathbf{y} = \boldsymbol{H}\mathbf{x} + \mathbf{v}$ would follow a multivariate t-distribution, which implicitly accounts for outliers [21].

Another remedy methodology is directly working on designing proper influence functions [99, 100], which is also known as the weighted-least-square M-estimation-based Kalman filter [71, 98]. For details, see Appendix A.4. In this category, the solutions for $\psi(\cdot)$ defined in Theorem 8 and Theorem 11 are particularly popular. Other possible influence functions are the maximum-correntropy-criterion (MCC) [129], IGG [130], Hampel's [55, 100], Tukey's Biweight [55], Andrew's Sine [55], etc. However, note that they are derived from other motivations and might no longer have clear perspectives of distributional robustness.

2.3.5 Experiments

In this section, we compare the state estimation performances of the existing filters and our newly proposed filter for linear systems subject to parameter uncertainties and measurement outliers. All the source data and codes are available online at GitHub: `https://github.com/Spratm-Asleaf/DRSE-Outlier`. Interested readers can reproduce and/or verify the claims in this section via changing the parameters or codes by themselves.

We continue studying the classical instance discussed in [27, 89, 91], 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 this state estimation problem, the

nominal system matrix is known as

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

Besides, we randomly add outliers for 5% measurements (i.e., we accordingly set $\epsilon = 0.05$ in the proposed method).

Candidate Filters

We implement the following filters to compare.

1. **TMKF**: the canonical Kalman filter with the true model. In the simulation we know the underlying true model $\boldsymbol{F}_k^{real}$ and the outlier-free true measurements. Therefore, this method theoretically gives the best estimate of state in the sense of minimum estimation error covariance;
2. **KF**: the canonical Kalman filter (with the nominal model $\boldsymbol{F}_k$);
3. **HKF**: the outlier-insensitive Kalman filter based on the Huber's influence function [97, 99];
4. **τ-KF**: the τ-divergence Kalman filter [90];
5. **WKF**: the Wasserstein Kalman filter [91];
6. **MKF**: The moment-based distributionally robust state estimator (see Theorem 11). We choose moment-based ambiguity sets because under them the problem is easier to solve (than that under Wasserstein ambiguity sets).

Parameters Setting

Algorithm 2.2 requires to initialize the parameters ϵ and θ's. Note that when ϵ is specified, K in (2.67), and a, b, and c in (2.68) will be uniquely determined; see Lemmas 1 and 2 and their proofs. Besides, if we use the Wasserstein ambiguity sets, we need to initialize $\theta_{\boldsymbol{x}}$ and $\theta_{\boldsymbol{v}}$ [see (2.61)]. If we use the moment-based ambiguity sets, we need to initialize $\theta_{2,\boldsymbol{x}}$ and $\theta_{2,\boldsymbol{v}}$ [see (2.65)] (n.b., Algorithm 2.2 is irrelevant to $\theta_{1,\boldsymbol{x}}$ and $\theta_{1,\boldsymbol{v}}$).

In all methods, we set the initial state estimate as $\hat{\boldsymbol{x}}_{0|0} = [0, 0]^\top$ and its corresponding estimation error covariance as $\boldsymbol{P}^*_{0|0} := \mathrm{diag}\{1, 1\}$, where $\mathrm{diag}\{\cdot\}$ denotes a diagonal matrix [27, 89, 91]. All parameters of each filter are directly taken from the original paper or tuned to perform (nearly) best for the studied instance when Δ_k randomly changes and $\alpha = 1$.

In the Huber-based outlier-insensitive Kalman filter, we use $K = 1.4$ [see (2.67)], because when ϵ is fixed to 0.05, K has to be 1.4 [cf. (2.57)]. In the τ-divergence Kalman filter [90], we set $\tau = 0$ (i.e., the τ-divergence filter specifies the Kullback-Leibler filter [89]), and the radius of the

ambiguity set as 1.5×10^{-4}. In the Wasserstein Kalman filter [91], the radius of the ambiguity set is set to 0.1. In the moment-based distributionally robust filter, we set $\theta_{2,\boldsymbol{x}} = \theta_{2,\boldsymbol{v}} = 1.02$, and $K = 1.4$. Namely, the influence function in (2.67) is used.

Suppose each simulation episode runs $T = 1000$ discrete-time steps. The overall estimation error of each episode is measured by the rooted mean square error (RMSE) as

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

where $x_{1,k}$ (resp. $x_{2,k}$) 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.

Results

Results are obtained by a laptop with 8G RAM and Intel(R) Core(TM) i7-8850H CPU @ 2.60GHz. We conduct the following three experiments, respectively. First, let Δ_k randomly take its value according to the uniform distribution from its support $\mathcal{U}$ at each step k, and let $\alpha = 1$. However, in this simulation, we do not add outliers in the measurements. The results are shown in Table 2.6. Second, let $\alpha = 0$ (i.e., there are no parameter uncertainties). Nevertheless, we add outliers for 5% measurements. The results are shown in Table 2.7. Third, both parameter uncertainties and measurement outliers are considered as above. The results are shown in Table 2.8.

Table 2.6: Results when $\alpha = 1$ but no outliers

Filter	RMSE	Avg Time	Filter	RMSE	Avg Time
TMKF	**3.25**	1.41e-5	τ-KF [90]	**9.90**	25.25e-5
KF	14.52	7.51e-6	WKF [91]	9.95	132.24e-5
HKF [97]	14.74	1.22e-5	MKF[Ours]	**9.91**	1.23e-5

Avg Time: Average Execution Time at each time step (seconds);

1e-5: 1×10^{-5}; **Note:** TMKF gives theoretically optimal solution.

From Tables 2.6, 2.7, and 2.8, the following observations can be outlined. When there only exist parameter uncertainties, the τ-divergence Kalman filter, the Wasserstein Kalman filter, and the proposed moment-based distributionally robust state estimator are relatively robust, while the Huber-based outlier-insensitive Kalman filter is not. In addition, the proposed moment-based distributionally robust state estimator is preferable since it is computationally efficient. When there only exist measurement outliers, the Huber-based outlier-insensitive Kalman filter is roughly optimal as expected. However, the τ-divergence Kalman filter and the Wasserstein Kalman filter perform badly, implying that they are not robust against measurement outliers.

Table 2.7: Results when $\alpha = 0$ and only outliers

Filter	RMSE	Avg Time	Filter	RMSE	Avg Time
TMKF	**7.64**	1.36e-5	τ-KF [90]	19.41	26.43e-5
KF	16.14	7.82e-6	WKF [91]	16.56	125.55e-5
HKF [97]	**7.70**	1.39e-5	MKF[Ours]	**8.19**	1.20e-5

See Table 2.6 for table notes.

Table 2.8: Results when $\alpha = 1$ and also outliers

Filter	RMSE	Avg Time	Filter	RMSE	Avg Time
TMKF	**3.23**	1.40e-5	τ-KF [90]	22.15	25.76e-5
KF	21.04	7.31e-6	WKF [91]	16.94	126.52e-5
HKF [97]	16.14	1.26e-5	MKF[Ours]	**11.72**	1.16e-5

See Table 2.6 for table notes.

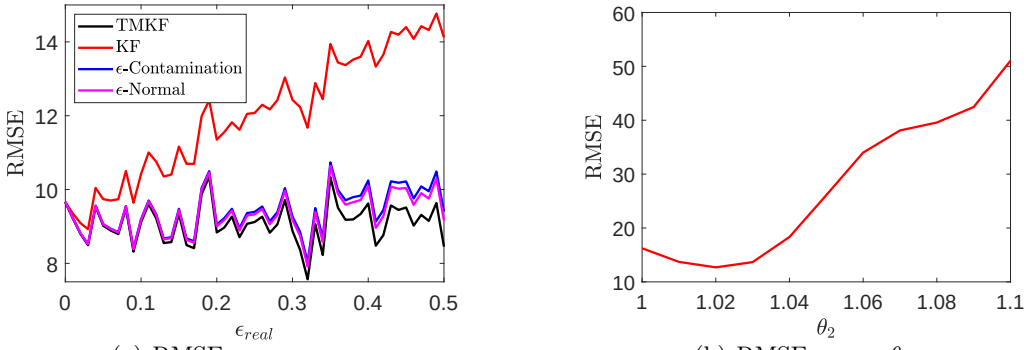
When both parameter uncertainties and measurement outliers exist, the proposed moment-based distributionally robust state estimator works better than other candidate filters; i.e., it is robust against both parameter uncertainties and measurement outliers. In Tables 2.6 and 2.8, the performances of the proposed method are far away from those of the TMKF because a relatively large uncertainty coefficient α is used (i.e., the true system model is far away from the nominal one). When α is set to be small, the difference will reduce (cf. Table 2.7). This reminds us that the robust filters are just remedial, but not once-for-all, solutions. In practice, continuing efforts need to be put on improving the accuracy of the nominal model, unless the model accuracy cannot be refined or robust solutions are satisfactory.

Sensitivity Analysis

In reality, it is hard to know the exact values of the true proportion of outliers (i.e., ϵ), and the true uncertainty level of the nominal model (i.e., $\theta_{\boldsymbol{x}}$, $\theta_{\boldsymbol{v}}$, $\theta_{2,\boldsymbol{x}}$, and $\theta_{2,\boldsymbol{v}}$). They cannot be learned to be optimal either because for a real system, the true state is unknown (i.e., training data set is unavailable). Hence, we need to investigate whether the proposed algorithm is sensitive to parameters ϵ and θ's, and explore the prior knowledge of tuning them for a real problem. Without loss of generality, we continue using the instance discussed above, where the moment-based ambiguity sets are adopted. As before, we set $\theta_{2,\boldsymbol{x}}$ and $\theta_{2,\boldsymbol{v}}$ to be the same, and $\theta_{2,\boldsymbol{x}} = \theta_{2,\boldsymbol{v}} := \theta_2$.

First, we let $\alpha = 0$ (i.e., no model uncertainty; specifically, no parameter uncertainty) and only study the sensitivity against the true proportion of outliers. For the case that we use the influence function in (2.67), we arbitrarily set $\epsilon = 0.01$ so that $K = 2$; for the case that we use the influence function in (2.68), we let $\epsilon = 0.03$ so that $a = 1.3496$, $b = 1.3496$, and $c = 1.2316$. Then, we let the real proportion of outliers ϵ_{real} change from 0 to 0.5. We have the results in Fig. 2.4 (a). It shows that the proposed method is not sensitive to ϵ_{real}. Thus, it is safe in practice to keep the values of ϵ, K, a, b, and c recommended above regardless of ϵ_{real}. (Other values are also viable; readers can validate this claim using the shared source codes themselves.) Besides, we show the breakdown properties of all the candidate filters. The results are shown in Fig. 2.5. We see that the HKF is better than the MKF when there are no parameter uncertainties [cf. Fig. 2.5 (a)], whereas the HKF is worse than the MKF when there exist parameter uncertainties [cf. Fig. 2.5 (b)]. This is because the MKF is the robustified version of the HKF against general model uncertainties (n.b., the MKF reduces to the HKF when $\theta_2 := 0$). Therefore, the price of the robustness in uncertain conditions (when $\alpha \neq 0$) is sacrificing the optimality in perfect conditions (when $\alpha = 0$).

Second, we fix $\epsilon_{real} = 0.05$ and study the sensitivity against the true degree of the model uncertainty. We let $\alpha = 1$, and θ_2 change from 1 to 1.1. We have the results in Fig. 2.4 (b). It shows that the performance of the proposed method depends heavily on the value of θ_2. If θ_2 is too small, the algorithm has no sufficient robustness against the uncertainty. Contrarily, if θ_2 is too large, the algorithm is too conservative to obtain a good performance as well. Therefore, one should carefully (and pragmatically) tune this parameter to achieve good performances for their specific real problems.

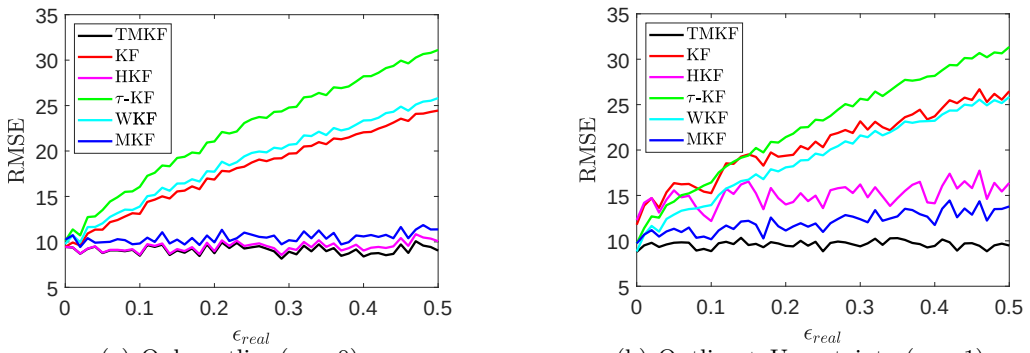


(a) RMSE versus ϵ_{real} (b) RMSE versus θ_2

Figure 2.4: Sensitivity results over ϵ_{real} and θ_2.

Student's t-Distributed Measurement Noise

In this subsection, we investigate the performances of the candidate filters for Student's t-distributed measurement noises. The degree of freedom of the used Student's t-distribution is set to be 3. But the covariance of measurement noise $\mathbf{v}_k$ at each time step is kept unchanged as $\boldsymbol{R}_k$. Note that although the variance σ^2 of a t-distribution is determined by its degree of freedom ν through $\sigma^2 = \frac{\nu}{\nu-2}$ for $\nu \geq 3$, it can be scaled by constant coefficients. For example,



(a) Only outlier ($\alpha = 0$) (b) Outlier + Uncertainty ($\alpha = 1$)

Figure 2.5: Breakdown test against ϵ_{real} with and without parameter uncertainty.

supposing a random variable t follows a t-distribution with degree of freedom ν, the variance of the transformed random variable $\frac{1}{\sqrt{\frac{\nu}{\nu-2}}}$t is unit.

Parameters of the candidate filters are tuned to perform best, respectively, for this new instance. (Details can be found in the shared source codes.) The results when only outliers exist are shown in Table 2.9, while those when both parameter uncertainties and outliers exist are shown in Table 2.10. Note that in this case, the TMKF designed for Gaussian-noise models is no longer optimal for the t-noise true model (i.e., it reduces to the KF when $\alpha = 0$).

Table 2.9: Results when $\alpha = 0$ and only outliers (t-distributed)

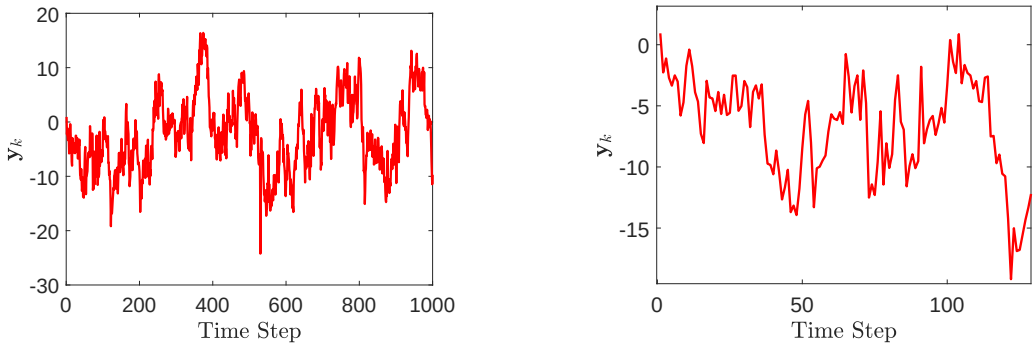
Filter	RMSE	Avg Time	Filter	RMSE	Avg Time
TMKF	**6.38**	1.58e-5	τ-KF [90]	7.07	26.00e-5
KF	**6.38**	1.18e-5	WKF [91]	**6.80**	121.51e-5
HKF [97]	**6.38**	1.66e-5	MKF[Ours]	**6.72**	1.31e-5

Table 2.10: Results when $\alpha = 1$ and also outliers (t-distributed)

Filter	RMSE	Avg Time	Filter	RMSE	Avg Time
TMKF	**3.53**	1.16e-5	τ-KF [90]	8.73	22.62e-5
KF	13.40	0.85e-5	WKF [91]	**8.34**	115.05-5
HKF [97]	13.75	1.25e-5	MKF[Ours]	**8.40**	1.28e-5

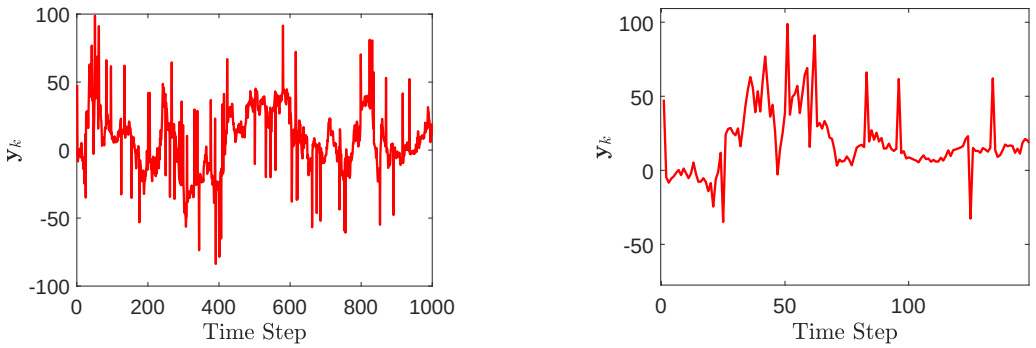
As we can see, when there are no parameter uncertainties (see Table 2.9), the TMKF, KF, and HKF have the same performance, and the τ-KF, WKF, and MKF perform worse than them. In other words, the Huber-based outlier-insensitive filter (HKF) no longer has advantage over the KF. This is because the measurements subject to t-distributed measurement noises do not have

significantly outstanding outliers; see Fig. 2.6. In contrast, in Fig. 2.7, we added significantly outstanding outliers. The two cases are all common in signal processing practices. Therefore, the outlier-robust methods (i.e., HKF and MKF) are more suitable for the cases where outliers significantly exist. (But an estimator that is suitable/optimal to t-distributed measurement noises might no longer be robust to other types of outliers, e.g., large-valued outliers.) Again, we see the price of robustness under uncertain conditions is sacrificing the optimality under perfect conditions because the MKF has larger RMSE than the HKF when there are no parameter uncertainties.



(a) Measurements (b) Closeup of (a)

Figure 2.6: Measurements contaminated by t-distributed noises.



(a) Measurements (b) Closeup of (a)

Figure 2.7: Measurements contaminated by significant outliers.

When there exist parameter uncertainties (see Table 2.10), as expected, the τ-KF, WKF, and MKF perform better because they are relatively robust against uncertainties. In this case, the MKF has smaller RMSE than the HKF, which verifies the claim that the sacrifice of optimality under perfect conditions might offer the robustness under uncertain conditions.

2.3.6 Section Conclusions

This section proposes the distributionally robust state estimation method that can account for both parameter uncertainties and measurement outliers. It offers a new perspective to

understand the robust state estimation problem under parameter uncertainties and measurement outliers and generalizes several classic methods into a unified framework. It uses only a few scalars to describe parameter uncertainties and measurement outliers and does not require structural information of uncertainties, especially useful when we have limited trust towards the nominal model and scarce knowledge about the uncertainties. Experiments show that the proposed method under moment-based ambiguity sets outperforms existing methods, which is not hard to expect because none of them is designed to simultaneously address both parameter uncertainties and measurement outliers. Although the method might be insensitive to the true proportion of outliers (i.e., the value of ϵ used in the algorithm does not significantly matter), it is sensitive to the true uncertainty level of the nominal model (i.e., the values of θ's used in the algorithm significantly matter). Practitioners have to carefully try appropriate θ's for their specific problems (n.b., θ's cannot be learned because the true state is unavailable). At last, three closing remarks need to be outlined.

1) Robust filters are just remedial solutions. Reducing modeling uncertainties is always important. Readers should not expect that the proposed method is optimal or satisfactory in all scenarios, e.g., for a model with t-distributed measurement noises (which implies that the true model is known).

2) The proposed outlier-robust filtering frameworks that use influence functions in Theorem 8 are more suitable for the cases that measurements contain significantly outstanding outliers and for the case that the measurement noise models are unknown. If measurement noises are t-distributed, it means that the system model is exactly known so that we can derive optimal filters for t-distributed noises (theoretically, this is still possible no matter whether the mathematical derivation is easy or not; cf. [95]). However, a filter that is optimal (or suitable) for t-distributed noise is likely to lose robustness for other types of noises.

3) The robustness under uncertain conditions comes with the cost of sacrificing the optimality under perfect conditions.

Chapter 3

State Estimation for Nonlinear Systems

From Theorem 5, Problem (2.60) subject to (2.61), and Problem (2.64) subject to (2.65), we are motivated to find worst-case prior state distributions and worst-case likelihood distributions (i.e., worst-case conditional measurement distributions given prior states) in robust Bayesian estimation settings; cf. (1.7), (2.35), and (2.54). To be specific, Theorem 5 shows that by raising variances of nominal Gaussian prior state distributions and nominal Gaussian likelihood distributions, the robust state estimation can be obtained. Intuitively, we recall the Bayesian posterior estimation principle:

$$p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x}) \cdot p(\boldsymbol{x}).$$

Thus, if we trust more the prior distribution $p(\boldsymbol{x})$ and doubt the likelihood distribution $p(\boldsymbol{y}|\boldsymbol{x})$, we should let $p(\boldsymbol{y}|\boldsymbol{x})$ be noninformative/uncertain and make main use of $p(\boldsymbol{x})$. Conversely, we should let $p(\boldsymbol{x})$ be noninformative/uncertain and mainly utilize $p(\boldsymbol{y}|\boldsymbol{x})$.

A natural measure of "uncertainty" of a random variable is entropy [131, 132]. A large entropy value implies that the distribution is not concentrated, and instead scattered/flat; i.e., we are less certain about happenings. Therefore, an eligible method to quantify the "worst-case" distribution is to use entropy; it is due to the principle of maximum entropy: "the probability distribution which best represents the current state of knowledge about a system is the one with largest entropy" [133]. Coincident with the implication of Theorem 5, the principle of maximum entropy is also popular in robust Bayesian methods [134–136], especially in choosing robust prior distributions [135, 137].[1] In this chapter, we utilize classical statistical metrics and divergences, such as Wasserstein metrics and Kullback-Leibler divergence, to construct balls containing a family of distributions near a nominal state prior distribution or near a nominal likelihood distribution. Then, we find maximum entropy distributions in the balls to generate new prior state particles and/or update their weights, and to evaluate the worst-case likelihoods of these prior state particles. As a result, the worst-case posterior state particles are immediate to be obtained by particle filters.

[1] According to [135, p.229], flat-tailed priors and noninformative priors can robustify a Bayesian statistical method. In fact, maximum-entropy distributions tend to be flat-tailed because maximizing the entropy of a variable (n.b., not fixed) distribution admits minimizing the Kullback–Leibler divergence of this distribution from a uniform distribution [cf. (3.38)], and uniform distributions are most flat-tailed. Moreover, maximum-entropy distributions also tend to be noninformative [136, Section 2.3].

As we can see, this chapter does not describe robustness from the min-max perspective as in Chapter 2. Instead, we directly pursue the original definition of robustness: a solution method is robust to a model if this method is not sensitive to small model perturbations [100, 128, 138]. Note that although the min-max robustness is a popular perspective, it is not the unique one.

3.1 Problem Formulation

Motivated by (1.5), we consider a nonlinear-system state estimation problem

$$\begin{cases} \mathbf{x}_k &= \boldsymbol{f}_k(\mathbf{x}_{k-1}, \mathbf{w}_{k-1}) \\ \mathbf{y}_k &= \boldsymbol{h}_k(\mathbf{x}_k, \mathbf{v}_k) \end{cases} \tag{3.1}$$

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

The first issue is that the nominal nonlinear system model (3.1) might be uncertain; recall Introduction 1.1 for motivating details. Specifically, for given k, at least one of the process dynamics function $\boldsymbol{f}_k(\cdot,\cdot)$, measurement dynamics function $\boldsymbol{h}_k(\cdot,\cdot)$, and types and/or parameters of distributions of $\mathbf{w}_k$ and $\mathbf{v}_k$ might be inexact. In designing robust state estimation solutions that are insensitive to these uncertainties, the challenge is quantifying and bounding such modeling uncertainties. Special cases when (3.1) takes linear forms have been discussed in Chapter 2. In this chapter, we exclusively investigate non-degenerate nonlinear cases. As measurements $\mathbf{y}_k$ sequentially arrive, we focus on a time-incremental state estimation problem: studying the problem at every k given the measurement set $\mathcal{Y}_k$ [46, 47]. Hence, it suffices to examine the following single-stage Bayesian inference problem:

$$\begin{cases} \mathbf{z} &\sim \mathbb{P}_{\mathbf{x}_{k-1}|\mathcal{Y}_{k-1}} \\ \mathbf{x} &= \boldsymbol{f}(\mathbf{z}, \mathbf{w}) \\ \mathbf{y} &= \boldsymbol{h}(\mathbf{x}, \mathbf{v}) \end{cases} \tag{3.2}$$

where $\mathbf{z}$ represents the conditional posterior state at $k-1$ given the past measurement set $\mathcal{Y}_{k-1}$, $\mathbf{x} := \mathbf{x}_k$ the state at k, $\mathbf{y} := \mathbf{y}_k$ the measurement at k, $\mathbf{w} := \mathbf{w}_{k-1}$ the process noise at $k-1$, and $\mathbf{v} := \mathbf{v}_k$ the measurement noise at k; nominal distributions $\mathbb{P}_{\mathbf{z}}$, $\mathbb{P}_{\mathbf{w}}$, and $\mathbb{P}_{\mathbf{v}}$ are known; nominal nonlinear dynamics functions $\boldsymbol{f}(\cdot,\cdot)$ and $\boldsymbol{h}(\cdot,\cdot)$ are known as well. The time index k is dropped

to avoid notational clutter. Note that $\mathbf{z}$ is random due to the randomness of $\mathcal{Y}_{k-1}$, but it is non-random in terms of $\mathbf{x}$ and $\mathbf{y}$; whenever $\mathcal{Y}_{k-1} = \boldsymbol{Y}_{k-1}$ is specified, $\mathbf{z} = \boldsymbol{z}$ becomes deterministic. In particle filters, all involved distributions $\mathbb{P}_{\mathbf{z}}$, $\mathbb{P}_{\mathbf{w}}$, $\mathbb{P}_{\mathbf{v}}$, $\mathbb{P}_{\mathbf{x}}$, and $\mathbb{P}_{\mathbf{y}}$ are represented/approximated by particles; they are discrete. Specifically, e.g., $p(\boldsymbol{z}) := \sum_{i=1}^{N_{\boldsymbol{z}}} u_{\boldsymbol{z}^i} \cdot \delta_{\boldsymbol{z}^i}(\boldsymbol{z})$ where $N_{\boldsymbol{z}}$ is the number of particles; particles $\boldsymbol{z}^i$ are sampled from $\mathbb{P}_{\mathbf{z}}$ and $u_{\boldsymbol{z}^i}$ are weights. Since uncertain process dynamics (resp. uncertain measurement dynamics) would let the true prior state distribution $\mathbb{P}_{\mathbf{x}}$ (resp. true likelihood distribution $\mathbb{P}_{\mathbf{y}|\mathbf{x}}$) deviate from the nominal prior state distribution (resp. nominal likelihood distribution), particle filters can be robustified by considering that prior state distributions (resp. likelihood distributions) are uncertain, and finding worst-case state priors (resp. likelihoods). To be specific, we propose to find the worst-case prior state distribution near the nominal prior state distribution $\bar{\mathbb{P}}_{\mathbf{x}}$ to generate worst-case prior state particles $\boldsymbol{x}^j$. Likewise, worst-case likelihood distributions near the nominal ones $\bar{\mathbb{P}}_{\mathbf{y}|\boldsymbol{x}^j}$ are leveraged to evaluate the worst-case likelihoods of prior state particles $\boldsymbol{x}^j$ at the measurement $\boldsymbol{y}$. The principle of maximum entropy supports us to explore and exploit the maximum entropy distribution when given limited information. Since the limited (i.e., inexact) prior state information is conveyed in $\bar{\mathbb{P}}_{\mathbf{x}}$, the following optimization problem has to be solved:

$$\begin{aligned} \max_{p(\boldsymbol{x})\in L^1} \quad & \int -p(\boldsymbol{x}) \ln p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ s.t. \quad & \begin{cases} D(\mathbb{P}_{\mathbf{x}}, \bar{\mathbb{P}}_{\mathbf{x}}) & \leq \theta \\ \int p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} & = 1 \end{cases} \end{aligned} \tag{3.3}$$

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

$$\begin{aligned} \max_{p(\boldsymbol{x})\in l^1} \quad & \sum_{j=1}^{M} -p(\boldsymbol{x}^j) \ln p(\boldsymbol{x}^j) \\ s.t. \quad & \begin{cases} D(\mathbb{P}_{\mathbf{x}}, \bar{\mathbb{P}}_{\mathbf{x}}) & \leq \theta \\ \sum_j p(\boldsymbol{x}^j) & = 1. \end{cases} \end{aligned} \tag{3.4}$$

Note that the support sets of the uncertain priors $p(\boldsymbol{x}) := \sum_{j=1}^{M} u_{\boldsymbol{x}^j} \delta_{\boldsymbol{x}^j}(\boldsymbol{x})$ and the nominal prior $\bar{p}(\boldsymbol{x}) := \sum_{i=1}^{N} u_{\boldsymbol{x}^i} \delta_{\boldsymbol{x}^i}(\boldsymbol{x})$ may not be the same: $\bar{p}(\boldsymbol{x})$ is supported on $\{\boldsymbol{x}^i\}$ for $i \in [N]$, while $p(\boldsymbol{x})$ is supported on $\{\boldsymbol{x}^j\}$ for $j \in [M]$. We call $\{\boldsymbol{x}^i\}_{i\in[N]}$ the nominal prior state particles and $\{\boldsymbol{x}^j\}_{j\in[M]}$ the worst-case prior state particles. Suppose that $p^*(\boldsymbol{x})$ solves (3.3). The worst-case prior state particles $\boldsymbol{x}^j$ can be sampled from $p^*(\boldsymbol{x})$. If $p^*(\boldsymbol{x})$ solves (3.4), $\boldsymbol{x}^j$ are worst-case prior state particles whose weights are $p^*(\boldsymbol{x}^j)$, respectively.

The second issue is evaluating likelihoods of particles $\boldsymbol{x}^j$ given the measurement $\boldsymbol{y}$. When measurement noises are additive [i.e., $\mathbf{y} = \boldsymbol{h}(\mathbf{x}) + \mathbf{v}$] or multiplicative [i.e., $\mathbf{y} = \boldsymbol{h}(\mathbf{x}) \cdot \mathbf{v}$],

the evaluation method is straightforward. For this reason, virtually all of the existing state-estimation literature tacitly takes the premise of additive/multiplicative measurement noises, which, however, is not always tenable in practice. This chapter, therefore, also aims to study a likelihood evaluation method for a general nonlinear measurement dynamics. The philosophy handling inexact likelihood information conveyed in nominal likelihood distributions $\bar{p}(\boldsymbol{y}|\boldsymbol{x}^j) := \sum_{r=1}^{R} u_{\boldsymbol{y}^r|\boldsymbol{x}^j}\delta_{\boldsymbol{y}^r|\boldsymbol{x}^j}(\boldsymbol{y}), \forall j \in [M]$ keeps consistent. Specifically, for every j, we need to solve

$$\begin{aligned} \max_{p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y})\in L^1} \quad & \int -p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}) \ln p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \\ s.t. \quad & \begin{cases} D(\mathbb{P}_{\mathbf{y}|\boldsymbol{x}^j}, \bar{\mathbb{P}}_{\mathbf{y}|\boldsymbol{x}^j}) & \leq \theta \\ \int p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} & = 1 \end{cases} \end{aligned} \tag{3.5}$$

or its discrete version

$$\begin{aligned} \max_{p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y})\in l^1} \quad & \sum_{t=1}^{T} -p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}^t) \ln p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}^t) \\ s.t. \quad & \begin{cases} D(\mathbb{P}_{\mathbf{y}|\boldsymbol{x}^j}, \bar{\mathbb{P}}_{\mathbf{y}|\boldsymbol{x}^j}) & \leq \theta \\ \sum_t p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}^t) & = 1. \end{cases} \end{aligned} \tag{3.6}$$

Likewise, the support sets of the uncertain likelihood distributions $p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}) := \sum_{t=1}^{T} u_{\boldsymbol{y}^t|\boldsymbol{x}^j}\delta_{\boldsymbol{y}^t|\boldsymbol{x}^j}(\boldsymbol{y})$ and the nominal likelihood distributions $\bar{p}_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}) := \sum_{r=1}^{R} u_{\boldsymbol{y}^r|\boldsymbol{x}^j}\delta_{\boldsymbol{y}^r|\boldsymbol{x}^j}(\boldsymbol{y})$ may not be the same: $p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y})$ is supported on $\{\boldsymbol{y}^t\}$ for $t \in [T]$, while $\bar{p}_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y})$ is supported on $\{\boldsymbol{y}^r\}$ for $r \in [R]$. We call $\{\boldsymbol{y}^r\}_{r\in[R]}$ the nominal likelihood particles and $\{\boldsymbol{y}^t\}_{t\in[T]}$ the worst-case likelihood particles. Suppose that $p^*_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y})$ solves (3.5). The worst-case likelihood of the prior state particle $\boldsymbol{x}^j$ given the measurement $\boldsymbol{y}$ can be evaluated by $p^*_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y})$. Instead, if $p^*_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y})$ solves (3.6) and one of $\boldsymbol{y}^t$ is the same as the given measurement $\boldsymbol{y}$, the worst-case likelihood of the prior state particle $\boldsymbol{x}^j$ given the measurement $\boldsymbol{y}$ can be evaluated by $p^*_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}^t)$. This is possible because we can let the collected $\boldsymbol{y}$ be a supporting point to solve (3.6); i.e., $\boldsymbol{y} \in \{\boldsymbol{y}^t\}_{t\in[T]}$. Note that the support set $\{\boldsymbol{y}^t\}_{t\in[T]}$ is specified by filter designers.

The third issue is to identify possible outliers in measurements and take actions to remove or attenuate them [58]. Motivated by the M-estimation theory [55], we claim that this can be done by evaluating the likelihoods of prior state particles at the given measurement: if the largest likelihood of the prior state particles is smaller than a threshold (e.g., 5%), we treat this measurement as an outlier because there exists no any prior state particle that can possibly generate this measurement. Then, this measurement can be directly trashed and all prior state particles directly become posterior (cf. re-descending influence functions, e.g., Hampel's [55, Eq. (4.90)], in M-estimation). This measurement can also be replaced by the nearest likelihood particle generated by the prior state particle that has the largest likelihood (cf. monotonic influence functions, e.g., Huber's [55, Eq. (4.53)], in M-estimation).

To the core, this chapter needs to find solutions of (3.3), (3.4), (3.5), and (3.6). In the following

sections, we first explicitly choose eligible forms of the statistical similarity measure $D(\cdot,\cdot)$. Then, we find maximum entropy distributions for generating worst-case prior state particles and evaluating their worst-case likelihoods. Third, we identify and handle measurement outliers. At last, the overall distributionally robust state estimation framework for nonlinear systems will be outlined.

3.2 Find Maximum Entropy Distributions

Mathematically, (3.3) and (3.5) are the same problem, so are (3.4) and (3.6). The former is a maximum entropy problem for a continuous distribution family given a discrete reference distribution, while the latter is a maximum entropy problem for a discrete distribution family given a discrete reference distribution. Therefore, for notation simplicity, we investigate a unified form for (3.3) and (3.5):

$$\begin{aligned} \max_{p(\boldsymbol{x})\in L^1} \quad & \int -p(\boldsymbol{x})\ln p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ s.t. \quad & \begin{cases} D[p(\boldsymbol{x}),q(\boldsymbol{x})] & \leq \theta \\ \int p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} & = 1 \end{cases} \end{aligned} \tag{3.7}$$

where $q(\boldsymbol{x}) = \sum_{i=1}^{N} u_{\boldsymbol{x}^i}\delta_{\boldsymbol{x}^i}(\boldsymbol{x})$ is a N-point discrete reference distribution whose probability measure is $\mathbb{Q}_{\mathbf{x}}$. Likewise, supposing $p(\boldsymbol{x}) = \sum_{j=1}^{M} u_{\boldsymbol{x}^j}\delta_{\boldsymbol{x}^j}(\boldsymbol{x})$ is a M-point discrete distribution whose probability measure is $\mathbb{P}_{\mathbf{x}}$, the unified form for (3.4) and (3.6) is

$$\begin{aligned} \max_{\boldsymbol{p}\in l^1} \quad & \sum_{j=1}^{M} -p_j \ln p_j \\ s.t. \quad & \begin{cases} D[\boldsymbol{p},\boldsymbol{q}] & \leq \theta \\ \sum_{j=1}^{M} p_j & = 1 \end{cases} \end{aligned} \tag{3.8}$$

where $p_j := u_{\boldsymbol{x}^j}$, $q_i := u_{\boldsymbol{x}^i}$, $\boldsymbol{p} := [p_1, p_2, ..., p_j, ...p_M]^\top$, and $\boldsymbol{q} := [q_1, q_2, ..., q_i, ...q_N]^\top$. In (3.8), M might be equal to N but this is not always the case. Besides, even when $M = N$, $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{Q}_{\mathbf{x}}$ can be supported on different discrete points; the former is $\{\boldsymbol{x}^j\}_{j=1,2,...,M}$ and the latter is $\{\boldsymbol{x}^i\}_{i=1,2,...,N}$.

Therefore, it suffices to consider only (3.7) and (3.8) in this section. In state-of-the-art distributionally robust optimization literature, the most commonly adopted statistical similarity measures are moments-based similarity [120, 139], Wasserstein distance [56], and ϕ-divergence [125]. We will find the solutions to (3.7) and (3.8) based on these three statistical similarity measures, respectively. In the following sections, to avoid notational clutter, we no longer emphasize that a density $p(\cdot) \in L^1$ or a mass $p(\cdot) \in l^1$; they are implicitly admitted instead.

3.2.1 Solutions Using Moments-Based Similarity

Moments-Based statistical similarity claims that two random vectors are similar (in distribution) if they have similar moments up to the order of O (e.g., when $O = 2$, two random vectors have the same mean and covariance). This measure is also widely used in information theory [132, Chapter 11]. The maximum entropy solutions to (3.7) and (3.8) using moments-based similarity are, therefore, hardly new. We repeat them because from which we can cast new insights into Gaussian approximation state estimators. Note that the moments of the discrete reference distribution $\mathbb{Q}_{\mathbf{x}}$ can be estimated from its particles (using any eligible approaches, e.g., weighted sample mean and weighted sample covariance). Suppose the first two sample moments of $\mathbb{Q}_{\mathbf{x}}$ are given by $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}} := \sum_{i=1}^{N} u_{\boldsymbol{x}^i} \cdot \boldsymbol{x}^i$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} := \sum_{i=1}^{N} u_{\boldsymbol{x}^i} \cdot (\boldsymbol{x}^i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(\boldsymbol{x}^i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})^\top$, respectively.

Solution to (3.7)

The theorem below gives the continuous maximum entropy distribution when the first two moments are specified.

Theorem 13 (Theorem 9.6.5 [132])**.** *If the first two moments of an absolutely continuous distribution $\mathbb{P}_{\mathbf{x}}$ are $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$, respectively, then the maximum entropy of $\mathbb{P}_{\mathbf{x}}$ is obtained by a Gaussian with mean $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$.*

Proof. See [132, Theorem 9.6.5] or [140, Theorem 4.1.2]. Note that a Gaussian distribution is translation-invariant, and absolute continuity of $\mathbb{P}_{\mathbf{x}}$ implies the existence of its density almost everywhere. □

Theorem 13 can be extended to take into account higher order moments; see [132, Section 11.1]. We do not consider moments with orders equal to or higher than 3 because they are tensors for multivariate problems, and they are unnecessary for this thesis's contexts. Theorem 13 reveals the distributional robustness of the Gaussian approximation state estimation framework.

Corollary 4. *The Gaussian approximation state estimation framework for nonlinear systems is distributionally robust in the sense that it uses maximum entropy distributions for prior states and their likelihoods.* □

Corollary 4 implies that when the nominal nonlinear system model is uncertain, Gaussian approximation filters, such as Unscented Kalman filter (UKF), Cubature Kalman filter (CKF), and Ensemble Kalman filter (EnKF) might outperform general particle filters. The benefit of such Gaussian approximation is that the induced filters (UKF, CKF, EnKF, etc.) are, strictly speaking, no longer computationally-intensive sequential Monte Carlo methods because they do not store prior state particles and explicitly evaluate their likelihoods. Instead, states and measurements are assumed to be marginally Gaussian and also jointly Gaussian, and therefore, closed-form solutions (i.e., canonical Kalman iterations) are applicable, which are computationally attractive.

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

Solution to (3.8)

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

Theorem 14. *Among all discrete distributions supported on $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$ with first two moments $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$, the maximum entropy distribution is*

$$p_j = \exp\{-1-\gamma-\boldsymbol{\lambda}^\top \boldsymbol{x}^j - (\boldsymbol{x}^j-\hat{\boldsymbol{\mu}}_{\boldsymbol{x}})^\top \boldsymbol{\Lambda}^\top (\boldsymbol{x}^j-\hat{\boldsymbol{\mu}}_{\boldsymbol{x}})\}, \tag{3.9}$$

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

$$\begin{cases} \sum_{j=1}^M p_j &= 1, \\ \sum_{j=1}^M \boldsymbol{x}^j \cdot p_j &= \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}, \\ \sum_{j=1}^M (\boldsymbol{x}^j-\hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(\boldsymbol{x}^j-\hat{\boldsymbol{\mu}}_{\boldsymbol{x}})^\top \cdot p_j &= \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}. \end{cases} \tag{3.10}$$

Proof. Applying the Lagrange multiplier method to (3.8), the statements are immediate. □

Theorem 14 gives the worst-case weights of particles $\boldsymbol{x}^j$; i.e., $u_{\boldsymbol{x}^j} = p_j$. Therefore, particles $\boldsymbol{x}^j$ together with their weights $u_{\boldsymbol{x}^j}$ represent a worst-case prior state distribution [cf. (3.4)] or a worst-case likelihood distribution [cf. (3.6)]. The nonlinear root-finding problem (3.10) is, however, complicated even when only the first two moments are considered and only equalities are involved. If higher order moments and inequalities exist in (3.8), the complexity would be inconceivable (due to, e.g., tensors). However, the solution of (3.10) is just theoretically meaningful. In practice, when we take Gaussian assumption, it is pointless to store particles and evaluate their likelihoods; we prefer to apply canonical closed-form Kalman iterations.

3.2.2 Solutions Using Wasserstein Distance

The definition of Wasserstein distance can be revisited in (2.27). The benefit to use Wasserstein distance is that it does not require the two involved distributions to have the same support. In other words, it is possible that either $\mathbb{P}_{\mathbf{x}}$ or $\mathbb{Q}_{\mathbf{x}}$ is continuous and the other one is discrete. Besides, the Wasserstein distance can also implicitly take higher-order-moment information of random vectors into consideration, unlike the Gaussian assumption that only focuses on the first two moments. In this section, as claimed, $\mathbb{Q}_{\mathbf{x}}$ is discrete and supported on $\{\boldsymbol{x}^i\}_{i=1,2,\ldots,N}$.

Solution to (3.7)

Let $\Pi_{\mathbf{a},\mathbf{b}}$ denote any possible product measures (i.e., joint distributions) whose marginals are $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{Q}_{\mathbf{x}}$; $\mathbf{a}$ is the random vector associated with $\mathbb{P}_{\mathbf{x}}$, while $\mathbf{b}$ is with $\mathbb{Q}_{\mathbf{x}}$. Suppose $\mathbb{P}_{\mathbf{x}}$ and $\Pi_{\mathbf{a},\mathbf{b}}$ are absolutely continuous, and the density of $\Pi_{\mathbf{a},\mathbf{b}}$ is $\pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}})$; $\pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}) = I(\boldsymbol{x}_{\mathbb{Q}}|\boldsymbol{x}_{\mathbb{P}})p(\boldsymbol{x}_{\mathbb{P}})$ where $I(\boldsymbol{x}_{\mathbb{Q}}|\boldsymbol{x}_{\mathbb{P}})$ is the conditional density. We solve (3.7) using the Wasserstein distance. Hence, (3.7) can be written as

$$\begin{aligned} \max_{p(\boldsymbol{x})} \quad & \int -p(\boldsymbol{x}) \ln p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ s.t. \quad & \begin{cases} \inf_{\pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}) \mathrm{d}\boldsymbol{x}_{\mathbb{P}} \mathrm{d}\boldsymbol{x}_{\mathbb{Q}} & \leq \theta \\ \int p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} & = 1. \end{cases} \end{aligned} \tag{3.11}$$

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

We first study the constraint $\inf_{\pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}) \mathrm{d}\boldsymbol{x}_{\mathbb{P}} \mathrm{d}\boldsymbol{x}_{\mathbb{Q}} \leq \theta$. The infimum optimization problem on the left hand side of this constraint is functional and infinite-dimensional. Therefore, we aim to transform it into a vector-valued and finite-dimensional equivalent.

Lemma 3. *The infinite-dimensional optimization problem*

$$\inf_{\pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \pi(\boldsymbol{x}_{\mathbb{P}}, \boldsymbol{x}_{\mathbb{Q}}) \mathrm{d}\boldsymbol{x}_{\mathbb{P}} \mathrm{d}\boldsymbol{x}_{\mathbb{Q}}$$

is equivalent to a finite-dimensional optimization problem

$$\max_{\boldsymbol{\lambda}} \int p(\boldsymbol{x}) \min_{i \in [N]} \{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\} \mathrm{d}\boldsymbol{x} + \sum_{i=1}^{N} q_i \lambda_i, \tag{3.12}$$

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

Proof. See Appendix C.1. □

We identify that (3.12) is a continuous-region partitioning problem for optimal transport [141]; intuitions can be found in Appendix C.1. Specifically, (3.12) is equivalent to

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & \int p(\boldsymbol{x}) \sigma(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \sum_{i=1}^{N} q_i \lambda_i \\ s.t. \quad & \begin{cases} \sigma(\boldsymbol{x}) & = \min_{i \in [N]} \{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\} \leq \|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i, \qquad \forall i \in [N], \\ \sigma(\boldsymbol{x}) & \geq 0, \end{cases} \end{aligned} \tag{3.13}$$

which has the same form with [141, Eq. (5)]. Note that in [141, Eq. (4)], an auxiliary variable t was used, which introduced λ_i to [141, Eq. (5)]. (In [141], if t were cancelled, λ_i would disappear.)

Note also that in [141], a generic measure $\mathrm{d}A$ was used. In the contexts of this section, it is instantiated to $\mathrm{d}A := p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$. Therefore, for any given $p(\boldsymbol{x})$ and $q(\boldsymbol{x}) = \sum_i q_i \delta_{\boldsymbol{x}^i}(\boldsymbol{x})$, an optimal partition exist [141]. For illustration, see Fig. 3.1, in which we suppose that $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are distributed over the whole rectangular region. However, $q(\boldsymbol{x})$ is discrete ($N = 9$), and supported on nine red dots. The optimal solution states that the optimal transport plan is to move all density of $p(\boldsymbol{x})$ in C_i to its centre $\boldsymbol{x}^i$. In other words, any density outside of C_i will strictly not be accepted at $\boldsymbol{x}^i$. Intuitively, this renders $\int I(\boldsymbol{x}^i|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = q_i,\ \forall i \in [N]$, and $I(\boldsymbol{x}^i|\boldsymbol{x})$ is in fact an indicator: $I(\boldsymbol{x}^i|\boldsymbol{x}) = 1$ if $\boldsymbol{x} \in C_i$ and $I(\boldsymbol{x}^i|\boldsymbol{x}) = 0$ otherwise (cf. Appendix C.1). Therefore, $\int_{\mathbb{R}^n} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \sum_{i=1}^N \int_{C_i} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \sum_{i=1}^N q_i = 1$.

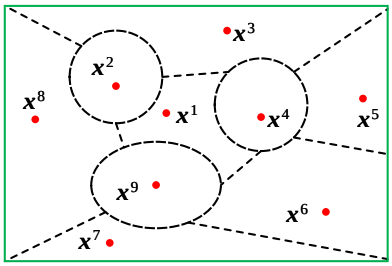


Figure 3.1: The whole rectangular region C is divided into 9 sub-regions C_1, C_2, ..., and C_9 whose centres (red dots) are $\boldsymbol{x}^1$, $\boldsymbol{x}^2$, ..., and $\boldsymbol{x}^9$, respectively. Boundaries of sub-regions are marked by dashed lines.

Lemma 3 transforms (3.11) to

$$\begin{aligned}
&\max_{p(\boldsymbol{x})} && \int -p(\boldsymbol{x}) \ln p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&s.t. && \\
&\begin{cases} \max\limits_{\boldsymbol{\lambda}} \int p(\boldsymbol{x}) \min\limits_{i\in[N]} \left\{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\right\} \mathrm{d}\boldsymbol{x} + \sum\limits_{i=1}^N q_i\lambda_i & \leq \theta \\ \int p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} & = 1. \end{cases}
\end{aligned} \tag{3.14}$$

The solution to problem (3.14) is given in theorem below.

Theorem 15. *The maximum entropy distribution solving (3.14) is*

$$p(\boldsymbol{x}) = \exp\left\{-v_0 \min_{i\in[N]} \left\{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\right\} - v_1 - 1\right\} \tag{3.15}$$

where $v_0 \in \mathbb{R}^1$, $v_1 \in \mathbb{R}^1$, and $\lambda_i \in \mathbb{R}^1, \forall i$ solve the following convex and smooth problem (n.b.,

almost-everywhere smooth in terms of λ_i; non-smooth only on zero-measure boundaries):

$$\begin{aligned} \min_{v_0,v_1,\boldsymbol{\lambda}} \quad & v_0 \cdot (\theta - \textstyle\sum_{i=1}^N \lambda_i q_i) + v_1 + \int \exp\left\{-v_0 \min_{i\in[N]} \left\{\|\boldsymbol{x}-\boldsymbol{x}^i\| - \lambda_i\right\} - v_1 - 1\right\} \mathrm{d}\boldsymbol{x} \\ s.t. \quad & v_0 \geq 0, \end{aligned} \tag{3.16}$$

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

Proof. See Appendix C.2. □

Suppose that v_0^*, v_1^*, and $\boldsymbol{\lambda}^*$ solve (3.16). We claim that $p(\boldsymbol{x})$ in (3.15) admits

$$p(\boldsymbol{x}) = \exp\left\{-v_0^* \cdot \left\{\|\boldsymbol{x}-\boldsymbol{x}^i\| - \lambda_i^*\right\} - v_1^* - 1\right\}, \ \forall \boldsymbol{x} \in C_i, \tag{3.17}$$

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

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

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

Since (3.16) is convex[2] and smooth[3], it can be solved using any first-order method (e.g., projected gradient descent). Let the objective of (3.16) be $f_{W-C}(v_0, v_1, \boldsymbol{\lambda})$; the subscripts "W" is for "Wasserstein" and "C" for "Continuous". By letting $g(\boldsymbol{x}, \boldsymbol{\lambda}) := \min_{i\in[N]} \left\{\|\boldsymbol{x}-\boldsymbol{x}^i\| - \lambda_i\right\}$, the gradients of $f_{W-C}(v_0, v_1, \boldsymbol{\lambda})$ with respect to v_0, v_1, and λ_i are, respectively,

$$\frac{\partial f_{W-C}}{\partial v_0} = \theta - \sum_{i=1}^N \lambda_i q_i - \int_{\mathbb{R}^n} g(\boldsymbol{x}, \boldsymbol{\lambda}) \exp\{-v_0 g(\boldsymbol{x}, \boldsymbol{\lambda}) - v_1 - 1\} \mathrm{d}\boldsymbol{x}, \tag{3.18}$$

$$\frac{\partial f_{W-C}}{\partial v_1} = 1 - \int_{\mathbb{R}^n} \exp\{-v_0 g(\boldsymbol{x}, \boldsymbol{\lambda}) - v_1 - 1\} \mathrm{d}\boldsymbol{x}, \tag{3.19}$$

and

$$\begin{aligned} \frac{\partial f_{W-C}}{\partial \lambda_i} &= -v_0 q_i + v_0 \int_{C_i} \exp\{-v_0 g(\boldsymbol{x}, \boldsymbol{\lambda}) - v_1 - 1\} \mathrm{d}\boldsymbol{x}, \\ &= -v_0 q_i + v_0 \int_{C_i} \exp\{-v_0 (\|\boldsymbol{x}-\boldsymbol{x}^i\| - \lambda_i) - v_1 - 1\} \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{3.20}$$

When the optimality reaches (i.e., all gradients vanish), (3.18) implies that the Wasserstein

[2] In the Lagrange duality sense, dual problems of any primal problems are always concave (resp. convex), no matter whether the primal problems are convex (resp. concave) or not [142, Chapter 5]. One can verify this point on (3.16) themselves by the definition of convexity. Note that for every two bounded functions f_1 and f_2 that have the same support, $\min(f_1 + f_2) \geq \min f_1 + \min f_2$.

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

distance is strictly equal to the prescribed budget θ, (3.19) indicates that $p(\boldsymbol{x})$ in (3.15) is indeed a density that is integrated to unit, and (3.20) means that a partition for optimal transport exists (i.e., $\int_{C_i} p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = q_i$). The projection step is straightforward in the gradient descent procedure: whenever $v_0 < 0$, let $v_0 = 0$.

In the projected gradient descent procedure, all involved integrals can be approximated by numerical methods, e.g., global adaptive quadrature [143] or Monte Carlo integration [144, 145], whichever is easier to be implemented for specific problems.

Solution to (3.8)

Suppose $\mathbb{P}_{\mathbf{x}}$ is also discrete and supported on $\{\boldsymbol{x}^j\}_{j=1,2,...,M}$. We solve (3.8) using the Wasserstein distance. Hence, (3.8) can be written as

$$\begin{aligned} \max_{\boldsymbol{p}} \quad & \sum_{j=1}^{M} -p_j \ln p_j \\ s.t. \quad & \begin{cases} \inf\limits_{\pi(\boldsymbol{x}_{\mathbb{P}},\boldsymbol{x}_{\mathbb{Q}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \pi(\boldsymbol{x}_{\mathbb{P}},\boldsymbol{x}_{\mathbb{Q}}) \mathrm{d}\boldsymbol{x}_{\mathbb{P}} \mathrm{d}\boldsymbol{x}_{\mathbb{Q}} & \leq \theta \\ \sum_{j=1}^{M} p_j & = 1, \end{cases} \end{aligned} \tag{3.21}$$

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

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

Lemma 4. *If both $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{Q}_{\mathbf{x}}$ are discrete, and supported on $\{\boldsymbol{x}^j\}_{j=1,2,...,M}$ and $\{\boldsymbol{x}^i\}_{i=1,2,...,N}$, respectively, the Wasserstein distance $\inf_{\pi(\boldsymbol{x}_{\mathbb{P}},\boldsymbol{x}_{\mathbb{Q}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \pi(\boldsymbol{x}_{\mathbb{P}},\boldsymbol{x}_{\mathbb{Q}}) \mathrm{d}\boldsymbol{x}_{\mathbb{P}} \mathrm{d}\boldsymbol{x}_{\mathbb{Q}}$ is equivalent to a linear program*

$$\begin{aligned} \min_{P_{ij}} \quad & \sum_{i=1}^{N} \sum_{j=1}^{M} \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \cdot P_{ij} \\ s.t. \quad & \begin{cases} \sum_{j=1}^{M} P_{ij} = q_i, & \forall i \in [N], \\ \sum_{i=1}^{N} P_{ij} = p_j, & \forall j \in [M], \\ P_{ij} \geq 0, & \forall i \in [N], \forall j \in [M]. \end{cases} \end{aligned} \tag{3.22}$$

In (3.22), P_{ij} denotes a joint discrete distribution supported on $\{(\boldsymbol{x}^i, \boldsymbol{x}^j)\}_{i\in[N],j\in[M]}$.

Proof. See Appendix C.3. □

Intuitively, (3.22) can also be seen as an optimal transport problem (cf. Lemma 3 and Fig. 3.1): resources are discretely distributed on some given points $\boldsymbol{x}^j$, whereas facilities are fixed at $\boldsymbol{x}^i$.

Lemma 4 transforms (3.21) to

$$\begin{aligned} \max_{\boldsymbol{p}} \quad & \sum_{j=1}^{M} -p_j \ln p_j \\ s.t. \quad & \begin{cases} \min\limits_{P_{ij}} \sum\limits_{i=1}^{N}\sum\limits_{j=1}^{M} \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \cdot P_{ij} \le \theta \\ \sum_{j=1}^{M} P_{ij} = q_i, \quad \forall i \in [N], \\ \sum_{i=1}^{N} P_{ij} = p_j, \quad \forall j \in [M], \\ P_{ij} \ge 0, \ \forall i \in [N], \forall j \in [M]. \end{cases} \end{aligned} \tag{3.23}$$

The constraint $\sum_{j=1}^{M} p_j = 1$ is dropped because it is redundant to (3.23).

Since the left hand side of the first constraint is a minimization problem, we can directly drop the minimization. Thus, (3.23) is equivalent to

$$\begin{aligned} \max_{P_{ij}} \quad & -\sum_{i=1}^{N}\sum_{j=1}^{M} P_{ij} \ln \sum_{i=1}^{N} P_{ij} \\ s.t. \quad & \begin{cases} \sum_{i=1}^{N}\sum_{j=1}^{M} \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \cdot P_{ij} \le \theta \\ \sum_{j=1}^{M} P_{ij} = q_i, \quad \forall i \in [N], \\ P_{ij} \ge 0, \ \forall i \in [N], \forall j \in [M]. \end{cases} \end{aligned} \tag{3.24}$$

The solution to problem (3.24) is given in theorem below.

Theorem 16. *If there exists a discrete distribution $\{P_{ij}^0\}_{\forall i,\forall j}$ that strictly satisfies the inequality $\sum_{i=1}^{N}\sum_{j=1}^{M} \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \cdot P_{ij}^0 < \theta$ and simultaneously satisfies the equality $\sum_{j=1}^{M} P_{ij}^0 = q_i$, the maximum entropy distribution solving (3.24) also solves*

$$\begin{aligned} \min_{v_0,\boldsymbol{\lambda}} \max_{P_{ij}} \quad & v_0\theta + \sum_{i=1}^{N} \lambda_i q_i + \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{P_{ij}^2}{\sum_{i=1}^{N} P_{ij}} \\ s.t. \quad & \begin{cases} -\ln(\sum_{i=1}^{N} P_{ij}) - \frac{P_{ij}}{\sum_{i=1}^{N} P_{ij}} - v_0\|\boldsymbol{x}^i - \boldsymbol{x}^j\| - \lambda_i = 0, \quad \forall i \in [N], \forall j \in [M], \\ P_{ij} \ge 0, \quad \forall i \in [N], \forall j \in [M], \\ v_0 \ge 0, \end{cases} \end{aligned} \tag{3.25}$$

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

Proof. See Appendix C.4. □

The problem (3.25) is intuitively uneasy to be solved because P_{ij} has no closed-form expression. Therefore, we try to relax the original maximum entropy problem (3.24). Since the entropy of a joint distribution is no larger than the sum of the entropy of marginals [132, Theorem 2.6.6]; i.e.,

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

and $-\sum_{i=1}^{N} q_i \ln q_i$ is a constant, we can use the entropy of the join distribution as a surrogate for optimization. Whenever the entropy of the join distribution is maximized, the entropy of $p(\boldsymbol{x})$ is improved as well. [Of course, under this approximation, the entropy of $p(\boldsymbol{x})$ induced from the optimal P_{ij} is not guaranteed to be maximal as in (3.23).] As a result, (3.24) can be relaxed as follows.

$$\begin{aligned} \max_{P_{ij}} \quad & -\sum_{i=1}^{N}\sum_{j=1}^{M} P_{ij}\ln P_{ij} \\ s.t. \quad & \begin{cases} \sum_{i=1}^{N}\sum_{j=1}^{M} \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \cdot P_{ij} \leq \theta \\ \sum_{j=1}^{M} P_{ij} = q_i, \quad \forall i \in [N]. \end{cases} \end{aligned} \tag{3.26}$$

The solution to (3.26) is given in the theorem below.

Theorem 17. *If there exists a discrete distribution $\{P_{ij}^0\}_{\forall i,\forall j}$ that strictly satisfies the inequality $\sum_{i=1}^{N}\sum_{j=1}^{M} \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \cdot P_{ij}^0 < \theta$ and simultaneously satisfies the equality $\sum_{j=1}^{M} P_{ij}^0 = q_i$, then the maximum entropy distribution solving (3.26) is*

$$P_{ij} = \exp\left\{-v_0\|\boldsymbol{x}^i - \boldsymbol{x}^j\| - \lambda_i - 1\right\}, \quad \forall i \in [N], \forall j \in [M], \tag{3.27}$$

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

$$\begin{aligned} \min_{v_0,\boldsymbol{\lambda}} \quad & v_0 \cdot \theta + \sum_{i=1}^{N} \lambda_i q_i + \sum_{i=1}^{N}\sum_{j=1}^{M} \exp\left\{-v_0\|\boldsymbol{x}^i - \boldsymbol{x}^j\| - \lambda_i - 1\right\} \\ s.t. \quad & v_0 \geq 0, \end{aligned} \tag{3.28}$$

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

Proof. Similar to the proof of Theorem 16. □

Since (3.28) is convex and smooth, it can be solved using any first-order method (e.g., projected gradient descent). Let the objective of (3.28) be $f_{W-D}(v_0, \boldsymbol{\lambda})$; the subscripts "W" is for "Wasserstein" and "D" for "Discrete". The gradients of $f_{W-D}(v_0, \boldsymbol{\lambda})$ with respect to v_0 and λ_i

are, respectively,

$$\frac{\partial f_{W-D}}{\partial v_0} = \theta - \sum_{i=1}^{N}\sum_{j=1}^{M} \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \exp\left\{-v_0\|\boldsymbol{x}^i - \boldsymbol{x}^j\| - \lambda_i - 1\right\}, \tag{3.29}$$

and

$$\frac{\partial f_{W-D}}{\partial \lambda_i} = q_i - \sum_{j=1}^{M} \exp\left\{-v_0\|\boldsymbol{x}^i - \boldsymbol{x}^j\| - \lambda_i - 1\right\}. \tag{3.30}$$

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

3.2.3 Solutions Using ϕ-Divergence

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

$$\int_{\mathcal{S}} \phi\left(\frac{\mathrm{d}\mathbb{P}_{\mathbf{x}}}{\mathrm{d}\mathbb{Q}_{\mathbf{x}}}\right) \mathrm{d}\mathbb{Q}_{\mathbf{x}} = \int_{\mathcal{S}} \phi\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) q(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{3.31}$$

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

$$\sum_{i=1}^{N} q_i \phi\left(\frac{p_i}{q_i}\right). \tag{3.32}$$

The ϕ-divergence is a generalization of the Kullback-Leibler divergence. Letting $\phi(t) := t\ln t$ or $\phi(t) := t\ln t - t + 1$, the ϕ-divergence degenerates to the Kullback-Leibler divergence. Other possible choice of $\phi(t)$ can be found in, e.g., [125, Table 2]. For the demonstration purpose only, results in this chapter are only based on the Kullback-Leibler divergence. This is because the Kullback-Leibler divergence is the most popular one which also has clear physical meaning in information theory [131, 146]. Interested readers may try other $\phi(\cdot)$ themselves.

Since the reference distribution $\mathbb{Q}_{\mathbf{x}}$ in this chapter is limited to be discrete, it is pointless to consider the continuity of $\mathbb{P}_{\mathbf{x}}$. Otherwise, $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{Q}_{\mathbf{x}}$ would have discrepant supports so that the ϕ-divergence is undefined. Thus, we only study the solution to (3.8) when $\mathbb{P}_{\mathbf{x}}$ is discrete and neglect the continuous case (3.7).

Solution to (3.8)

We solve (3.8) using the Kullback-Leibler divergence. Hence, (3.8) can be written as

$$\begin{aligned} \max_{\boldsymbol{p}} \quad & \sum_{i=1}^{N} -p_i \ln p_i \\ s.t. \quad & \begin{cases} \sum_{i=1}^{N} p_i \ln\left(\frac{p_i}{q_i}\right) & \leq \theta \\ \sum_{i=1}^{N} p_i & = 1, \end{cases} \end{aligned} \tag{3.33}$$

where $\boldsymbol{p} := [p_1, p_2, ..., p_j, ..., p_N]^\top$ (n.b., $M = N$). The solution to (3.33) is outlined in the theorem below.

Theorem 18. *The distribution solving (3.33) is given by*

$$p_i = \exp\left\{\frac{-\lambda_0 \ln(q_i) + \lambda_1}{-(\lambda_0 + 1)} - 1\right\}, \qquad \forall i \in [N], \tag{3.34}$$

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

$$\begin{aligned} \min_{\lambda_0, \lambda_1} \quad & \lambda_0 \theta + \lambda_1 + (\lambda_0 + 1) \sum_{i=1}^{N} p_i \\ s.t. \quad & \lambda_0 \geq 0. \end{aligned} \tag{3.35}$$

Proof. See Appendix C.5. □

Since (3.35) is convex and smooth, it can be solved using any first-order method (e.g., projected gradient descent). Let the objective of (3.35) be $f_{KL-D}(\lambda_0, \lambda_1)$; the subscripts "KL" is for "Kullback-Leibler" and "D" for "Discrete". The gradients of $f_{KL-D}(\lambda_0, \lambda_1)$ with respect to λ_0 and λ_1 are, respectively,

$$\frac{\partial f_{KL-D}(\lambda_0, \lambda_1)}{\partial \lambda_0} = \theta + \sum_{i=1}^{N} \left[1 + \frac{\ln(q_i) + \lambda_1}{\lambda_0 + 1}\right] p_i, \tag{3.36}$$

and

$$\frac{\partial f_{KL-D}(\lambda_0, \lambda_1)}{\partial \lambda_1} = 1 - \sum_{i=1}^{N} p_i. \tag{3.37}$$

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

3.2.4 Comparisons for the Three Statistical Similarity Measures

As we can see, the moments-based similarity and Wasserstein distance do not require that the two distributions to have the same support. Therefore, a discrete distribution and a continuous distribution can be discussed in a same maximum entropy problem, so can be two discrete distributions with different supports. In addition, the advantage of the Wasserstein distance and the ϕ-divergence is that they can implicitly take into account high-order moments of random variables even for multivariate problems. However, using the Wasserstein distance and the ϕ-divergence implies that computationally intensive numerical problems have to be solved (cf. Theorem 15, Theorem 17, and Theorem 18). Instead, using the moments-based similarity gives the Gaussian approximation state estimation framework which means that closed-form solutions exist (i.e., canonical Kalman iterations).

3.2.5 Projected Gradient Descent Algorithm for Maximum Entropy Problems

Since all maximum entropy problems subject to the Wasserstein distance and the ϕ-divergence can be solved by the projected gradient descent algorithm, we depict it in Algorithm 3.1. Without loss of generality, we use the problem under the Kullback-Leibler divergence [i.e. (3.33)] as an example; see Theorem 18.

3.3 Distributionally Robust State Estimation

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

3.3.1 Generate Worst-Case Prior State Particles

We use the solutions to (3.3) and (3.4) to generate worst-case prior state particles. Solutions under the moments-based similarity measure are just used to argue for the distributional robustness of the Gaussian approximation framework; see Corollary 4. Therefore, we do not cover them in this subsection. Suppose the worst-case prior state particles are $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$.

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

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

1. *If the two sets $\{\boldsymbol{x}^j\}$ and $\{\boldsymbol{x}^i\}$ are identical, the worst-case weights $u_{\boldsymbol{x}^j}$ of particles $\boldsymbol{x}^j$ can be determined by Theorem 17 or Theorem 18.*

Algorithm 3.1: Projected Gradient Descent Method for Maximum Entropy Problem Under the Kullback-Leibler Divergence

Definition: S as maximum allowed iteration steps and s the current iteration step; α as step size; ϵ as numerical precision threshold; abs$(\cdot)$ returns absolute value.

Remark: Since (3.35) is convex, in principle, any initial values for $\lambda_0 \geq 0$ and λ_1 are acceptable. If early stopping is applied (i.e., S is not sufficiently large for time-saving purpose), a normalization procedure is necessary to guarantee $1 = \sum_i p_i$.

Input : $S, \alpha, \epsilon, \lambda_0, \lambda_1$

$s \leftarrow 0$
while *true* **do**
  // Gradient Descent
  $\lambda_0 \leftarrow \lambda_0 - \alpha \cdot \frac{\partial f_{KL-D}}{\partial \lambda_0}$ *// See (3.36)*
  $\lambda_1 \leftarrow \lambda_1 - \alpha \cdot \frac{\partial f_{KL-D}}{\partial \lambda_1}$ *// See (3.37)*
  // Projection
  if $\lambda_0 < 0$ **then**
    $\lambda_0 \leftarrow 0$
  end
  // Next Iteration
  $s \leftarrow s + 1$;
  if $s > S$ ***or*** $\text{abs}(\frac{\partial f_{KL-D}}{\partial \lambda_1}) < \epsilon$ **then**
    // Early Stopping Applied
    if $1 \neq \sum_i p_i$ **then**
      $p_i \leftarrow p_i / \sum_i p_i$ *// Normalization*
    end
    Exit Algorithm
  end
end

Output : p_i in (3.34)

2. *If the two sets are different, the worst-case weights $u_{\boldsymbol{x}^j}$ of particles $\boldsymbol{x}^j$ can be determined by Theorem 17.*

3. *No matter whether the two sets are identical or not, the worst-case weights $u_{\boldsymbol{x}^j}$ of particles $\boldsymbol{x}^j$ can also be determined by Theorem 15 by letting $u_{\boldsymbol{x}^j} \propto p(\boldsymbol{x}^j)$, where $p(\boldsymbol{x})$ is defined in (3.17). In this case, a normalization procedure is necessary; $u_{\boldsymbol{x}^j} \leftarrow u_{\boldsymbol{x}^j} / \sum_j u_{\boldsymbol{x}^j}$.* □

Second, we suppose $\{\boldsymbol{x}^j\}_{j=1,2,\ldots,M}$ are not preset. Hence, we can directly sample M particles from $p(\boldsymbol{x})$ in (3.17). Since $p(\boldsymbol{x})$ is defined in a partitioned region/space, the first step is to choose a sub-region, and the second step is to draw a worst-case prior state particle from this sub-region. We have the following method.

Method 2. *First, draw an integer $j \in [N]$ according to the discrete reference distribution $\mathbb{Q}_{\mathbf{x}}$ (i.e., choose a sub-region C_j whose probability being chosen is q_j). Second, draw a sample $\boldsymbol{x}^j$ from C_j using $p(\boldsymbol{x})$ defined in (3.17). Repeat the two steps above M times to obtain M worst-case prior state particles. In this case, all particles $\boldsymbol{x}^j$ have the same weight $u_{\boldsymbol{x}^j} = 1/M$.* □

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

$$-\sum_{j=1}^{M} p_j \ln p_j = \ln M - \sum_{j=1}^{M} p_j \ln \frac{p_j}{1/M}. \tag{3.38}$$

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

$$-\sum_{j=1}^{M} p_j \ln p_j \geq \sum_{j=1}^{M} p_j(1-p_j) = 1 - \sum_{j=1}^{M} p_j^2. \tag{3.39}$$

It means that any methods reducing the variance of weights (i.e., improving the effective sample size) implicitly elevate the entropy of weights of prior state particles; cf. [45, (51)] or [50, (5)].

3.3.2 Evaluate Worst-Case Likelihoods

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

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

By multiplying $\boldsymbol{\Sigma}$ by a scalar $\theta \geq 1$, a worst-case maximum-entropy likelihood distribution can be obtained because the entropy of the m-dimensional Gaussian distribution $p_{\mathbf{v}}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$\frac{m}{2} + \frac{m}{2}\ln(2\pi) + \frac{1}{2}\ln(|\mathbf{\Sigma}|)$. Hence, improving the covariance implies raising the entropy. Method 3 can be straightforwardly extended to other noise distributions such as the Student's t distribution. We do not cover details here.

However, when the nominal measurement noise is non-additive and non-multiplicative, such closed-form evaluation methods are unavailable. Therefore, numerical methods are indispensable. The first step is to generate nominal likelihood particles $\{\boldsymbol{y}^r|\boldsymbol{x}^j\}_{r=1,2,...,R}$ for each worst-case prior state particle $\boldsymbol{x}^j$ (n.b., when the process dynamics is exact, worst-case $\{\boldsymbol{x}^j\}_{j\in[M]}$ and nominal $\{\boldsymbol{x}^i\}_{i\in[N]}$ are the same). This can be done by the nominal measurement dynamics $\mathbf{y} = \boldsymbol{h}(\boldsymbol{x}^j, \mathbf{v})$. Specifically, we need to generate R samples from $\mathbb{P}_{\mathbf{v}}$, say $\boldsymbol{v}^r$, and obtain $\{\boldsymbol{y}^r|\boldsymbol{x}^j\}_{r\in[R]}$ by $\boldsymbol{y}^r|\boldsymbol{x}^j := \boldsymbol{h}(\boldsymbol{x}^j, \boldsymbol{v}^r), \forall r \in [R]$. Since $\mathbf{v}$ is high-dimensional, we use the importance sampling method [17, Section 11.1.4]: 1) uniformly draw R samples in the support of $\mathbf{v}$, and 2) use $p_{\mathbf{v}}(\boldsymbol{v})$ to determine their weights; $u_{\boldsymbol{v}^r} \propto p_{\mathbf{v}}(\boldsymbol{v}^r)$ (n.b., a normalization procedure is hence necessary). Based on nominal likelihood particles $\{\boldsymbol{y}^r|\boldsymbol{x}^j\}_{r=1,2,...,R}$ whose weights are $u_{\boldsymbol{y}^r|\boldsymbol{x}^j} = u_{\boldsymbol{v}^r}$, the worst-case likelihood of $\boldsymbol{x}^j$ is ready to be evaluated. Suppose the support set of the worst-case likelihood distribution is $\{\boldsymbol{y}^t|\boldsymbol{x}^j\}_{t=1,2,...,T}$. As the case that generates worst-case prior state particles in Subsection 3.3.1, $\{\boldsymbol{y}^t|\boldsymbol{x}^j\}_{t=1,2,...,T}$ can be just a copy of $\{\boldsymbol{y}^r|\boldsymbol{x}^j\}_{r=1,2,...,R}$ (thus $T := R$) or uniformly sampled from a subset of $\mathbb{R}^m$. The subset can be the smallest hyperrectangle or hyperellipsoid containing $\{\boldsymbol{y}^r|\boldsymbol{x}^j\}_{r=1,2,...,R}$.

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

Method 4. *Suppose $p^*_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y})$ solves (3.5) using the Wasserstein distance. The worst-case likelihood of $\boldsymbol{x}^j$ given the measurement $\boldsymbol{y}$ is $p^*_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y})$.* □

Method 5. *Augment $\boldsymbol{y}$ into the support sets of worst-case likelihood distributions for $\boldsymbol{x}^j, j \in [M]$; i.e., let $\{\boldsymbol{y}^t|\boldsymbol{x}^j\}_{t=1,2,...,T+1} := \{\boldsymbol{y}\} \bigcup \{\boldsymbol{y}^t|\boldsymbol{x}^j\}_{t=1,2,...,T}$. Suppose $p^*_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y})$ solves (3.6) using the Wasserstein distance (n.b., the Kullback-Leibler Divergence is not applicable because after augmentation, the two support sets are hardly identical). The worst-case likelihood of $\boldsymbol{x}^j$ given the measurement $\boldsymbol{y}$ is $p^*_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y})$.* □

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

Remark 5. *In practice, when there do not exist model uncertainties in measurement dynamics, we may have a heuristic method to evaluate the likelihoods of prior particles*

$$p(\boldsymbol{y}|\boldsymbol{x}^j) := \frac{\exp\left(-\alpha \sum_{r=1}^{R} u_{\boldsymbol{y}^r|\boldsymbol{x}^j} \cdot \|\boldsymbol{y} - \boldsymbol{y}_j^r\|\right)}{\sum_{j=1}^{M} \exp\left(-\alpha \sum_{r=1}^{R} u_{\boldsymbol{y}^r|\boldsymbol{x}^j} \cdot \|\boldsymbol{y} - \boldsymbol{y}_j^r\|\right)},$$

where $\alpha > 0$ is a scaling coefficient and $\boldsymbol{y}_j^r$ is a notational shorthand for the likelihood particle

$\boldsymbol{y}^r|\boldsymbol{x}^j$. *Since there are no model uncertainties in measurement dynamics, nominal likelihood particles are directly used. Another strategy can be* $p(\boldsymbol{y}|\boldsymbol{x}^j) := \frac{\exp\left(-\alpha\|\boldsymbol{y}-\sum_{r=1}^{R} u_{\boldsymbol{y}^r|\boldsymbol{x}^j}\cdot\boldsymbol{y}_j^r\|\right)}{\sum_{j=1}^{M}\exp\left(-\alpha\|\boldsymbol{y}-\sum_{r=1}^{R} u_{\boldsymbol{y}^r|\boldsymbol{x}^j}\cdot\boldsymbol{y}_j^r\|\right)}$. □

3.3.3 Outlier Treatment

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

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

The outlier treatment method is given below.

Method 7. *The identified outlier can be directly trashed and all prior state particles directly become posterior, during which associated weights keep unchanged. This idea is motivated by re-descending influence functions in M-estimation, e.g., Hampel's influence function [55, Eq. (4.90)]. The outlier can also be replaced by the nearest likelihood particle generated by the prior state particle that has the largest likelihood or replaced by the nearest likelihood particle generated by the weighted mean. This philosophy is motivated by monotonic influence functions in M-estimation, e.g., Huber's influence function [55, Eq. (4.53)].* □

3.3.4 Overall Method

The distributionally robust particle filtering framework is summarized in Algorithm 3.2. Algorithm 3.2 is a robustified version of the popular canonical particle filter in [45, Algorithm 3]. The used proposal density (i.e., importance density) for importance sampling is the prior distribution as in [45, Eq. (63)].

Remark 6 (Symbols in Algorithm 3.2)**.** *k as discrete time index; N as number of nominal prior state particles; M as number of worst-case prior (and also posterior) state particles; R as number of nominal likelihood particles for every (worst-case) prior state particle, and T as number of worst-case likelihood particles for the same (worst-case) prior state particle; $\boldsymbol{x}_0^i$ as posterior state particles at $k=0$ and $u_{\boldsymbol{x}_0^i}$ the associated weights, $\forall i \in [N]$; $p^*(\boldsymbol{y}_k|\boldsymbol{x}_k^j)$ as worst-case likelihood of $\boldsymbol{x}_k^j$ given $\boldsymbol{y}_k$; $\hat{N}_{eff}$ as effective sample size and N_{thres} its threshold.* □

Remark 7. *If measurement noises are additive or multiplicative, ignore Step 3, and use Method 3 in Step 4. If there are no process model uncertainties, ignore Step 2. If resampling is applied at every time k, M and N can be different; cf. Line 28. Otherwise, M and N must be identical to guarantee the number of posterior state particles at time $k-1$ is the same as the number of prior state particles at time k; cf. Step 1.* □

Algorithm 3.2: Distributionally Robust Particle Filtering for Nonlinear Systems

Remarks: See Remark 6 and Remark 7. For every k, execute the following 5 steps.

Initialization: N, M, R, T, N_{thres}, and $\{\boldsymbol{x}_0^i, u_{\boldsymbol{x}_0^i}\}_{i\in[N]}$.

Input : $\boldsymbol{y}_k,\ k = 1, 2, 3, ...$

// Step 1: Generate Nominal Prior State Particles

for $i = 1 : N$ **do**

  Sample $\boldsymbol{w}_{k-1}^i$ from the distribution of $\mathbf{w}_{k-1}$

  $\boldsymbol{x}_k^i = \boldsymbol{f}_k(\boldsymbol{x}_{k-1}^i, \boldsymbol{w}_{k-1}^i)$

end

// Step 2: Obtain Worst-Case Prior State Particles

Use Method 1 or Method 2 to generate worst-case prior state particles $\{\boldsymbol{x}_k^j\}_{j\in[M]}$ and obtain their weights $\{u_{\boldsymbol{x}_k^j}\}_{j\in[M]}$

// Step 3: Evaluate Worst-Case Likelihood for Every $\boldsymbol{x}_k^j$

for $j = 1 : M$ **do**

  //Generate Nominal Likelihood Particles $\boldsymbol{y}_k^r, \forall r \in [R]$

  for $r = 1 : R$ **do**

    Sample $\boldsymbol{v}_k^r$ from the distribution of $\mathbf{v}_k$

    $\boldsymbol{y}_k^r = \boldsymbol{h}_k(\boldsymbol{x}_k^j, \boldsymbol{v}_k^r)$

  end

  //Evaluate Worst-Case Likelihood of $\boldsymbol{x}_k^j$ *at* $\boldsymbol{y}_k$

  Use Method 4 or Method 5 for likelihood evaluation

  //Outlier Identification and Treatment

  Use Method 6 for outlier identification and Method 7 for outlier treatment

end

// Step 4: Generate Posterior State Particles $\boldsymbol{x}_k^j$

for $j = 1 : M$ **do**

  Keep $\boldsymbol{x}_k^j$ unchanged; Update weights by $u_{\boldsymbol{x}_k^j} \leftarrow u_{\boldsymbol{x}_k^j} \cdot p^*(\boldsymbol{y}_k|\boldsymbol{x}_k^j)$

end

Normalize weights $u_{\boldsymbol{x}_k^j}, \forall j \in [M]$

// Step 5: Resampling

$\hat{N}_{eff} \leftarrow 1/\sum_{j=1}^{M} u_{\boldsymbol{x}_k^j}^2$

if $\hat{N}_{eff} < N_{thres}$ **then**

  Resample N times from $\{\boldsymbol{x}_k^j, u_{\boldsymbol{x}_k^j}\}_{j\in[M]}$

end

Output : Worst-case posterior state particles $\{\boldsymbol{x}_k^i\}$ and weights $\{u_{\boldsymbol{x}_k^i}\}$, $\forall i \in [N]$.

3.3.5 Computational Complexity

As we can see, the proposed generic robustified particle filter is computationally intensive: if S in Algorithm 3.1 and N, M, R, and T in Algorithm 3.2 are large, the calculation burden is heavy as well. The worst-case complexity order of Algorithm 3.1 is $\mathcal{O}(S)$. However, the complexity order of Algorithm 3.2 is hard to be specified because it depends on which sampling method (e.g., the importance sampling and the fundamental theorem of simulation) is used, which resampling method (e.g., systematic and multinomial) is used, and which maximum-entropy method (among Methods 1-5) is used. The burden, however, is unavoidable to robustify particle-based filters and to evaluate likelihoods under non-additive and non-multiplicative measurement noises. If the process dynamics is exact and measurement noise densities fortunately have closed-form expressions (see, e.g., Method 3), then the computation burden can be limited and the resulted robust particle filter has the same computational complexity as the canonical particle filter (because no extra computation burden is introduced in Step 2 and Step 3).

3.3.6 Sizes of Ambiguity Sets

The sizes of ambiguity sets (i.e., θ's in Theorems 15-18) need to be specified in implementing the robustified particle filter. However, this cannot be theoretically conducted because for a real state estimation problem, the true states are unknown. In other words, the training dataset is unavailable so that the sizes of ambiguity sets cannot be tuned to be (nearly) optimal. Therefore, signal processing practitioners are expected to try appropriate values for their specific problems. The general principle is that the sizes can be neither too large nor too small: an extremely large value renders the robust filter being too conservative, while the robust filter with an extremely small value cannot provide sufficient robustness. Since this tuning principle remains the same for linear systems [57, 58], we do not repeat experimental details in this section; details can be seen in Subsection 2.2.6 (*Suggestions on Tuning the Size of the Ambiguity Set*) and Subsection 2.3.5 (*Sensitivity Analysis*).

3.4 Experiments

All the source data and codes are available online at GitHub: https://github.com/Spratm-Asleaf/DRSE-Nonlinear.

3.4.1 Find Maximum Entropy Distributions

Continuous Maximum Entropy Distribution Using Wasserstein Distance

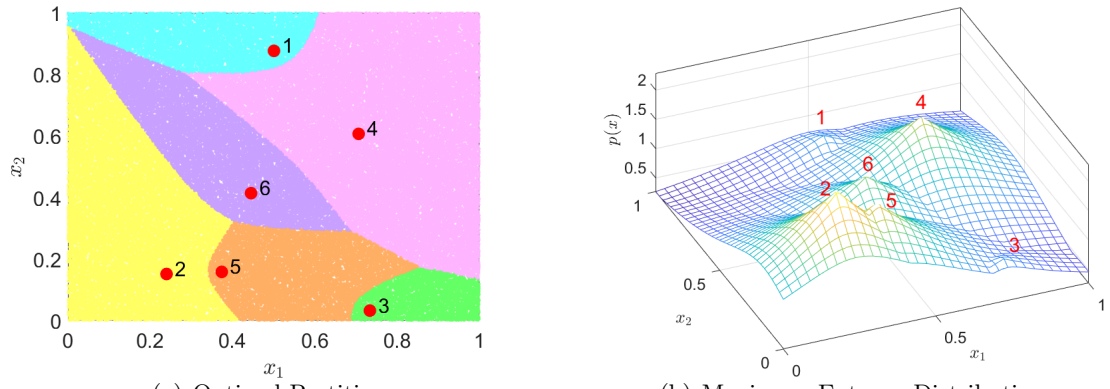
We consider a two-dimensional continuous rectangular region $[0, 1] \times [0, 1]$. Let $\boldsymbol{x}$ be a 2-dimensional prior state particle: x_1 denote the horizontal axis and x_2 the vertical axis. Suppose the reference discrete prior state distribution $\boldsymbol{q}$ is supported on six points, which are randomly sampled from the rectangle. Their weights are also randomly determined. The points and their weights are displayed in Table 3.1.

Table 3.1: The reference distribution

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

We use Theorem 15 and its corresponding projected gradient descent method to find the continuous maximum entropy distribution. The uncertainty budget θ is set to $\theta := 0.025$ (only for a possible demonstration; other values also applicable). In the projected gradient descent procedure, the step size $\alpha := 0.05$ and the maximum allowed iteration steps $S := 500$. The results are shown in Fig. 3.2. The Monte Carlo integration method is used to evaluate integrals in (3.18), (3.19), and (3.20); for every Monte Carlo sample $\boldsymbol{x}$, it belongs to C_i if

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



(a) Optimal Partition. (b) Maximum Entropy Distribution.

Figure 3.2: Optimal partition and maximum entropy distribution. The whole rectangular region is partitioned into six sub-regions. Red-filled circles in (a) indicate the supports of the reference distribution $\boldsymbol{q}$. Peaks in (b) correspond to the supporting points of $\boldsymbol{q}$.

Discrete Maximum Entropy Distribution Using Kullback-Leibler Divergence

The reference distribution $\boldsymbol{q}$ and the induced maximum entropy distribution $\boldsymbol{p}$ are displayed in Table 3.2 and Fig. 3.3. $\boldsymbol{p}$ is calculated by Theorem 18. Since they have the same support set, we do not explicitly demonstrate what the particles $\boldsymbol{x}^i$ are. The uncertainty budget θ is set to $\theta := 0.0075$ (only for a possible demonstration; other values also applicable). In the projected gradient descent procedure, the step size $\alpha := 0.05$ and the maximum allowed iteration steps $S := 500$. From Table 3.2 and Fig. 3.3, we can see that $\boldsymbol{p}$ are more balanced than $\boldsymbol{q}$: the

minimum of $\boldsymbol{p}$ is larger than that of $\boldsymbol{q}$ (when $i = 4$), while the maximum of $\boldsymbol{p}$ is smaller than that of $\boldsymbol{q}$ (when $i = 2$).

Table 3.2: The reference distribution and its induced maximum entropy distribution (Using Kullback-Leibler Divergence)

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

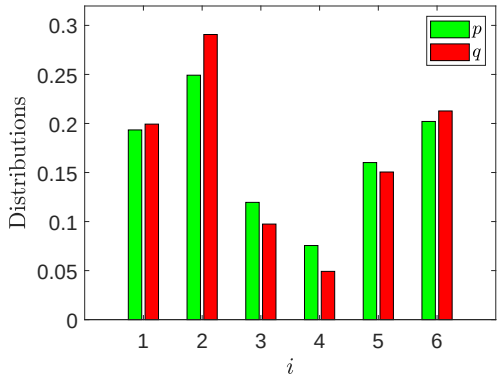


Figure 3.3: The maximum entropy distribution $\boldsymbol{p}$ (left bar at each i) induced by the reference distribution $\boldsymbol{q}$ (right bar at each i) using the Kullback-Leibler Divergence.

Discrete Maximum Entropy Distribution Using Wasserstein Distance

We let the reference discrete distribution $\boldsymbol{q}$ explicitly be a likelihood distribution of one (worst-case) prior state particle $\boldsymbol{x}$. Suppose $\boldsymbol{q}$ and its induced maximum entropy distribution $\boldsymbol{p}$ have different support sets, as displayed in Fig. 3.4. The support set $\{\boldsymbol{y}^r|\boldsymbol{x}\}_{r\in[R]}$ of $\boldsymbol{q}$ consists of particles propagated from a 2-dimensional nonlinear measurement dynamics

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

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

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

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

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

where $\boldsymbol{p}$ is obtained by Theorem 17. The uncertainty budget θ is set to $\theta := 0.325$ (only for a possible demonstration; other values also applicable). In the projected gradient descent procedure, the step size $\alpha := 0.05$ and the maximum allowed iteration steps $S := 500$. As expected, although the support sets are different, we can still calculate the weights of new supporting points of $\boldsymbol{p}$, and the worst-case likelihood of the new measurement is evaluated as 0.0064. This small-valued likelihood result coincides with our intuition because the new point No. 6 is far away from the supports (i.e., red-filled circles) of $\boldsymbol{q}$.

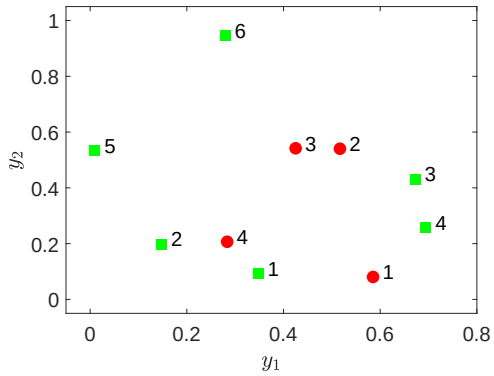


Figure 3.4: The maximum entropy distribution $\boldsymbol{p}$ induced by the reference distribution $\boldsymbol{q}$ using the Wasserstein distance. Red-filled circles are supports of $\boldsymbol{q}$, while green-filed squares are supports of $\boldsymbol{p}$.

Alternatively, we may suppose the support set of $\boldsymbol{p}$ is constructed by the union of the support set of $\boldsymbol{q}$ and the new measurement. The supporting points of $\boldsymbol{q}$ are uniformly sampled from $[0,1] \times [0,1]$. We have the results in Table 3.3, in which $\boldsymbol{y}^5$ is a new measurement uniformly sampled from $[0,1] \times [0,1]$ as well. The uncertainty budget θ is set to $\theta := 0.01$ (only for a possible demonstration; other values also applicable). In the projected gradient descent procedure, the step size $\alpha := 0.05$ and the maximum allowed iteration steps $S := 500$. From Table 3.3, it can be seen that the likelihood (of the associated worst-case prior state particle) at this new measurement is 0.0260.

3.4.2 A Target Tracking Example

In this section, we consider a target tracking problem under uncertain conditions; see Fig. 3.5. The target moves along the curved-orange-dotted trajectory and its true (but unknown) speed is $v = 2m/s$. The sensor is able to obtain real-time distance ρ and relative orientation φ from the target to itself; it moves along the vertical axis from the origin and its speed is $v_0 = 1m/s$.

Table 3.3: The reference distribution and its induced maximum entropy distribution (Using Wasserstein Distance)

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

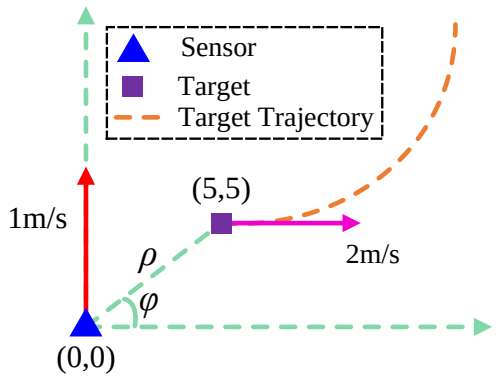


Figure 3.5: A target tracking diagram. The initial position of the target is $(5, 5)$ and of the sensor is $(0, 0)$.

Therefore, the nominal process model (i.e., target maneuver model) is $\mathbf{x}_k = \boldsymbol{F}\mathbf{x}_{k-1} + \boldsymbol{G}\mathbf{w}_{k-1}$ and

$$\mathbf{x}_k := \begin{bmatrix} \mathrm{x}_{1,k} \\ \mathrm{s}_{1,k} \\ \mathrm{x}_{2,k} \\ \mathrm{s}_{2,k} \end{bmatrix}, \boldsymbol{F} := \begin{bmatrix} 1 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \Delta t \\ 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{G} := \begin{bmatrix} \Delta t & 0 \\ 1 & 0 \\ 0 & \Delta t \\ 0 & 1 \end{bmatrix},$$

where $\Delta t := 0.5s$ is the sampling time; x_1 and s_1 (resp. x_2 and s_2) denote the real-time position and speed of the target in the horizontal (resp. vertical) axis, respectively; white-Gaussian-distributed $\mathbf{w}_{k-1}$ is the speed noise vector whose mean is zero and covariance is $\boldsymbol{Q}_k := \mathrm{diag}\{0.5, 0.5\}$. On the other hand, the nominal measurement model is $\mathbf{y}_k := [\rho_k, \varphi_k]^\top$ and

$$\rho_k = \sqrt{(\mathrm{x}_{1,k} - \mathrm{x}^0_{1,k})^2 + (\mathrm{x}_{2,k} - \mathrm{x}^0_{2,k})^2} + \mathrm{v}_1,$$

$$\varphi_k = \arctan\left(\frac{\mathrm{x}_{2,k} - \mathrm{x}^0_{2,k}}{\mathrm{x}_{1,k} - \mathrm{x}^0_{1,k}}\right) + \mathrm{v}_2,$$

where $\mathrm{x}^0_{1,k}$ and $\mathrm{x}^0_{2,k}$ denote the real-time position of the sensor in the horizontal axis and the vertical axis, respectively; v_1 is the ranging error and v_2 is the heading error. Both v_1 and v_2 are white Gaussian with zero mean. The measurement noise covariance is $\boldsymbol{R}_k := \begin{bmatrix} 0.1 & 0 \\ 0 & 0.0001 \end{bmatrix}$. (Namely, the error range of v_1 is $\pm 3\sqrt{0.1} = \pm 0.95m$ and of v_2 is $\pm 3\sqrt{0.0001} = \pm 0.03rad = \pm 1.7deg.$) The unit of all position variables is meter, the unit of all speed variables is meter per second, and the unit of all angle variables is radian.

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

$$\rho_k = \sqrt{(\mathrm{x}_{1,k} - \mathrm{x}^0_{1,k} - \eta_{1,k})^2 + (\mathrm{x}_{2,k} - \mathrm{x}^0_{2,k} - \eta_{2,k})^2} + \mathrm{v}_1,$$

$$\varphi_k = \arctan\left(\frac{\mathrm{x}_{2,k} - \mathrm{x}^0_{2,k} - \eta_{2,k}}{\mathrm{x}_{1,k} - \mathrm{x}^0_{1,k} - \eta_{1,k}}\right) + \mathrm{v}_2,$$

where $\eta_{1,k}$ and $\eta_{2,k}$ are positioning errors. In this experiment, they are assumed to be Gaussian having the same mean of zero and the same variance of 0.3; i.e., the error range is $\pm 3\sqrt{0.3} = \pm 1.6m$.

We conduct ten episodes of Monte Carlo simulation and each episode runs 100 discrete time steps. For each episode, the target tracking accuracy is measured by rooted mean square error (RMSE) along 100 time steps, i.e.,

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

where $\hat{x}$ denotes the estimate of x. The overall target tracking accuracy is measured by the averaged RMSEs of ten episodes. We implement the canonical particle filter (PF) in [45, Algorithm 3], the Gaussian approximation method (GA) in Theorem 13 [specifically, the Ensemble Kalman filter (EnKF)], and the robust particle filter (RPF) in Algorithm 3.2 for comparison. In this example, since the measurement noises v_1 and v_2 are additive and Gaussian, Method 3 with $\theta := 5$ is used to evaluate worst-case likelihoods for the proposed robust particle filter. (θ can be set to other possible values; we just set $\theta := 5$ as an example.) For all methods, we assume that the initial state particles are sampled from a 4-dimensional Gaussian distribution with mean of $[5, 0, 5, 0]^\top$ and covariance of $\mathrm{diag}\{0.2, 0.2, 0.2, 0.2\}$. In Algorithm 3.1, $S := 500$, $\alpha := 0.05$, $\epsilon := 1 \times 10^{-4}$, $\lambda_0 := 2$, and $\lambda_1 := 0$. In Algorithm 3.2, $N = M := 1000$. (We do not initialize R and T because for this closed-form likelihood evaluation case, they are not used.) For the

demonstration purpose and without loss of generality, all the involved parameter values in this experiment are arbitrarily set, one may also try other values for comparison.

The results with and without model uncertainties (i.e., η_1 and η_2) are shown in Table 3.4. In Table 3.4, "Time" denotes average execution time at each time step (unit: seconds).

Table 3.4: The target tracking results with and without uncertainties

	PF		GA (EnKF)		RPF	
	RMSE	Time	RMSE	Time	RMSE	Time
With	1.49	0.013	1.32	**0.0023**	**1.25**	0.0097
Without	**1.01**	0.011	1.13	**0.0021**	1.05	0.0090

As we can see, when there exist model uncertainties, GA and RPF have smaller averaged RMSEs because they are distributionally robust. Besides, RPF outperforms GA since GA assumes Gaussianity of prior state distributions and likelihood distributions. On the other hand, when there are no model uncertainties (i.e., the nominal model exactly coincides with the true model), the canonical PF works best because it is optimal for the nominal model. However, the benefit of using GA is that it is always computationally easy.

3.5 Chapter Conclusions

This chapter studies the distributionally robust state estimation scheme for nonlinear systems subject to model uncertainties. Attention has been paid to the particle filtering framework due to its flexibility. The maximum entropy prior state distributions and the maximum entropy likelihood distributions are leveraged to robustify the particle filter. The existing Gaussian approximation framework is proven to be distributionally robust. In addition, a generic likelihood evaluation method is presented under non-additive and non-multiplicative measurement noises. However, extra computation burden is required to obtain worst-case prior state particles even when worst-case likelihoods can be analytically evaluated. Another issue is to properly choose the radii of ambiguity sets, i.e., θ's in Theorems 15-18. Nevertheless, these radii cannot be trained to be (nearly) optimal because for real state estimation problems, true states (i.e., training dataset) are unknown. Therefore, in practice, practitioners have to try appropriate values for their specific problems.

Chapter 4

Conclusions

This thesis studies distributionally robust state estimation frameworks for both linear systems and nonlinear systems. In a big view, the modeling uncertainties are quantified by families of distributions, and the worst-case distributions are leveraged to find robust state estimators that are insensitive to model uncertainties. The families of distributions are constructed as balls centered at nominal distributions with radii defined by statistical similarity measures such as Wasserstein distance, Kullback-Leibler divergence, and moment-based similarity. Comprehensive comparisons with existing robust state estimation solutions are made through showing the advantages and disadvantages of existing methods.

For the linear case, the following key points can be summarized.

1) The proposed framework can account for both parameter uncertainties and measurement outliers. It offers a new perspective to understand the robust state estimation problem under parameter uncertainties and measurement outliers, and generalizes several classic methods into a unified framework.

2) The 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.

3) 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 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.

4) 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.

For the nonlinear case, attention has been paid to the particle filtering framework to counteract model uncertainties due to its flexibility. The key points are as follows.

1) The maximum entropy prior state distributions and the maximum entropy likelihood distributions are leveraged to robustify the particle filter.

2) The proposed maximum-entropy strategies can also provide weight-balancing mechanism to reduce particle degeneracy and new-sample-generating mechanism to diminish particle impoverishment.

3) The existing Gaussian approximation framework is shown to be distributionally robust.

4) A generic likelihood evaluation method is presented under non-additive and non-multiplicative measurement noises.

For both the linear case and the nonlinear case, three remarks below have to be outlined.

1) Robust filters are just remedial solutions. Reducing modeling uncertainties is always important. Readers should not expect that the proposed methods are optimal or satisfactory in all scenarios, e.g., for a model with t-distributed measurement noises (which implies that the true model is known to be with t-distributed measurement noises).

2) It is better to take into consideration model uncertainties from immediate sources where uncertainties occur because higher-level treatments tend to be more conservative (i.e., loss of flexibility). For example, see Section 2.2.5: if parameter uncertainties can be directly modeled/quantified, the corresponding specific-purpose robust solutions are likely to outperform the proposed general-purpose distributionally robust solution.

3) The robustness under uncertain conditions comes with the cost of sacrificing the optimality under perfect conditions.

However, the proposed algorithms are not robust with respect to the sizes of the ambiguity sets, i.e., θ_2 in Algorithm 2.1, θ's in Algorithm 2.2, and θ's in Theorems 15-18. Unfortunately, the optimal or convincing tuning methods for the sizes of ambiguity sets have yet to be found. Nevertheless, these radii cannot be trained to be (nearly) optimal because for real state estimation problems, true states (i.e., training dataset) are unknown. Therefore, in practice, practitioners have to try appropriate values for their specific problems. We invite scholars in this field to collaborate with the author on addressing the two issues below in the future.

1) How can these radii/sizes 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 time-varying radii/sizes?

Although imperfect, the proposed method is still promising because, for example, tuning a scalar parameter, e.g., θ_2 in Algorithm 2.1 is easier than tuning structural matrices $\boldsymbol{\Gamma}_{k-1}$ in (2.31), $\boldsymbol{M}_{k-1}$, $\boldsymbol{E}_{f,k-1}$, and $\boldsymbol{E}_{g,k-1}$ in (2.33), and $\boldsymbol{F}_{i,k-1}$ and $\boldsymbol{G}_{i,k-1}$ in (2.34).

References

[1] N. Wahlström and E. Özkan, "Extended target tracking using Gaussian processes," *IEEE Transactions on Signal Processing*, vol. 63, no. 16, pp. 4165–4178, 2015.

[2] X. R. Li and V. P. Jilkov, "Survey of maneuvering target tracking. part i. dynamic models," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1333–1364, 2003.

[3] J. Zhao, A. Gómez-Expósito, M. Netto, L. Mili, A. Abur, V. Terzija, I. Kamwa, B. Pal, A. K. Singh, J. Qi, *et al.*, "Power system dynamic state estimation: Motivations, definitions, methodologies, and future work," *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 3188–3198, 2019.

[4] W. Peng, Z.-S. Ye, and N. Chen, "Joint online RUL prediction for multivariate deteriorating systems," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 2870–2878, 2018.

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

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

[7] C. Papachristos, T. Dang, S. Khattak, F. Mascarich, N. Khedekar, K. Alexis, *et al.*, *Modeling, Control, State Estimation and Path Planning Methods for Autonomous Multirotor Aerial Robots*. Now Publishers, 2018.

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

[9] Y. Aviv, "Gaining benefits from joint forecasting and replenishment processes: The case of auto-correlated demand," *Manufacturing & Service Operations Management*, vol. 4, no. 1, pp. 55–74, 2002.

[10] Y. Aviv, "A time-series framework for supply-chain inventory management," *Operations Research*, vol. 51, no. 2, pp. 210–227, 2003.

[11] Y. Ho and R. Lee, "A Bayesian approach to problems in stochastic estimation and control," *IEEE Transactions on Automatic Control*, vol. 9, no. 4, pp. 333–339, 1964.

[12] I. Gorynin, *Bayesian State Estimation in Partially Observable Markov Processes*. PhD thesis, Université Paris-Saclay, 2017.

[13] 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 Processing*, vol. 135, pp. 218–238, 2017.

[14] V. A. Nguyen, S. Shafieezadeh Abadeh, M.-C. Yue, D. Kuhn, and W. Wiesemann, "Optimistic distributionally robust optimization for nonparametric likelihood approximation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

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

[16] F. Gustafsson, "Particle filter theory and practice with positioning applications," *IEEE Aerospace and Electronic Systems Magazine*, vol. 25, no. 7, pp. 53–82, 2010.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[18] D. Simon, *Optimal State Estimation: Kalman, H_∞, and Nonlinear Approaches*. John Wiley & Sons, 2006.

[19] B. D. Anderson and J. B. Moore, *Optimal Filtering*. Prentice-Hall, 1979.

[20] G. Chen, *Approximate Kalman Filtering*, vol. 2. World Scientific, 1993.

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

[22] I. Urteaga, M. F. Bugallo, and P. M. Djurić, "Sequential Monte Carlo methods under model uncertainty," in *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 1–5, IEEE, 2016.

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

[24] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Prentice Hall, 2000.

[25] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, 1960.

[26] P. Tichavsky, C. H. Muravchik, and A. Nehorai, "Posterior cramér-rao bounds for discrete-time nonlinear filtering," *IEEE Transactions on Signal Processing*, vol. 46, no. 5, pp. 1386–1396, 1998.

[27] A. H. Sayed, "A framework for state-space estimation with uncertain models," *IEEE Transactions on Automatic Control*, vol. 46, no. 7, pp. 998–1013, 2001.

[28] B. Hassibi, A. H. Sayed, and T. Kailath, "Linear estimation in Krein spaces. I. theory," *IEEE Transactions on Automatic Control*, vol. 41, no. 1, pp. 18–33, 1996.

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

[30] A. Pourkabirian and M. H. Anisi, "Robust channel estimation in multiuser downlink 5G systems under channel uncertainties," *IEEE Transactions on Mobile Computing*, 2021.

[31] Y. Liang, D. Zhou, L. Zhang, and Q. Pan, "Adaptive filtering for stochastic systems with generalized disturbance inputs," *IEEE Signal Processing Letters*, vol. 15, pp. 645–648, 2008.

[32] L. Han, Z. Ren, and D. S. Bernstein, "Maneuvering target tracking using retrospective-cost input estimation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 5, pp. 2495–2503, 2016.

[33] 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.

[34] K. Dedecius and O. Tichỳ, "Collaborative sequential state estimation under unknown heterogeneous noise covariance matrices," *IEEE Transactions on Signal Processing*, vol. 68, pp. 5365–5378, 2020.

[35] S. Wang, Z. Wu, and A. Lim, "Denoising, outlier/dropout correction, and sensor selection in range-based positioning," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

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

[37] K. Fujii, "Extended Kalman filter," *Refernce Manual*, pp. 14–22, 2013.

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

[39] E. A. Wan and R. Van Der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*, pp. 153–158, IEEE, 2000.

[40] I. Arasaratnam and S. Haykin, "Cubature Kalman filters," *IEEE Transactions on Automatic Control*, vol. 54, no. 6, pp. 1254–1269, 2009.

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

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

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

[44] J. Courts, A. Wills, and T. B. Schon, "Gaussian variational state estimation for nonlinear state-space models," *IEEE Transactions on Signal Processing*, 2021.

[45] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

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

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

[48] T. Li, M. Bolic, and P. M. Djuric, "Resampling methods for particle filtering: classification, implementation, and strategies," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 70–86, 2015.

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

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

[51] O. Cappé and E. Moulines, "On-line expectation–maximization algorithm for latent data models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 593–613, 2009.

[52] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 3, pp. 269–342, 2010.

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

[54] K. Myers and B. Tapley, "Adaptive sequential estimation with unknown noise statistics," *IEEE Transactions on Automatic Control*, vol. 21, no. 4, pp. 520–523, 1976.

[55] P. J. Huber, *Robust Statistics (2nd edition)*. John Wiley & Sons, 2009.

[56] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," *INFORMS Tutorials in Operations Research*, pp. 130–166, 2019.

[57] S. Wang, Z. Wu, and L. Andrew, "Robust state estimation for linear systems under distributional uncertainty," *IEEE Transactions on Signal Processing*, 2021. DOI: 10.1109/TSP.2021.3118540.

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

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

[60] A. D. Kshemkalyani and M. Singhal, *Distributed computing: principles, algorithms, and systems*. Cambridge University Press, 2011.

[61] M. Cetin, L. Chen, J. W. Fisher, A. T. Ihler, R. L. Moses, M. J. Wainwright, and A. S. Willsky, "Distributed fusion in sensor networks," *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 42–55, 2006.

[62] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1-2, pp. 115–166, 2018.

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

[64] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, "Regularization via mass transportation," *Journal of Machine Learning Research*, vol. 20, no. 103, pp. 1–68, 2019.

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

[66] R. Chen, I. C. Paschalidis, *et al.*, "Distributionally robust learning," *Foundations and Trends® in Optimization*, vol. 4, no. 1-2, pp. 1–243, 2020.

[67] J.-C. Bertein, R. Ceschi, and J.-C. Bertein, *Discrete Stochastic Processes and Optimal Filtering*. Wiley Online Library, 2007.

[68] A. W. Van der Vaart, *Asymptotic Statistics*. Cambridge University Press, 1998.

[69] V. A. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, "Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization," *Mathematics of Operations Research*, 2021.

[70] I. R. Petersen and A. V. Savkin, *Robust Kalman Filtering for Signals and Systems with Large Uncertainties*. Springer Science & Business Media, 1999.

[71] M. A. Gandhi and L. Mili, "Robust Kalman filter based on a generalized maximum-likelihood-type estimator," *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2509–2520, 2010.

[72] G. Gawrys and V. Vandelinde, "Divergence and the fading memory filter," in *1975 IEEE Conference on Decision and Control including the 14th Symposium on Adaptive Processes*, pp. 66–68, IEEE, 1975.

[73] B. Hassibi, A. H. Sayed, and T. Kailath, "Linear estimation in Krein spaces. II. applications," *IEEE Transactions on Automatic Control*, vol. 41, no. 1, pp. 34–49, 1996.

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

[75] J. Speyer, J. Deyst, and D. Jacobson, "Optimization of stochastic linear systems with additive measurement and process noise using exponential performance criteria," *IEEE Transactions on Automatic Control*, vol. 19, no. 4, pp. 358–366, 1974.

[76] D. Bertsekas and I. Rhodes, "Recursive state estimation for a set-membership description of uncertainty," *IEEE Transactions on Automatic Control*, vol. 16, no. 2, pp. 117–128, 1971.

[77] X. Shen and L. Deng, "Game theory approach to discrete H_∞ filter design," *IEEE Transactions Signal Processing*, vol. 45, pp. 1092–1095, 1997.

[78] R. Mehra, "On the identification of variances and adaptive Kalman filtering," *IEEE Transactions on Automatic Control*, vol. 15, no. 2, pp. 175–184, 1970.

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

[80] Y. Huang, Y. Zhang, Z. Wu, N. Li, and J. Chambers, "A novel adaptive Kalman filter with inaccurate process and measurement noise covariance matrices," *IEEE Transactions on Automatic Control*, vol. 63, no. 2, pp. 594–601, 2018.

[81] Y. Huang, Y. Zhang, P. Shi, and J. Chambers, "Variational adaptive Kalman filter with Gaussian-inverse-Wishart mixture distribution," *IEEE Transactions on Automatic Control*, 2020.

[82] E. Mazor, A. Averbuch, Y. Bar-Shalom, and J. Dayan, "Interacting multiple model methods in target tracking: a survey," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 34, no. 1, pp. 103–123, 1998.

[83] Y. Ma, S. Zhao, and B. Huang, "Multiple-model state estimation based on variational Bayesian inference," *IEEE Transactions on Automatic Control*, vol. 64, no. 4, pp. 1679–1685, 2018.

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

[85] 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.

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

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

[88] W. Liu and P. Shi, "Convergence of optimal linear estimator with multiplicative and time-correlated additive measurement noises," *IEEE Transactions on Automatic Control*, vol. 64, no. 5, pp. 2190–2197, 2018.

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

[90] M. Zorzi, "Robust Kalman filtering under model perturbations," *IEEE Transactions on Automatic Control*, vol. 62, no. 6, pp. 2902–2907, 2017.

[91] S. S. Abadeh, V. A. Nguyen, D. Kuhn, and P. M. M. Esfahani, "Wasserstein distributionally robust Kalman filtering," in *Advances in Neural Information Processing Systems*, pp. 8474–8483, 2018.

[92] H. W. Sorenson and D. L. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 7, no. 4, pp. 465–479, 1971.

[93] Y. Huang, Y. Zhang, Z. Wu, N. Li, and J. Chambers, "A novel adaptive kalman filter with inaccurate process and measurement noise covariance matrices," *IEEE Transactions on Automatic Control*, vol. 63, no. 2, pp. 594–601, 2017.

[94] Y. Huang, Y. Zhang, Y. Zhao, and J. A. Chambers, "A novel robust gaussian-student's t mixture distribution based Kalman filter," *IEEE Transactions on Signal Processing*, vol. 67, no. 13, pp. 3606–3620, 2019.

[95] L. Sun, W. K. Ho, K. V. Ling, T. Chen, and J. Maciejowski, "Recursive maximum likelihood estimation with t-distribution noise model," *Automatica*, vol. 132, p. 109789, 2021.

[96] J. Neri, P. Depalle, and R. Badeau, "Approximate inference and learning of state space models with Laplace noise," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3176–3189, 2021.

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

[98] Z. M. Durovic and B. D. Kovacevic, "Robust estimation with unknown noise statistics," *IEEE Transactions on Automatic Control*, vol. 44, no. 6, pp. 1292–1296, 1999.

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

[100] F. R. Hampel, "The influence curve and its role in robust estimation," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 383–393, 1974.

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

[102] D. A. Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions*. Wiley, 1954.

[103] K. Kim and G. Shevlyakov, "Why Gaussianity?," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 102–113, 2008.

[104] S. N. Diggavi and T. M. Cover, "The worst additive noise under a covariance constraint," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 3072–3081, 2001.

[105] D. Guo, Y. Wu, S. S. Shitz, and S. Verdú, "Estimation in Gaussian noise: Properties of the minimum mean-square error," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2371–2385, 2011.

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

[107] 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.

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

[109] M. Salvoldi and D. Choukroun, "Process noise covariance design in Kalman filtering via bounds optimization," *IEEE Transactions on Automatic Control*, vol. 64, no. 2, pp. 834–840, 2018.

[110] 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.

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

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

[113] 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.

[114] 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 Transactions on Signal Processing*, vol. 54, no. 6, pp. 2110–2118, 2006.

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

[116] 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," *Mechanical Systems and Signal Processing*, vol. 151, p. 107353, 2021.

[117] A. Subramanian and A. H. Sayed, "A robust minimum-variance filter for time varying uncertain discrete-time systems," in *Proceedings of the 2003 American Control Conference, 2003.*, vol. 3, pp. 1885–1889, IEEE, 2003.

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

[119] 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.

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

[121] G. Li and Y. Gu, "Restricted isometry property of gaussian random projection for finite set of subspaces," *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1705–1720, 2017.

[122] L. V. Kantorovich and S. Rubinshtein, "On a space of totally additive functions," *Vestnik of the St. Petersburg University: Mathematics*, vol. 13, no. 7, pp. 52–59, 1958.

[123] C. Villani, *Topics in Optimal Transportation*, vol. 58. American Mathematical Society, 2003.

[124] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, pp. 214–223, PMLR, 2017.

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

[126] P. J. Huber, "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73 – 101, 1964.

[127] Z. Li, Y. Yao, J. Wang, and J. Gao, "Application of improved robust Kalman filter in data fusion for PPP/INS tightly coupled positioning system," *Metrology and Measurement Systems*, vol. 24, no. 2, 2017.

[128] F. Hampel, *Contributions to the Theory of Robust Estimation.* PhD thesis, University of California, Berkeley, Sep. 1968.

[129] B. Chen, X. Liu, H. Zhao, and J. C. Principe, "Maximum correntropy Kalman filter," *Automatica*, vol. 76, pp. 70–77, 2017.

[130] Y. Yuanxi, "Robust estimation for dependent observations," *Manuscripta Geodaetica*, vol. 19, no. 1, pp. 10–17, 1994.

[131] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[132] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* John Wiley & Sons, 1991.

[133] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Transactions on Information Theory*, vol. 26, no. 1, pp. 26–37, 1980.

[134] P. D. Grünwald and A. P. Dawid, "Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory," *The Annals of Statistics*, vol. 32, no. 4, pp. 1367–1433, 2004.

[135] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis (2nd. edition).* Springer Science & Business Media, 1985.

[136] J. O. Berger, E. Moreno, L. R. Pericchi, M. J. Bayarri, J. M. Bernardo, J. A. Cano, J. De la Horra, J. Martín, D. Ríos-Insúa, B. Betrò, *et al.*, "An overview of robust bayesian analysis," *Test*, vol. 3, no. 1, pp. 5–124, 1994.

[137] C. P. Robert *et al.*, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation (2nd. edition)*, vol. 2. Springer, 2007.

[138] D. Z. Long, M. Sim, and M. Zhou, "Robust satisficing," *Operations Research*, 2022.

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

[140] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

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

[142] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[143] L. F. Shampine, "Vectorized adaptive quadrature in MATLAB," *Journal of Computational and Applied Mathematics*, vol. 211, no. 2, pp. 131–140, 2008.

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

[145] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, 1992.

[146] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

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

[148] M. Staib and S. Jegelka, "Distributionally robust optimization and generalization in kernel methods," in *Advances in Neural Information Processing Systems*, pp. 9134–9144, 2019.

[149] 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.

[150] X. Shen and P. K. Varshney, "Sensor selection based on generalized information gain for target tracking in large sensor networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 2, pp. 363–375, 2013.

[151] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 61–80, 2012.

Appendix A

Preliminaries

A.1 Distributionally Robust Optimization

Distributionally robust optimization, originating from statistical game theory (cf. mixed strategy) [102] and robust statistics [126], is currently popular in academic communities, such as the fields of operations research [147], machine learning [148], and systems control [65]. 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 $\boldsymbol{\Xi}$. In many application scenarios, when $\boldsymbol{\xi}$ is random, 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

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}_{\boldsymbol{\xi}}\in\mathcal{F}} \mathbb{E}[f(\boldsymbol{x},\boldsymbol{\xi})], \tag{A.1}$$

where the expectation is taken over $\mathbb{P}_{\boldsymbol{\xi}}$ and $f(\cdot,\cdot)$ is the objective function. Here, $\mathcal{F}$ is termed as an **ambiguity set**. Thus, the ambiguity set $\mathcal{F}$ forms a distributional uncertainty space for the modeling uncertainties of the objective function $f(\cdot,\cdot)$. Typically, $\mathcal{F}$ can be constructed using the moments of $\boldsymbol{\xi}$ [120] or a metric/divergence of distributions such as the Kullback–Leibler (KL) divergence [125]

$$\mathcal{F}_{\boldsymbol{\xi}}(\theta) = \left\{ \mathbb{P}_{\boldsymbol{\xi}} \in \mathcal{P}(\boldsymbol{\Xi}) \,\middle|\, \mathrm{KL}(\mathbb{P}_{\boldsymbol{\xi}} \| \bar{\mathbb{P}}_{\boldsymbol{\xi}}) \le \theta \right\}, \tag{A.2}$$

or the Wasserstein metric [56]

$$\mathcal{F}_{\boldsymbol{\xi}}(\theta) = \left\{ \mathbb{P}_{\boldsymbol{\xi}} \in \mathcal{P}(\boldsymbol{\Xi}) \,\middle|\, \mathrm{W}(\mathbb{P}_{\boldsymbol{\xi}}, \bar{\mathbb{P}}_{\boldsymbol{\xi}}) \le \theta \right\}, \tag{A.3}$$

or others including the τ-divergence [90], ϕ-divergence [125], $\alpha/\beta/\gamma$-divergence [149], 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, $\mathcal{F}_{\boldsymbol{\xi}}(\theta)$ means 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 the radius of θ.

Suppose that $\boldsymbol{x}^*$ and $\mathbb{P}^*_{\boldsymbol{\xi}}$ solve the distributionally robust optimization problem (A.1). We term $\boldsymbol{x}^*$ the worst-case robust solution and $\mathbb{P}^*_{\boldsymbol{\xi}}$ the least-favorable (i.e., worst-case) distribution.

A.2 Optimal Estimation

The linear system model (2.1) induces two stochastic vector processes $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$, where $k = 0, 1, 2, \cdots$. For every k, $\mathbf{x}_k$ and $\mathbf{y}_k$ have finite second moments. Note that $\mathbf{y}_0 := \mathbf{0}$ (i.e., a degeneration distribution). Let $\mathcal{H}'_{\mathcal{Y}_k}$ denote a collection of functions (i.e., estimators) defined by

$$\mathcal{H}'_{\mathcal{Y}_k} := \left\{ \boldsymbol{\phi}(\mathbf{y}_1, ..., \mathbf{y}_k) \,\middle|\, \begin{array}{l} \boldsymbol{\phi} : \underbrace{\mathbb{R}^m \times \cdots \times \mathbb{R}^m}_{k} \to \mathbb{R}^n \\ \\ \boldsymbol{\phi} \text{ is Borel-measurable} \\ \\ \displaystyle\int_{(\mathbb{R}^m)^k} [\boldsymbol{\phi}(\boldsymbol{Y}_k)]^\top [\boldsymbol{\phi}(\boldsymbol{Y}_k)] \mathrm{d}\mathbb{P}_{\mathcal{Y}_k}(\boldsymbol{Y}_k) < \infty \end{array} \right\}. \tag{A.4}$$

Meanwhile, let $\mathcal{H}_{\mathcal{Y}_k}$ denote a set of linear functions of $\mathbf{1}$ and $\mathcal{Y}_k$ (i.e., linear estimators):

$$\begin{aligned} \mathcal{H}_{\mathcal{Y}_k} &:= \left\{ \boldsymbol{B}_k \mathbf{1} + \textstyle\sum_{i=1}^k \boldsymbol{A}_i \mathbf{y}_i \,\middle|\, \boldsymbol{B}_k, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_k \in \mathbb{R}^{n \times m} \right\}, \\ &= \left\{ \boldsymbol{b}_k + \textstyle\sum_{i=1}^k \boldsymbol{A}_i \mathbf{y}_i \,\middle|\, \boldsymbol{b}_k \in \mathbb{R}^n, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_k \in \mathbb{R}^{n \times m} \right\}. \end{aligned} \tag{A.5}$$

It is known that the optimal estimate of $\mathbf{x}_k$ given $\mathcal{Y}_k$ in the sense of minimum mean square error is the unique orthogonal projection of $\mathbf{x}_k$ onto $\mathcal{H}'_{\mathcal{Y}_k}$ [67, 68]. For the special case when $\{\boldsymbol{x}_k\} \cup \{\boldsymbol{y}_k\}$ are jointly Gaussian, the optimal estimate of $\mathbf{x}_k$ given $\mathcal{Y}_k$ in the sense of minimum mean square error is the unique orthogonal projection of $\mathbf{x}_k$ onto $\mathcal{H}_{\mathcal{Y}_k}$. However, no matter whether it is Gaussian or not, the unique orthogonal projection of $\mathbf{x}_k$ onto $\mathcal{H}_{\mathcal{Y}_k}$ gives the optimal **linear** estimation [67]. In view of the optimal Bayesian posterior estimation theory [11, 25] (cf. Sherman's theorem), this projection point is the same as the conditional mean of $\mathbf{x}_k$ given $\mathcal{Y}_k$, i.e., $\hat{\mathbf{x}}_k = \mathbb{E}(\mathbf{x}_k \mid \mathcal{Y}_k)$, which minimizes the mean square estimation error [28]

$$\hat{\mathbf{x}}_k = \operatorname*{arginf}_{\boldsymbol{\phi} \in \mathcal{H}'_{\mathcal{Y}_k}} \operatorname{Tr} \mathbb{E}[\mathbf{x}_k - \boldsymbol{\phi}(\mathcal{Y}_k)][\mathbf{x}_k - \boldsymbol{\phi}(\mathcal{Y}_k)]^\top, \tag{A.6}$$

where the expectation is taken over $\mathbb{P}_{\mathbf{x}_k, \mathcal{Y}_k}$.

In particular, in the linear case (e.g., jointly Gaussian), this optimal Bayesian estimator (i.e., the conditional mean) admits a linear form [28]

$$\hat{\mathbf{x}}_k = \bar{\boldsymbol{x}}_k + \boldsymbol{\Sigma}_{xY,k} \boldsymbol{\Sigma}_{YY,k}^{-1} \left[\mathcal{Y}_k - \bar{\boldsymbol{Y}}_k \right], \tag{A.7}$$

where $\bar{\boldsymbol{x}}_k$ and $\bar{\boldsymbol{Y}}_k$ are *a priori* expectations of $\mathbf{x}_k$ and $\mathcal{Y}_k$, respectively,

$$\boldsymbol{\Sigma}_{xY,k} := \mathbb{E} \left(\mathbf{x}_k - \bar{\boldsymbol{x}}_k \right) \left(\mathcal{Y}_k - \bar{\boldsymbol{Y}}_k \right)^\top,$$

and $\boldsymbol{\Sigma}_{YY,k} := \mathbb{E} \left(\mathcal{Y}_k - \bar{\boldsymbol{Y}}_k \right) \left(\mathcal{Y}_k - \bar{\boldsymbol{Y}}_k \right)^\top$. With a slight abuse of notation, we note that in (A.7),

$\mathcal{Y}_k - \bar{\boldsymbol{Y}}_k := \mathrm{col}\{\mathbf{y}_i - \bar{\boldsymbol{y}}_i\}_{0\le i\le k}$,

$$\mathbb{E}\mathbf{x}_k\mathcal{Y}_k^\top := \left[\mathbb{E}\mathbf{x}_k\mathbf{y}_0^\top, \mathbb{E}\mathbf{x}_k\mathbf{y}_1^\top, ..., \mathbb{E}\mathbf{x}_k\mathbf{y}_k^\top\right],$$

and

$$\mathbb{E}\mathcal{Y}_k\mathcal{Y}_k^\top := \left[\mathbb{E}\mathbf{y}_i\mathbf{y}_j^\top\right]_{0\le i,j\le k}.$$

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

$$\mathbb{E}\left(\mathbf{x}_k - \hat{\mathbf{x}}_k\right)\left(\mathbf{x}_k - \hat{\mathbf{x}}_k\right)^\top = \boldsymbol{\Sigma}_{xx,k} - \boldsymbol{\Sigma}_{xY,k}\boldsymbol{\Sigma}_{YY,k}^{-1}\boldsymbol{\Sigma}_{Yx,k}, \tag{A.8}$$

where $\boldsymbol{\Sigma}_{Yx,k} = \boldsymbol{\Sigma}_{xY,k}^\top$, $\boldsymbol{\Sigma}_{xx,k} := \mathbb{E}\left(\mathbf{x}_k - \bar{\boldsymbol{x}}_k\right)\left(\mathbf{x}_k - \bar{\boldsymbol{x}}_k\right)^\top$, and the expectation is taken over $\mathbb{P}_{\mathbf{x}_k,\mathcal{Y}_k}$. Eq. (A.8) implies that the introduction of the information of $\mathbf{x}_k$ from $\mathcal{Y}_k$ helps reduce (resp. improve) the estimation error (resp. performance) of $\mathbf{x}_k$. In contrast, if $\mathbf{x}_k$ is statistically independent of $\mathcal{Y}_k$, we have $\boldsymbol{\Sigma}_{xY,k} \equiv \mathbf{0}$, admitting $\hat{\mathbf{x}}_k = \bar{\boldsymbol{x}}_k$ and $\mathbb{E}\left(\mathbf{x}_k - \hat{\mathbf{x}}_k\right)\left(\mathbf{x}_k - \hat{\mathbf{x}}_k\right)^\top = \boldsymbol{\Sigma}_{xx,k}$; i.e., there is no improvement in estimation performance after introducing $\mathcal{Y}_k$.

However, as a state estimation problem, the measurements $\mathbf{y}_k$ are available in sequence one by one, not in block as $\mathcal{Y}_k$. Therefore, we need to design a time-incremental version [89] (i.e., recursive form [28]) of the optimal estimator (A.7). Namely,

$$\inf_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}_k}} \mathrm{Tr}\,\mathbb{E}_{\mathcal{Y}_{k-1}}\mathbb{E}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}\left\{\left[\mathbf{x}_k - \boldsymbol{\phi}(\mathbf{y}_k)\right]\left[\mathbf{x}_k - \boldsymbol{\phi}(\mathbf{y}_k)\right]^\top \middle| \mathcal{Y}_{k-1}\right\}, \tag{A.9}$$

where the inner expectation is taken over the joint distribution of $\mathbf{x}_k$ and $\mathbf{y}_k$ conditioned on the past measurements $\mathcal{Y}_{k-1}$, i.e., $\mathbb{P}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$, and the outer expectation is taken over $\mathbb{P}_{\mathcal{Y}_{k-1}}$. Formally, $\mathbb{P}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$ is a random measure (due to the randomness of $\mathcal{Y}_{k-1}$) and is called a Markov kernel or a probability kernel: for every fixed $\boldsymbol{Y}_{k-1}$, $\mathbb{P}_{\mathbf{x}_k,\mathbf{y}_k|\boldsymbol{Y}_{k-1}}$ is a probability measure. Note that $\mathcal{H}'_{\mathbf{y}_k}$ is different from $\mathcal{H}'_{\mathcal{Y}_k}$. In fact, the optimization problem (A.9) and the problem

$$\inf_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}_k}} \mathrm{Tr}\,\mathbb{E}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}\left\{\left[\mathbf{x}_k - \boldsymbol{\phi}(\mathbf{y}_k)\right]\left[\mathbf{x}_k - \boldsymbol{\phi}(\mathbf{y}_k)\right]^\top \middle| \mathcal{Y}_{k-1}\right\}, \tag{A.10}$$

have the same solution, i.e., $\hat{\mathbf{x}}_k = \mathbb{E}(\mathbf{x}_k|\mathcal{Y}_k)$. The fact that $\hat{\mathbf{x}}_k = \mathbb{E}(\mathbf{x}_k|\mathcal{Y}_k)$ is the solution to (A.10) is obvious: for a given distribution $\mathbb{P}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$, the minimizer of (A.10) is $\hat{\mathbf{x}}_k = \mathbb{E}(\mathbf{x}_k|\mathbf{y}_k,\mathcal{Y}_{k-1})$. Further, we can show that any minimizer of (A.10) also solves (A.9); see [68, Example 11.5].

According to [28], (A.9) is equivalent to (A.6). Therefore, $\hat{\mathbf{x}}_k$ in (A.7) also reads

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

where $\bar{\mathbf{x}}_k$ and $\bar{\mathbf{y}}_k$ are **conditional** *a priori* expectations of $\mathbf{x}_k$ and $\mathbf{y}_k$ given $\mathcal{Y}_{k-1}$, respectively, which are also random vectors. However, they are non-random in terms of $\mathbf{x}_k$ and $\mathbf{y}_k$ (whenever $\boldsymbol{Y}_{k-1}$ is specified, $\bar{\mathbf{x}}_k$ and $\bar{\mathbf{y}}_k$ become deterministic); $\boldsymbol{\Sigma}_{xy,k} := \mathbb{E}\left\{(\mathbf{x}_k - \bar{\mathbf{x}}_k)(\mathbf{y}_k - \bar{\mathbf{y}}_k)^\top \middle| \mathcal{Y}_{k-1}\right\}$;

$\boldsymbol{\Sigma}_{yy,k} := \mathbb{E}\left\{(\mathbf{y}_k - \bar{\mathbf{y}}_k)(\mathbf{y}_k - \bar{\mathbf{y}}_k)^\top \middle| \mathcal{Y}_{k-1}\right\}$. Hence, $\bar{\mathbf{x}}_k = \mathbb{E}(\mathbf{x}_k|\mathcal{Y}_{k-1}) = \boldsymbol{F}_{k-1}\hat{\mathbf{x}}_{k-1}$ and $\bar{\mathbf{y}}_k = \mathbb{E}(\mathbf{y}_k|\mathcal{Y}_{k-1}) = \boldsymbol{H}_k\boldsymbol{F}_{k-1}\hat{\mathbf{x}}_{k-1}$, leading (A.11) to

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

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

$$\mathbb{E}\left\{(\mathbf{x}_k - \hat{\mathbf{x}}_k)(\mathbf{x}_k - \hat{\mathbf{x}}_k)^\top \middle| \mathcal{Y}_{k-1}\right\} = \boldsymbol{\Sigma}_{xx,k} - \boldsymbol{\Sigma}_{xy,k}\boldsymbol{\Sigma}_{yy,k}^{-1}\boldsymbol{\Sigma}_{yx,k}, \tag{A.13}$$

where $\boldsymbol{\Sigma}_{yx,k} = \boldsymbol{\Sigma}_{xy,k}^\top$ and the prior estimation error conditioned on $\mathcal{Y}_{k-1}$ is

$$\boldsymbol{\Sigma}_{xx,k} := \mathbb{E}\left\{(\mathbf{x}_k - \bar{\mathbf{x}}_k)(\mathbf{x}_k - \bar{\mathbf{x}}_k)^\top \middle| \mathcal{Y}_{k-1}\right\}.$$

If we take the expectation of (A.13) over $\mathbb{P}_{\mathcal{Y}_{k-1}}$, we can recover the overall minimum mean square estimation error in (A.8). Note that the value in (A.13) is random due to the random sequence $\mathcal{Y}_{k-1}$, while that in (A.8) is deterministic. Since the estimator (A.12) is unbiased, the minimum mean square estimation error matrix coincides with the minimum estimation error covariance matrix. The filter (A.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.

A.3 On Matrix-Type Objective

In the state estimation literature, some people directly work on minimizing a covariance matrix (see, e.g., [28] and [24, Chapter 3]), while others work on minimizing its trace (see, e.g., [18, 19]). In fact, minimizing a matrix objective is equivalent to minimizing its trace [28, 150]. Note that $\min_{\boldsymbol{X}\in\mathcal{X}} \boldsymbol{X}$ and $\min_{\boldsymbol{X}\in\mathcal{X}} \mathrm{Tr}\,[\boldsymbol{X}]$ over a convex and compact matrix set $\mathcal{X}$ have the same matrix-valued solution $\boldsymbol{X}^*$ because the trace operator is monotonically increasing. In this thesis, we study on minimizing the traces of estimation error covariance matrices.

A.4 Some Statistical Concepts

Suppose the density of interest $p(\mu;\theta)$ is parameterized by unknown mean θ. For mean estimation (a.k.a. location estimation [126]) problems, in general, $p(\mu;\theta) = p(\mu-\theta)$; recall, e.g., the Gaussian distribution $p(\mu;\theta) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(\mu-\theta)^2\right]$ supposing the variance is unit. Strictly speaking, the score function is defined with respect to the unknown parameter θ as $\frac{d}{d\theta}\ln p(\mu;\theta)$. Since $\frac{d}{d\theta}\ln p(\mu;\theta) = -\frac{d}{d\mu}\ln p(\mu;\theta)$, in statistics, some authors also directly define the score function with respect to μ as $-\frac{d}{d\mu}\ln p(\mu;\theta)$. As a result, the Fisher information has two equivalent definitions as well: $\mathbb{E}[-\frac{d^2}{d\theta^2}\ln p(\mu;\theta)]$ and $\mathbb{E}[-\frac{d^2}{d\mu^2}\ln p(\mu;\theta)]$.

In statistics, the three concepts, score function, influence function, and weight function, are

closely related but different. Score function is well-known in maximum likelihood estimation, influence function in general (outlier-) robust statistics [100], [55, Chap. 3], and weight function in (outlier-) robust linear regression [71, 98, 151], [55, Chap. 7]. Influence function is a property of an estimator designed for a distribution, while score function is that of the distribution itself. However, in M-estimation, influence function is just a multiple of score function and the constant multiplier is the Fisher information associated with the distribution. Let $T_\theta(\mathbb{P}_\mathrm{u})$ be the M-estimator of the mean of the distribution $\mathbb{P}_\mathrm{u}$ whose density is $p_\mathrm{u}(\mu)$. Supposing a score function is given by $\psi(\mu) := -\frac{d}{d\mu}\ln p(\mu)$, the influence function $IF(\mu)$ equals to [100]

$$\begin{aligned} IF(\mu) &:= \lim_{\epsilon\downarrow 0}\frac{T_\theta[(1-\epsilon)\mathbb{P}_\mathrm{u}+\epsilon\Delta_\mu]-T_\theta[\mathbb{P}_\mathrm{u}]}{\epsilon} \\ &= \frac{\psi(\mu)}{-\int\psi'(\mu)p(\mu)d\mu}, \end{aligned}$$

where Δ_μ is a point mass distribution concentrated at μ, $\psi'(\cdot)$ is the derivative of $\psi(\cdot)$, and the denominator is the Fisher information. In particular, if the Fisher information of the distribution is unit (e.g., standard Gaussian), the score function coincides with the influence function. For this reason, in M-estimation contexts, practitioners first derive score function and then equate it to influence function because a score function is mathematically easier to obtain. On the other hand, the weight function in (outlier-) robust linear regression is defined by $\psi(\mu)/\mu$. In statistical theory the three concepts are distinguished because they have different backgrounds, meanings, and definitions, but in signal processing practice we consider them to be equivalent (in the sense that one uniquely implies another) because they have similar mathematical forms. With this implication in mind, it is not confusing that the score function $\psi(\cdot)$ shown in Theorem 8 and Theorem 11 is directly termed as "influence function" in literature such as [21, 35, 97, 101]. This is more intuitively understandable for signal processing practitioners because $\psi(\cdot)$ limits the "influence" that a (contaminated) measurement $\boldsymbol{y}_k$ may bring to the estimator.

In M-estimation contexts, when we mention to design an influence function, we mean to design the score function $\psi(\cdot)$ [100]. Besides, when we design a weight function in robust linear regression contexts, we also uniquely obtain the corresponding score function in M-estimation counterpart [99]. The score function, in turn, implicitly determines the distribution for the studied population (which includes both ordinary points and outliers); $p(\mu)\propto\exp\left[-\int_{-\infty}^{\mu}\psi(\mu)d\mu\right]$ because $\psi(\mu)=-\frac{d}{d\mu}\ln p(\mu)$. For additional information, see Appendix A.5.

A.5 Formal Definitions for Terminologies in State Estimation

Bayesians and Frequentists have different philosophies towards statistical parameter estimation. Suppose a linear measurement system is given as $\mathrm{y}=\theta+\mathrm{v}$ where $\mathrm{y}\in\mathbb{R}$ is the measurement, $\theta\in\mathbb{R}$ is the unknown parameter to be estimated, and $\mathrm{v}\in\mathbb{R}$ is the noise term which is zero-mean Gaussian. Frequentist statisticians treat θ as a fixed number although it is unknown, and

then use collected samples $y_i,\ i = 1, 2, \cdots, N$ to estimate θ, during which properties, such as asymptotic variance, (Fisher) consistency, confidence interval, are also studied. In contrast, Bayesian statisticians treat θ as a random variable which has a prior distribution, say, $p(\theta)$, and then find the posterior distribution $p(\theta|y)$ based on the prior distribution $p(\theta)$ and measurement likelihood $p(y|\theta)$ [i.e., $p(\theta|y)$ (or its mean, mode, etc.) is the estimate of θ]. Therefore, in principle, a state estimation problem is more a Bayesian estimation problem than a Frequentist estimation problem. Likewise, the canonical Kalman filter is a Bayesian method; see its origin [25]. On the other hand, it is well-known that the maximum *a-posteriori* (MAP) estimation in Bayesians is equivalent to a regularized regression, while the maximum likelihood estimation in Frequentists is equivalent to a (conventional) regression; see [17, Chap. 3]. Since the term "influence function" was intentionally invented for Frequentist statistical inference problems [128], using influence functions are natural for robust (conventional) regression problems [55, Chap. 7; cf. Eq. (7.39)], especially when deriving the associated asymptotic variances [55, Chap. 7.6; cf. Eq. (7.78)]). In contrast, for Bayesian statistical inference problems, influence functions are applied over the innovation vectors (cf. Theorem 8 in this thesis), and therefore, the measurement residual term for robust regularized regression problems (cf. [99]). As we can see, in both Frequentists and Bayesians, an influence function can be understood as a qualitative robustness measure: it describes how much "influence" that a wild measurement may bring for an estimator.

Suppose an influence function $\psi(\mu) := -\frac{d}{d\mu} \ln p(\mu)$ is used. In Frequentists, the quantitative robustness (against measurement outliers) can be measured by the **minimax** asymptotic variance of a M-estimator [126], [55, pp. 11], i.e.,

$$\min_{\phi(\cdot)} \max_{p(\cdot)} V := \frac{1}{-\int \psi'(\mu) p(\mu) d\mu} = \frac{1}{\mathbb{E}\left[-\frac{d^2}{d\mu^2} \ln p(\mu)\right]},$$

where $\phi(\cdot)$ is the optimal robust estimator. In contrast, in Bayesians, the quantitative robustness (against measurement outliers) can be measured by the **minimax** posterior estimation error covariance [see (2.35), (2.38), and (2.43) in this thesis], i.e.,

$$\min_{\phi(\cdot)} \max_{p(\cdot)} \operatorname{Tr} \boldsymbol{P} := \operatorname{Tr} \left\{ \boldsymbol{M} - \boldsymbol{M}\boldsymbol{H}^\top \boldsymbol{S}^{-1} \boldsymbol{H}\boldsymbol{M} \cdot \mathbb{E}\left[-\frac{d^2}{d\mu^2} \ln p(\mu)\right] \right\}.$$

Note that in the definition of $\boldsymbol{P}$ above, $\boldsymbol{M}$ and $\boldsymbol{S}$ are exact (cf. Theorem 3 of this thesis) because we are only investigating the influence of measurement outliers. As we can see, both $\min\max V$ and $\min\max \boldsymbol{P}$ require smallest Fisher information $\mathbb{E}\left[-\frac{d^2}{d\mu^2} \ln p(\mu)\right]$. This implies another side contribution of this thesis, i.e., the design/definition of robustness measure (against measurement outliers) for Kalman-type filters. Note also that in Frequentists, the quantitative robustness measure is not limited to the asymptotic variance. It may also include other candidates such as breakdown point, gross-error-sensitivity, rejection point [100].

As above, we mentioned a term "Kalman-type". In state estimation community, two terms

"Kalman-type" and "Kalman-like" are widely used. However, the exact definition of "Kalman-type" has not been rigorously given in literature. We complement it in Definition 2.

Definition 2. *A state estimator is called "Kalman-type" or "Kalman-like", if the following three conditions are satisfied.*

1. *It is a Bayesian statistical inference method.*
2. *It is a closed-form solution at each time step (i.e., no numerical iterations are required).*
3. *It operates recursively along the time axis.* □

Although state estimation problems are Bayeisan statistical inference problems and Kalman-type filters are Bayesian methods (recall Definition 2), it does not imply that optimal state estimate for linear systems cannot be obtained by Frequentist methods. For example, [71, 98] amazingly reformulate a state estimation problem into a pure M-estimation problem (n.b., pure M-estimation is a Frequentist method). The merit of this reformulation is that the robustness measure (against measurement outliers) of the state estimator can be derived using the minimax asymptotic variance of the M-estimator [71, Eq. (38)], instead of the minimax posterior estimation error covariance. Therefore, strictly speaking, the state estimator proposed in [71] is not a Kalman-type filter because at each time step, a pure M-estimation (rather than a Bayesian estimation) problem is solved. Besides, the state estimator in [71, Eq. (21)] requires to numerically solve a nonlinear root-finding problem.

In state estimation literature, another term "M-estimation-based Kalman filter" is popular, which has not been rigorously defined as well. We complement it in Definition 3.

Definition 3. *A Kalman-type filter is called M-estimation-based if an influence function is applied over the innovation vector (or transformed/normalized innovation vector) to limit the "influence" that a wild measurement may bring. See, e.g., Theorem 8 in this thesis. Therefore, a Kalman-type filter is said to be measurement-outlier-robust if it is M-estimation-based.* □

In this sense, the state estimator proposed in Theorem 8 of this thesis is a M-estimation-based Kalman-type filter. Likewise, the state estimator proposed in [97] is also a M-estimation-based Kalman-type filter. However, state estimators in [71, 95] are not M-estimation-based Kalman-type filters, because they are pure Frequentist M-estimators, and therefore, not Bayesian Kalman-type. In this sense, names "M-type filter" and "M-type state estimator" should be more reasonable.

Appendix B

Proofs and Derivations in Chapter 2

B.1 Derive (2.5)

By (2.1), we have

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

namely,

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

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

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

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

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

and covariance of

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

where

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

In summary,

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

where

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

which is (2.5).

B.2 Proof of Theorem 1

Since the optimal estimator $\boldsymbol{\phi}(\cdot)$ is parameterized by $\boldsymbol{A}_k$ and $\boldsymbol{b}_k$, and the distributions $\mathbb{P}_{\mathbf{z}_k|\mathcal{Y}_{k-1}}$ in (2.11) are parameterized by $\boldsymbol{c}_k$ and $\boldsymbol{S}_k$, (2.4) is equivalent to the left-hand side of the equality in (2.13).

Let $\boldsymbol{S}_{xx,k} := \mathbb{E}\left(\mathbf{x}_k - \boldsymbol{c}_{x,k}\right)\left(\mathbf{x}_k - \boldsymbol{c}_{x,k}\right)^\top$, $\boldsymbol{S}_{yx,k}^\top = \boldsymbol{S}_{xy,k} := \mathbb{E}\left(\mathbf{x}_k - \boldsymbol{c}_{x,k}\right)\left(\mathbf{y}_k - \boldsymbol{c}_{y,k}\right)^\top$, $\boldsymbol{S}_{yy,k} := \mathbb{E}\left(\mathbf{y}_k - \boldsymbol{c}_{y,k}\right)\left(\mathbf{y}_k - \boldsymbol{c}_{y,k}\right)^\top$ where the three expectations are taken over $\mathbb{P}_{\mathbf{x}_k|\mathcal{Y}_{k-1}}$, $\mathbb{P}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$, and $\mathbb{P}_{\mathbf{y}_k|\mathcal{Y}_{k-1}}$, respectively. As a result, in particular by the definitions of $\boldsymbol{c}_k$ and $\boldsymbol{S}_k$ in (2.11), we have

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

Since $\boldsymbol{\Sigma}_k \succ \mathbf{0}$, we have $\boldsymbol{S}_k \succ \mathbf{0}$. By Schur complement, we further have $\boldsymbol{S}_{xx,k} \succ \mathbf{0}$ and $\boldsymbol{S}_{yy,k} \succ \mathbf{0}$.

This means that (2.15) is equivalent to

$$\begin{cases} \boldsymbol{S}_k \preceq \theta_2 \boldsymbol{\Sigma}_k, \\ \boldsymbol{S}_k \succeq \theta_1 \boldsymbol{\Sigma}_k. \end{cases} \tag{B.1}$$

With the affine optimal estimator (2.12), straightforward algebraic manipulations on the objective function of (2.4), i.e., $\operatorname{Tr}\mathbb{E}\left[\mathbf{x}_k - (\boldsymbol{A}_k\mathbf{y}_k + \boldsymbol{b}_k)\right]\left[\mathbf{x}_k - (\boldsymbol{A}_k\mathbf{y}_k + \boldsymbol{b}_k)\right]^\top$, gives

$$\begin{aligned} \min_{\boldsymbol{A}_k, \boldsymbol{b}_k} \max_{\boldsymbol{c}_k, \boldsymbol{S}_k} \; & \left\langle \boldsymbol{I}, \boldsymbol{S}_{xx,k} + \boldsymbol{c}_{x,k}\boldsymbol{c}_{x,k}^\top \right\rangle + \left\langle \boldsymbol{A}_k^\top \boldsymbol{A}_k, \boldsymbol{S}_{yy,k} + \boldsymbol{c}_{y,k}\boldsymbol{c}_{y,k}^\top \right\rangle - \left\langle \boldsymbol{A}_k, \boldsymbol{S}_{xy,k} + \boldsymbol{c}_{x,k}\boldsymbol{c}_{y,k}^\top \right\rangle \\ & - \left\langle \boldsymbol{A}_k^\top, \boldsymbol{S}_{yx,k} + \boldsymbol{c}_{y,k}\boldsymbol{c}_{x,k}^\top \right\rangle + 2\left\langle \boldsymbol{b}_k, \boldsymbol{A}_k\boldsymbol{c}_{y,k} - \boldsymbol{c}_{x,k} \right\rangle + \left\langle \boldsymbol{b}_k, \boldsymbol{b}_k \right\rangle. \end{aligned} \tag{B.2}$$

For details of derivation, see Appendix B.3. Hence, the problem (2.4) can be reformulated as solving (B.2) subject to (2.11). Since (B.2) 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.,

$$\boldsymbol{b}_k^\star = \boldsymbol{c}_{x,k} - \boldsymbol{A}_k\boldsymbol{c}_{y,k}. \tag{B.3}$$

This equality simplifies (B.2) to

$$\min_{\boldsymbol{A}_k} \max_{\boldsymbol{S}_k} \left\langle \boldsymbol{I}, \boldsymbol{S}_{xx,k} \right\rangle + \left\langle \boldsymbol{A}_k^\top \boldsymbol{A}_k, \boldsymbol{S}_{yy,k} \right\rangle - \left\langle \boldsymbol{A}_k, \boldsymbol{S}_{xy,k} \right\rangle - \left\langle \boldsymbol{A}_k^\top, \boldsymbol{S}_{yx,k} \right\rangle, \tag{B.4}$$

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

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

The objective function (B.4) can be further written in a compact form as

$$\min_{\boldsymbol{A}_k} \max_{\boldsymbol{S}_k} \left\langle \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{A}_k \\ -\boldsymbol{A}_k^\top & \boldsymbol{A}_k^\top \boldsymbol{A}_k \end{bmatrix}, \boldsymbol{S}_k \right\rangle, \tag{B.5}$$

which is subject to (2.11). To avoid notational clutter, we rewrite (2.11) as

$$\begin{cases} (\boldsymbol{c}_k - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{c}_k - \boldsymbol{\mu}_k) \leq \theta_3, \\ \boldsymbol{S}_k + (\boldsymbol{c}_k - \boldsymbol{\mu}_k)(\boldsymbol{c}_k - \boldsymbol{\mu}_k)^\top \preceq \theta_2 \boldsymbol{\Sigma}_k, \\ \boldsymbol{S}_k + (\boldsymbol{c}_k - \boldsymbol{\mu}_k)(\boldsymbol{c}_k - \boldsymbol{\mu}_k)^\top \succeq \theta_1 \boldsymbol{\Sigma}_k. \end{cases} \tag{B.6}$$

Since the ambiguity set (B.6) is convex and compact in terms of $(\boldsymbol{c}_k, \boldsymbol{S}_k)$ and the objective

function in (B.5)

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

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.,

$$\min_{\boldsymbol{A}_k} \max_{\boldsymbol{S}_k} \left\langle \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{A}_k \\ -\boldsymbol{A}_k^\top & \boldsymbol{A}_k^\top \boldsymbol{A}_k \end{bmatrix}, \boldsymbol{S}_k \right\rangle = \max_{\boldsymbol{S}_k} \min_{\boldsymbol{A}_k} \left\langle \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{A}_k \\ -\boldsymbol{A}_k^\top & \boldsymbol{A}_k^\top \boldsymbol{A}_k \end{bmatrix}, \boldsymbol{S}_k \right\rangle.$$

This gives the min-max equality (2.13). In view that the optimization problem (B.5) 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

$$\boldsymbol{A}_k^\star = \boldsymbol{S}_{xy,k} \cdot \boldsymbol{S}_{yy,k}^{-1}. \tag{B.7}$$

This equality simplifies (B.5) to (2.14). Note that the objective function (2.14) is irrelevant to $\boldsymbol{c}_k$. Therefore, to maximize (2.14), the larger the feasible set of $\boldsymbol{S}_k$, the better. This gives the optimal solution of $\boldsymbol{c}_k$ as

$$\boldsymbol{c}_k^\star = \boldsymbol{\mu}_k. \tag{B.8}$$

This equality simplifies (B.6) to (B.1), which is equivalent to (2.15). This completes the proof. □

B.3 Derive (B.2)

Note that all the expectations in this appendix are conditional on $\mathcal{Y}_{k-1}$: the involved distribution is $\mathbb{P}_{\mathbf{x}_k|\mathcal{Y}_{k-1}}$, $\mathbb{P}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$, or $\mathbb{P}_{\mathbf{y}_k|\mathcal{Y}_{k-1}}$, wherever it is needed. Recall that $\boldsymbol{c}_k$ and $\boldsymbol{S}_k$ are conditioned on $\mathcal{Y}_{k-1}$ but they are non-random in terms of $\mathbf{x}_k$ and $\mathbf{y}_k$.

We have[1]

$$\begin{aligned} \boldsymbol{S}_{xx,k} &:= \mathbb{E}(\mathbf{x}_k - \boldsymbol{c}_{x,k})(\mathbf{x}_k - \boldsymbol{c}_{x,k})^\top \\ &= \mathbb{E}\mathbf{x}_k\mathbf{x}_k^\top - \mathbb{E}\mathbf{x}_k\boldsymbol{c}_{x,k}^\top - \boldsymbol{c}_{x,k}\mathbb{E}\mathbf{x}_k^\top + \boldsymbol{c}_{x,k}\boldsymbol{c}_{x,k}^\top \\ &= \mathbb{E}\mathbf{x}_k\mathbf{x}_k^\top - \boldsymbol{c}_{x,k}\boldsymbol{c}_{x,k}^\top. \end{aligned}$$

[1] To put it strict, we should write $\boldsymbol{S}_{xx,k} := \mathbb{E}\left\{ (\mathbf{x}_k - \boldsymbol{c}_{x,k})(\mathbf{x}_k - \boldsymbol{c}_{x,k})^\top \middle| \mathcal{Y}_{k-1} \right\}$ and the expectation is taken over $\mathbb{P}_{\mathbf{x}_k|\mathcal{Y}_{k-1}}$. However, to avoid notational clutter, we do not explicitly write the full forms of conditional expectations. Always keeping in mind that all the expectations in this appendix are conditional on $\mathcal{Y}_{k-1}$.

Hence,

$$\mathbb{E}\mathbf{x}_k\mathbf{x}_k^\top = \boldsymbol{S}_{xx,k} + \boldsymbol{c}_{x,k}\boldsymbol{c}_{x,k}^\top.$$

Similarly, we have

$$\mathbb{E}\mathbf{x}_k\mathbf{y}_k^\top = \boldsymbol{S}_{xy,k} + \boldsymbol{c}_{x,k}\boldsymbol{c}_{y,k}^\top,$$

$$\mathbb{E}\mathbf{y}_k\mathbf{x}_k^\top = \boldsymbol{S}_{yx,k} + \boldsymbol{c}_{y,k}\boldsymbol{c}_{x,k}^\top,$$

and

$$\mathbb{E}\mathbf{y}_k\mathbf{y}_k^\top = \boldsymbol{S}_{yy,k} + \boldsymbol{c}_{y,k}\boldsymbol{c}_{y,k}^\top.$$

As a result, we have

$$\begin{aligned}
& \mathbb{E}\left[\mathbf{x}_k - (\boldsymbol{A}_k\mathbf{y}_k + \boldsymbol{b}_k)\right]\left[\mathbf{x}_k - (\boldsymbol{A}_k\mathbf{y}_k + \boldsymbol{b}_k)\right]^\top \\
= \;& \mathbb{E}\mathbf{x}_k\mathbf{x}_k^\top - \mathbb{E}\left[\boldsymbol{A}_k\mathbf{y}_k + \boldsymbol{b}_k\right]\left[\mathbf{x}_k\right]^\top - \mathbb{E}\left[\mathbf{x}_k\right]\left[\boldsymbol{A}_k\mathbf{y}_k + \boldsymbol{b}_k\right]^\top + \mathbb{E}\left[\boldsymbol{A}_k\mathbf{y}_k + \boldsymbol{b}_k\right]\left[\boldsymbol{A}_k\mathbf{y}_k + \boldsymbol{b}_k\right]^\top \\
= \;& \mathbb{E}\mathbf{x}_k\mathbf{x}_k^\top - \boldsymbol{A}_k \cdot (\mathbb{E}\mathbf{y}_k\mathbf{x}_k^\top) - \mathbb{E}\boldsymbol{b}_k\mathbf{x}_k^\top - (\mathbb{E}\mathbf{x}_k\mathbf{y}_k^\top)\boldsymbol{A}_k^\top - \mathbb{E}\mathbf{x}_k\boldsymbol{b}_k^\top + \boldsymbol{A}_k \cdot (\mathbb{E}\mathbf{y}_k\mathbf{y}_k^\top) \cdot \boldsymbol{A}_k^\top + \\
& \boldsymbol{A}_k \cdot (\mathbb{E}\mathbf{y}_k\boldsymbol{b}_k^\top) + (\mathbb{E}\boldsymbol{b}_k\mathbf{y}_k^\top)\boldsymbol{A}_k^\top + \mathbb{E}\boldsymbol{b}_k\boldsymbol{b}_k^\top \\
= \;& \mathbb{E}\mathbf{x}_k\mathbf{x}_k^\top + \boldsymbol{A}_k(\mathbb{E}\mathbf{y}_k\mathbf{y}_k^\top)\boldsymbol{A}_k^\top - (\mathbb{E}\mathbf{x}_k\mathbf{y}_k^\top)\boldsymbol{A}_k^\top - \boldsymbol{A}_k(\mathbb{E}\mathbf{y}_k\mathbf{x}_k^\top) - \mathbb{E}\boldsymbol{b}_k\mathbf{x}_k^\top - \mathbb{E}\mathbf{x}_k\boldsymbol{b}_k^\top + \\
& \boldsymbol{A}_k(\mathbb{E}\mathbf{y}_k\boldsymbol{b}_k^\top) + (\mathbb{E}\boldsymbol{b}_k\mathbf{y}_k^\top)\boldsymbol{A}_k^\top + \mathbb{E}\boldsymbol{b}_k\boldsymbol{b}_k^\top \\
= \;& \mathbb{E}\mathbf{x}_k\mathbf{x}_k^\top + \boldsymbol{A}_k(\mathbb{E}\mathbf{y}_k\mathbf{y}_k^\top)\boldsymbol{A}_k^\top - (\mathbb{E}\mathbf{x}_k\mathbf{y}_k^\top)\boldsymbol{A}_k^\top - \boldsymbol{A}_k(\mathbb{E}\mathbf{y}_k\mathbf{x}_k^\top) + \mathbb{E}\left(\boldsymbol{b}_k\right)\left(\boldsymbol{A}_k\mathbf{y}_k - \mathbf{x}_k\right)^\top + \\
& \mathbb{E}\left(\boldsymbol{A}_k\mathbf{y}_k - \mathbf{x}_k\right)\left(\boldsymbol{b}_k\right)^\top + \mathbb{E}\boldsymbol{b}_k\boldsymbol{b}_k^\top \\
= \;& (\boldsymbol{S}_{xx,k} + \boldsymbol{c}_{x,k}\boldsymbol{c}_{x,k}^\top) + \boldsymbol{A}_k(\boldsymbol{S}_{yy,k} + \boldsymbol{c}_{y,k}\boldsymbol{c}_{y,k}^\top)\boldsymbol{A}_k^\top - (\boldsymbol{S}_{xy,k} + \boldsymbol{c}_{x,k}\boldsymbol{c}_{y,k}^\top)\boldsymbol{A}_k^\top - \\
& \boldsymbol{A}_k(\boldsymbol{S}_{yx,k} + \boldsymbol{c}_{y,k}\boldsymbol{c}_{x,k}^\top) + 2(\boldsymbol{A}_k\boldsymbol{c}_{y,k} - \boldsymbol{c}_{x,k})\boldsymbol{b}_k^\top + \boldsymbol{b}_k\boldsymbol{b}_k^\top.
\end{aligned} \tag{B.9}$$

Applying the trace operator on the both sides of (B.9) gives (B.2). Note that

$$\mathrm{Tr}\left[\boldsymbol{A}_k(\boldsymbol{S}_{yy,k} + \boldsymbol{c}_{y,k}\boldsymbol{c}_{y,k}^\top)\boldsymbol{A}_k^\top\right] = \mathrm{Tr}\left[\boldsymbol{A}_k^\top\boldsymbol{A}_k(\boldsymbol{S}_{yy,k} + \boldsymbol{c}_{y,k}\boldsymbol{c}_{y,k}^\top)\right].$$

B.4 Proof of Theorem 2

The NSDP (2.14) subject to (2.15) is equivalent to

$$\max_{\boldsymbol{S}_k} \mathrm{Tr}\left[\boldsymbol{S}_{xx,k} - \boldsymbol{S}_{xy,k}\boldsymbol{S}_{yy,k}^{-1}\boldsymbol{S}_{yx,k}\right] \tag{B.10}$$

subject to (B.1). Let $f(\boldsymbol{S}_k) := \operatorname{Tr}\left[\boldsymbol{S}_{xx,k} - \boldsymbol{S}_{xy,k}\boldsymbol{S}_{yy,k}^{-1}\boldsymbol{S}_{yx,k}\right]$. The gradient of $f(\boldsymbol{S}_k)$ with respect to $\boldsymbol{S}_k$ admits

$$\nabla_{\boldsymbol{S}_k} f(\boldsymbol{S}_k) = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{S}_{xy,k}\boldsymbol{S}_{yy,k}^{-1} \\ -\boldsymbol{S}_{yy,k}^{-1}\boldsymbol{S}_{yx,k} & \boldsymbol{S}_{yy,k}^{-1}\boldsymbol{S}_{yx,k}\boldsymbol{S}_{xy,k}\boldsymbol{S}_{yy,k}^{-1} \end{bmatrix}. \tag{B.11}$$

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 (B.10) subject to (B.1), we must have $\boldsymbol{S}_k^\star = \theta_2 \boldsymbol{\Sigma}_k$. This completes the proof; see also Remarks 8 and 9 below. □

Remark 8. *In the proof of Theorem 2, the following facts are involved.*

1) *For a block matrix* $\boldsymbol{M} := \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$, *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.,* $\boldsymbol{C} = \boldsymbol{B}^\top$*) 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:* $\nabla_{\boldsymbol{S}} \operatorname{Tr}[\boldsymbol{S}] = \boldsymbol{I}$; $\nabla_{\boldsymbol{S}} \operatorname{Tr}[\boldsymbol{A}\boldsymbol{S}] = \nabla_{\boldsymbol{S}} \operatorname{Tr}[\boldsymbol{S}\boldsymbol{A}] = \boldsymbol{A}^\top$; *and* $\nabla_{\boldsymbol{S}} \operatorname{Tr}\left[\boldsymbol{A}^\top\boldsymbol{S}^{-1}\boldsymbol{A}\right] = \nabla_{\boldsymbol{S}} \operatorname{Tr}\left[\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{A}^\top\right] = -(\boldsymbol{S}^{-1})^\top(\boldsymbol{A}\boldsymbol{A}^\top)^\top(\boldsymbol{S}^{-1})^\top = -\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{S}^{-1}$. □

Remark 9. *Another proof to Theorem 2 can be obtained by analogy with Theorem 7 and Appendix B.9.* □

B.5 Proof of Theorem 3

Given a specific measurement $\boldsymbol{y}$, the conditional mean (i.e., optimal **estimate**) of $\mathbf{x}$ is

$$\begin{aligned} \hat{\boldsymbol{x}} &= \int \boldsymbol{x}\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{y})}\mathrm{d}\boldsymbol{x} \\ &= [p(\boldsymbol{y})]^{-1}\int \boldsymbol{x}p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \bar{\boldsymbol{x}} + \bar{\boldsymbol{x}} \\ &= [p(\boldsymbol{y})]^{-1}\int (\boldsymbol{x}-\bar{\boldsymbol{x}})p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \bar{\boldsymbol{x}} \\ &= [p(\boldsymbol{y})]^{-1}\boldsymbol{M}\int p_{\mathbf{v}}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})\boldsymbol{M}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \bar{\boldsymbol{x}}. \end{aligned}$$

Since the prior distribution of $\mathbf{x}$ is Gaussian, we have

$$-\boldsymbol{M}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}})p(\boldsymbol{x}) = \frac{\mathrm{d}p(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}},$$

giving

$$\hat{\boldsymbol{x}} = \bar{\boldsymbol{x}} - \boldsymbol{M}[p(\boldsymbol{y})]^{-1} \int p_{\mathbf{v}}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}) \frac{\mathrm{d}p(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} \mathrm{d}\boldsymbol{x}.$$

By partial integration, we have

$$\begin{aligned}
\hat{\boldsymbol{x}} &= \bar{\boldsymbol{x}} + \boldsymbol{M}[p(\boldsymbol{y})]^{-1} \int \frac{\partial p_{\mathbf{v}}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})}{\partial \boldsymbol{x}} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= \bar{\boldsymbol{x}} - \boldsymbol{M}\boldsymbol{H}^{\top}[p(\boldsymbol{y})]^{-1} \int \frac{\partial p_{\mathbf{v}}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})}{\partial \boldsymbol{y}} p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= \bar{\boldsymbol{x}} - \boldsymbol{M}\boldsymbol{H}^{\top}[p(\boldsymbol{y})]^{-1} \int \frac{\partial p(\boldsymbol{x}, \boldsymbol{y})}{\partial \boldsymbol{y}} \mathrm{d}\boldsymbol{x} \\
&= \bar{\boldsymbol{x}} - \boldsymbol{M}\boldsymbol{H}^{\top}[p(\boldsymbol{y})]^{-1} \frac{\mathrm{d}p(\boldsymbol{y})}{\mathrm{d}\boldsymbol{y}} \\
&= \bar{\boldsymbol{x}} + \boldsymbol{M}\boldsymbol{H}^{\top} \left[-\frac{\mathrm{d} \ln p(\boldsymbol{y})}{\mathrm{d}\boldsymbol{y}} \right].
\end{aligned}$$

Furthermore, the conditional covariance of the estimation error, which is evaluated over $\mathbb{P}_{\mathbf{x}|\boldsymbol{y}}$, is

$$\begin{aligned}
\mathbb{E}\left\{ (\hat{\boldsymbol{x}} - \mathbf{x})(\cdots)^{\top} \middle| \boldsymbol{y} \right\} &= \mathbb{E}\left\{ (\bar{\boldsymbol{x}} - \mathbf{x})(\cdots)^{\top} \middle| \boldsymbol{y} \right\} - \mathbb{E}\left\{ (\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}})(\cdots)^{\top} \middle| \boldsymbol{y} \right\} \\
&= \mathbb{E}\left\{ (\bar{\boldsymbol{x}} - \mathbf{x})(\cdots)^{\top} \middle| \boldsymbol{y} \right\} - \boldsymbol{M}\boldsymbol{H}^{\top} \left[-\frac{\mathrm{d} \ln p(\boldsymbol{y})}{\mathrm{d}\boldsymbol{y}} \right] [\cdots]^{\top} \boldsymbol{H}\boldsymbol{M}.
\end{aligned}$$

Since $\mathbf{u} = \boldsymbol{S}^{-1/2}(\mathbf{y} - \boldsymbol{H}\bar{\boldsymbol{x}})$, we have

$$\begin{aligned}
p_{\mathbf{u}}(\boldsymbol{\mu}) &= p_{\mathbf{y}}(\boldsymbol{S}^{1/2}\boldsymbol{\mu} + \boldsymbol{H}\bar{\boldsymbol{x}}) \cdot \det \left[\frac{\mathrm{d}(\boldsymbol{S}^{1/2}\boldsymbol{\mu} + \boldsymbol{H}\bar{\boldsymbol{x}})}{\mathrm{d}\boldsymbol{\mu}} \right] \\
&= p_{\mathbf{y}}(\boldsymbol{S}^{1/2}\boldsymbol{\mu} + \boldsymbol{H}\bar{\boldsymbol{x}}) \cdot \det(\boldsymbol{S}^{1/2}).
\end{aligned}$$

As a result,

$$-\frac{\mathrm{d} \ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}} = -\frac{\mathrm{d} \ln p_{\mathbf{y}}(\boldsymbol{S}^{1/2}\boldsymbol{\mu} + \boldsymbol{H}\bar{\boldsymbol{x}})}{\mathrm{d}\boldsymbol{\mu}} = -\boldsymbol{S}^{1/2} \frac{\mathrm{d} \ln p_{\mathbf{y}}(\boldsymbol{y})}{\mathrm{d}\boldsymbol{y}},$$

implying

$$-\frac{\mathrm{d} \ln p_{\mathbf{y}}(\boldsymbol{y})}{\mathrm{d}\boldsymbol{y}} = \boldsymbol{S}^{-1/2}[-\frac{\mathrm{d} \ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}].$$

Therefore,

$$[-\frac{\mathrm{d} \ln p_{\mathbf{y}}(\boldsymbol{y})}{\mathrm{d}\boldsymbol{y}}][\cdots]^{\top} = \boldsymbol{S}^{-1/2}[-\frac{\mathrm{d} \ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}][\cdots]^{\top} \boldsymbol{S}^{-1/2}.$$

Note that both $-\frac{\mathrm{d} \ln p_{\mathbf{y}}(\boldsymbol{y})}{\mathrm{d}\boldsymbol{y}}$ and $-\frac{\mathrm{d} \ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}$ are functions of $\boldsymbol{y}$. Hence, if $\boldsymbol{y}$ were not specified, $-\frac{\mathrm{d} \ln p_{\mathbf{y}}(\boldsymbol{y})}{\mathrm{d}\boldsymbol{y}}\Big|_{\boldsymbol{y}=\mathbf{y}}$ [2] and $-\frac{\mathrm{d} \ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}\Big|_{\boldsymbol{y}=\mathbf{y}}$ would be functions of the random vector $\mathbf{y}$. Meanwhile, $-\frac{\mathrm{d} \ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}$ is also a function of $\boldsymbol{\mu}$. Therefore, when $\boldsymbol{y}$ were not specified, $\boldsymbol{\mu} := \boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}$ were not

[2] The notation $-\frac{\mathrm{d} \ln p_{\mathbf{y}}(\boldsymbol{y})}{\mathrm{d}\boldsymbol{y}}\Big|_{\boldsymbol{y}=\mathbf{y}}$ means $\boldsymbol{y}$ is replaced with $\mathbf{y}$ in $-\frac{d \ln p(\boldsymbol{y})}{d\boldsymbol{y}}$ so that $-\frac{\mathrm{d} \ln p_{\mathbf{y}}(\boldsymbol{y})}{\mathrm{d}\boldsymbol{y}}\Big|_{\boldsymbol{y}=\mathbf{y}}$ is a function of the random vector $\mathbf{y}$.

specified as well, and $-\frac{\mathrm{d}\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}\Big|_{\boldsymbol{\mu}=\mathbf{u}}$ would also be a function of the random vector $\mathbf{u}$.

Consequently, the optimal **estimator** $\hat{\mathbf{x}}$ of $\mathbf{x}$ would be a function of the random vector $\mathbf{y}$: i.e.,

$$\hat{\mathbf{x}} = \bar{\boldsymbol{x}} + \boldsymbol{M}\boldsymbol{H}^\top \left[-\frac{\mathrm{d}\ln p(\boldsymbol{y})}{\mathrm{d}\boldsymbol{y}}\right]_{\boldsymbol{y}=\mathbf{y}},$$

and the associated estimation error covariance conditioned on $\mathbf{y}$

$$\mathbb{E}\left\{(\hat{\mathbf{x}}-\mathbf{x})(\cdots)^\top\Big|\mathbf{y}\right\} = \mathbb{E}\left\{(\bar{\boldsymbol{x}}-\mathbf{x})(\cdots)^\top\Big|\mathbf{y}\right\} - \boldsymbol{M}\boldsymbol{H}^\top\left[-\frac{\mathrm{d}\ln p(\boldsymbol{y})}{\mathrm{d}\boldsymbol{y}}\right]_{\boldsymbol{y}=\mathbf{y}}\left[-\frac{\mathrm{d}\ln p(\boldsymbol{y})}{\mathrm{d}\boldsymbol{y}}\right]_{\boldsymbol{y}=\mathbf{y}}^\top \boldsymbol{H}\boldsymbol{M}.$$

By noting that

$$\mathbb{E}\left\{[p(\boldsymbol{\mu})]^{-1}\frac{\mathrm{d}^2 p(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top}\Big|_{\boldsymbol{\mu}=\mathbf{u}}\right\} = \int [p(\boldsymbol{\mu})]^{-1}\frac{\mathrm{d}^2 p(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top}p(\boldsymbol{\mu})\mathrm{d}\boldsymbol{\mu} = \frac{\mathrm{d}^2\int p(\boldsymbol{\mu})\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top} = \frac{\mathrm{d}^2 1}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top} = \mathbf{0},$$

we have

$$\begin{aligned}
&\mathbb{E}\left\{-\frac{\mathrm{d}^2\ln p(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top}\Big|_{\boldsymbol{\mu}=\mathbf{u}}\right\} \\
&= -\mathbb{E}\left\{[p(\boldsymbol{\mu})]^{-1}\frac{\mathrm{d}^2 p(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top}\Big|_{\boldsymbol{\mu}=\mathbf{u}}\right\} + \mathbb{E}\left\{[p(\boldsymbol{\mu})]^{-2}\left[-\frac{\mathrm{d}p(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}\right]\left[-\frac{\mathrm{d}p(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}\right]^\top\Big|_{\boldsymbol{\mu}=\mathbf{u}}\right\} \\
&= \mathbb{E}\left\{\left[-\frac{\mathrm{d}\ln p(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}\right]\left[-\frac{\mathrm{d}\ln p(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}\right]^\top\Big|_{\boldsymbol{\mu}=\mathbf{u}}\right\},
\end{aligned}$$

where the expectations are taken over $\mathbb{P}_{\mathbf{u}}$. Combining the results above, we have

$$\hat{\mathbf{x}} = \bar{\boldsymbol{x}} + \boldsymbol{M}\boldsymbol{H}^\top\boldsymbol{S}^{-1/2}\left[-\frac{\mathrm{d}\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}\right]_{\boldsymbol{\mu}=\mathbf{u}},$$

and the associated estimation error covariance, evaluated over $\mathbb{P}_{\mathbf{x},\mathbf{u}}$,

$$\begin{aligned}
&\mathbb{E}(\hat{\mathbf{x}}-\mathbf{x})(\hat{\mathbf{x}}-\mathbf{x})^\top = \mathbb{E}_{\mathbf{y}}\mathbb{E}_{\mathbf{x}|\mathbf{y}}\left\{(\hat{\mathbf{x}}-\mathbf{x})(\cdots)^\top\big|\mathbf{y}\right\} \\
&= \boldsymbol{M} - \boldsymbol{M}\boldsymbol{H}^\top\boldsymbol{S}^{-1/2}\mathbb{E}\left\{\left[-\frac{\mathrm{d}\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}\right]\left[-\frac{\mathrm{d}\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}}\right]^\top\Big|_{\boldsymbol{\mu}=\mathbf{u}}\right\}\boldsymbol{S}^{-1/2}\boldsymbol{H}\boldsymbol{M} \\
&= \boldsymbol{M} - \boldsymbol{M}\boldsymbol{H}^\top\boldsymbol{S}^{-1/2}\mathbb{E}\left\{\left[-\frac{\mathrm{d}^2\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top}\right]_{\boldsymbol{\mu}=\mathbf{u}}\right\}\boldsymbol{S}^{-1/2}\boldsymbol{H}\boldsymbol{M}.
\end{aligned}$$

The expectations inside of the last two lines are taken over $\mathbb{P}_{\mathbf{u}}$. This completes the proof. □

B.6 Proof of Lemma 1

See [126, pp. 80] for the solution of $p(\mu)$. As a result, $\min \mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln p(\mu)\right] = \int_{-K}^{K} p(\mu)\mathrm{d}\mu = (1-\epsilon)\int_{-K}^{K}\mathrm{d}\Phi(\mu) = (1-\epsilon)[1-2\Phi(-K)]$. For any given ϵ, the value of K can be found in [126, Table I] or [55, Exhibit 4.3]. □

B.7 Proof of Lemma 2

See [126, pp. 91] for the solution of $p(\mu)$. As a result,

$$\begin{aligned}\min \mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln p(\mu)\right] &= 2\times\left[\int_0^a \frac{1}{2}\frac{c^2}{\cos^2(\frac{1}{2}c\mu)}p(\mu)\mathrm{d}\mu + \int_a^b p(\mu)\mathrm{d}\mu\right] \\ &= 2\left[\frac{1}{2}\frac{c^2}{\cos^2(\frac{1}{2}ca)}p(a)\int_0^a \mathrm{d}\mu + \int_a^b \mathrm{d}\Phi(\mu)\right].\end{aligned}$$

For any given $0 \leq \epsilon \lesssim 0.0303$, the values of a, b, and c can be found in [126, Table II] or [55, Exhibit 4.6]. □

B.8 Proof of Theorem 6

The squared constraint $\mathrm{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2(\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}})^{\frac{1}{2}}] \leq \theta_{\boldsymbol{x}}^2$ is convex and compact, so is the squared constraint for $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ (as $\boldsymbol{R} \succ \boldsymbol{0}$) [69]. Therefore, the following equivalent feasible set is convex and also compact.

$$\begin{cases} \mathrm{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2\left(\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \leq \theta_{\boldsymbol{x}}^2 \\ \mathrm{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{v}} + \boldsymbol{R} - 2\left(\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{R}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \leq \theta_{\boldsymbol{v}}^2 \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \boldsymbol{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \boldsymbol{0}. \end{cases} \tag{B.12}$$

Due to $\boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \boldsymbol{0}$, the existence of the inverse in the objective function (2.60) is guaranteed. As the trace of the objective (2.60) is continuous, smooth (i.e., differentiable), and joint concave in terms of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$, the problem (2.60) subject to (2.61) is solvable (i.e., the optimal solutions exist and are finite).

In order to simplify the objective function, let $\boldsymbol{U} \succeq \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \boldsymbol{0}$. By Schur complement, it is equivalent to require

$$\begin{bmatrix} \boldsymbol{U} & \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top \\ \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}} \end{bmatrix} \succeq \boldsymbol{0}.$$

In order to simplify the constraints, let $\boldsymbol{V}_{\boldsymbol{x}} \preceq (\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}})^{\frac{1}{2}}$, i.e., $\boldsymbol{V}_{\boldsymbol{x}}^2 \preceq \boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}}$. By Schur complement, it is equivalent to require

$$\begin{bmatrix} \boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}} & \boldsymbol{V}_{\boldsymbol{x}} \\ \boldsymbol{V}_{\boldsymbol{x}} & \boldsymbol{I} \end{bmatrix} \succeq \boldsymbol{0}.$$

Likewise, let $\boldsymbol{V}_{\boldsymbol{v}} \preceq (\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{R}^{\frac{1}{2}})^{\frac{1}{2}}$, i.e., $\boldsymbol{V}_{\boldsymbol{v}}^2 \preceq \boldsymbol{R}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{R}^{\frac{1}{2}}$. By Schur complement, it is equivalent to require

$$\begin{bmatrix} \boldsymbol{R}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{R}^{\frac{1}{2}} & \boldsymbol{V}_{\boldsymbol{v}} \\ \boldsymbol{V}_{\boldsymbol{v}} & \boldsymbol{I} \end{bmatrix} \succeq \boldsymbol{0}.$$

Note that $\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}} \succeq \boldsymbol{0}$ and $\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{R}^{\frac{1}{2}} \succeq \boldsymbol{0}$. □

B.9 Proof of Theorem 7

In order to simplify the objective function, let $\boldsymbol{U} \preceq \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_\mu^{\min}$. By Schur complement, the problem (2.64) subject to (2.65) is equivalent to

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}, \boldsymbol{U}} \operatorname{Tr} \boldsymbol{U},$$

subject to

$$\begin{cases} \begin{bmatrix} (\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{U})/i_\mu^{\min} & \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top \\ \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}} \end{bmatrix} \succeq \boldsymbol{0} \\ \boldsymbol{U} \succeq \boldsymbol{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \preceq \theta_{2,x}\boldsymbol{M} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \theta_{1,x}\boldsymbol{M} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \preceq \theta_{2,v}\boldsymbol{R} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succeq \theta_{1,v}\boldsymbol{R} \succ \boldsymbol{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \boldsymbol{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \boldsymbol{0}. \end{cases}$$

Namely,

$$\begin{bmatrix} \boldsymbol{U}/i_\mu^{\min} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \preceq \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x}}/i_\mu^{\min} & \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top \\ \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}} \end{bmatrix}.$$

Since $\boldsymbol{I} \succ \mathbf{0}$ and $\boldsymbol{I}/i_\mu^{\min} - \boldsymbol{H}(\boldsymbol{H}^\top\boldsymbol{H})^{-1}\boldsymbol{H}^\top \succeq \boldsymbol{I} - \boldsymbol{H}(\boldsymbol{H}^\top\boldsymbol{H})^{-1}\boldsymbol{H}^\top \succeq \mathbf{0}$, by Schur complement, we have

$$\frac{d\,\mathrm{Tr}\begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{x}}/i_\mu^{\min} & \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top \\ \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}}\end{bmatrix}}{d\boldsymbol{\Sigma}_{\boldsymbol{x}}} = \begin{bmatrix}\boldsymbol{I}/i_\mu^{\min} & \boldsymbol{H} \\ \boldsymbol{H}^\top & \boldsymbol{H}^\top\boldsymbol{H}\end{bmatrix} \succeq \mathbf{0},$$

and

$$\frac{d\,\mathrm{Tr}\begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{x}}/i_\mu^{\min} & \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top \\ \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}}\end{bmatrix}}{d\boldsymbol{\Sigma}_{\boldsymbol{v}}} = \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}\end{bmatrix} \succeq \mathbf{0},$$

implying the upper bound of $\begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{x}}/i_\mu^{\min} & \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top \\ \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} & \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}}\end{bmatrix}$ is reached by the upper bounds of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$. Note that $\boldsymbol{H}(\boldsymbol{H}^\top\boldsymbol{H})^{-1}\boldsymbol{H}^\top$ is an idempotent matrix (a.k.a. projection matrix in linear regression) whose eigenvalues only contain zeros and ones (therefore, $\boldsymbol{I} - \boldsymbol{H}(\boldsymbol{H}^\top\boldsymbol{H})^{-1}\boldsymbol{H}^\top \succeq \mathbf{0}$).

As a result,

$$\begin{bmatrix}\boldsymbol{U}/i_\mu^{\min} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix} \preceq \begin{bmatrix}\theta_{2,\boldsymbol{x}}\boldsymbol{M}/i_\mu^{\min} & \theta_{2,\boldsymbol{x}}\boldsymbol{M}\boldsymbol{H}^\top \\ \boldsymbol{H}\theta_{2,\boldsymbol{x}}\boldsymbol{M} & \boldsymbol{H}\theta_{2,\boldsymbol{x}}\boldsymbol{M}\boldsymbol{H}^\top + \theta_{2,\boldsymbol{v}}\boldsymbol{R}\end{bmatrix},$$

giving

$$\begin{bmatrix}(\theta_{2,\boldsymbol{x}}\boldsymbol{M} - \boldsymbol{U})/i_\mu^{\min} & \theta_{2,\boldsymbol{x}}\boldsymbol{M}\boldsymbol{H}^\top \\ \boldsymbol{H}\theta_{2,\boldsymbol{x}}\boldsymbol{M} & \boldsymbol{H}\theta_{2,\boldsymbol{x}}\boldsymbol{M}\boldsymbol{H}^\top + \theta_{2,\boldsymbol{v}}\boldsymbol{R}\end{bmatrix} \succeq \mathbf{0}.$$

Therefore, the upper bound of $\boldsymbol{U}$ is

$$\theta_{2,\boldsymbol{x}}\boldsymbol{M} - \theta_{2,\boldsymbol{x}}^2\boldsymbol{M}\boldsymbol{H}^\top(\boldsymbol{H}\theta_{2,\boldsymbol{x}}\boldsymbol{M}\boldsymbol{H}^\top + \theta_{2,\boldsymbol{v}}\boldsymbol{R})^{-1}\boldsymbol{H}\boldsymbol{M}\cdot i_\mu^{\min},$$

reached by $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \theta_{2,\boldsymbol{x}}\boldsymbol{M}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \theta_{2,\boldsymbol{v}}\boldsymbol{R}$. □

B.10 Proof of Theorem 8

By noting that $\boldsymbol{\psi}(\cdot) := -\frac{d}{d\boldsymbol{\mu}}\ln p(\boldsymbol{\mu})$ and recalling (2.57) and (2.58) in the worst case, (2.67) and (2.68) are immediate. For the worst-case distribution of $\mathbf{v}$, it is not simply $\mathcal{N}_m(\boldsymbol{c}_{\boldsymbol{v}}^*, \boldsymbol{\Sigma}_{\boldsymbol{v}}^*)$ where $\boldsymbol{c}_{\boldsymbol{v}}^* = \mathbf{0}$ because we have $\mathbf{v} = \boldsymbol{S}^{\frac{1}{2}}\mathbf{u} - \boldsymbol{H}(\mathbf{x} - \bar{\boldsymbol{x}})$. From Highlight 1, the distribution of $\mathbf{v}$ suffers from two types of deviations, i.e., outlier-related and outlier-unrelated. $\mathcal{N}_m(\boldsymbol{c}_{\boldsymbol{v}}^*, \boldsymbol{\Sigma}_{\boldsymbol{v}}^*)$ is just the worst-case distribution for the outlier-unrelated part. The integrated worst-case distribution $\mathbb{P}_{\mathbf{v}}^*$ of $\mathbf{v}$ is determined by the convolution of $\mathbb{P}_{\mathbf{u}}^*$ and $\mathbb{P}_{\mathbf{x}}^*$ through $\mathbf{v}^* = \boldsymbol{S}^{*\frac{1}{2}}\mathbf{u}^* - \boldsymbol{H}(\mathbf{x}^* - \bar{\boldsymbol{x}})$. It

is non-trivial to explicitly compute this convolution. However, fortunately, we do not need to pursue its exact expression (or numerical value). The other statements are straightforward from Lemmas 1, 2 and Theorems 3, 4, 5, 6, 7. □

B.11 Proof of Theorem 9

The weak min-max property admits

$$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}}} V(\boldsymbol{\phi},\mathbb{P}) \leq \min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}}} \max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} V(\boldsymbol{\phi},\mathbb{P}).$$

Supposing the estimator $\boldsymbol{\phi}^*$ and the worst case distribution $\mathbb{P}^*$ solve the max-min problem which are available from Theorem 8, we have $V(\boldsymbol{\phi}^*,\mathbb{P}^*) \leq \min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}}} \max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} V(\boldsymbol{\phi},\mathbb{P})$.

In fact, $(\boldsymbol{\phi}^*,\mathbb{P}^*)$ forms a saddle point of $V(\boldsymbol{\phi},\mathbb{P})$ because there is an one-to-one correspondence between $\boldsymbol{\phi}^*$ and $\mathbb{P}^*$:

$$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} V(\boldsymbol{\phi}^*,\mathbb{P}) = V(\boldsymbol{\phi}^*,\mathbb{P}^*) = \min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}}} V(\boldsymbol{\phi},\mathbb{P}^*).$$

Hence,

$$\min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}}} \max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} V(\boldsymbol{\phi},\mathbb{P}) \leq \max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} V(\boldsymbol{\phi}^*,\mathbb{P}) = V(\boldsymbol{\phi}^*,\mathbb{P}^*).$$

As a result,

$$\min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}}} \max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} V(\boldsymbol{\phi},\mathbb{P}) = V(\boldsymbol{\phi}^*,\mathbb{P}^*).$$

This shows the strong min-max property, completing the proof. □

Appendix C

Proofs in Chapter 3

C.1 Proof of Lemma 3

This lemma is a special case of [123, Theorem 1.3]. With the facts in [123, Remark 1.12], the statements in this lemma can be obtained. However, the proof of [123, Theorem 1.3] is rather complicated because it dealt with a more general problem and conducted many advanced analyses; it is not motivational for the contexts of this thesis. Below gives a new and concise proof because it is necessary for insights in Fig. 3.1.

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

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

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

another. Since $I(\boldsymbol{x}^i|\boldsymbol{x})$ are conditional distributions, implicit constraints are

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

Second, we write the Lagrange dual problem

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

The sup-inf objective function also writes

$$\sup_{\lambda_i} \inf_{I(\boldsymbol{x}^i|\boldsymbol{x})} \quad \int \sum_{i=1}^{N} (\|\boldsymbol{x}-\boldsymbol{x}^i\| - \lambda_i) I(\boldsymbol{x}^i|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \sum_{i=1}^{N} \lambda_i q_i.$$

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

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

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

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

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

C.2 Proof of Theorem 15

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

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

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

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

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

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

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

almost everywhere. This gives the form of $p(\boldsymbol{x})$ in (3.15). Substituting $p(\boldsymbol{x})$ back into the objective of the Lagrange dual problem gives (3.16). The strong duality holds because (3.14) is concave and $\mathbb{Q}_{\mathbf{x}}$ is a relative interior point at which the inequality constraint in (3.14) is strictly satisfied (due to $\theta > 0$) and the equality constraint in (3.14) simultaneously holds (i.e., the Slater's conditions are met).

When $\theta = 0$, the gradient in (3.18) vanishes if and only if $\mathbb{P}_{\mathbf{x}} = \mathbb{Q}_{\mathbf{x}}$. Therefore, (3.15) and (3.16) also work for $\theta = 0$. In summary, this theorem works for all $\theta \geq 0$. □

C.3 Proof of Lemma 4

This lemma is a special case of [123, Theorem 1.3]. One can also prove it using the standard Lagrange dual theory (cf. Appendix C.1). We do not give details due to necessity. □

C.4 Proof of Theorem 16

The proof is straightforward by writing the Lagrange dual problem and differentiating with respect to P_{ij}. The strong duality holds: (3.24) is concave and $\{P^0_{ij}\}_{\forall i,\forall j}$ is assumed to be a relative interior point satisfying the Slater's conditions. In the special case when $M = N$, and $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{Q}_{\mathbf{x}}$ have the same support, P^0_{ij} can be constructed as follow:

$$P^0_{ij} = \begin{cases} q_i, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases}$$

which is resulted from letting $\mathbb{P}_{\mathbf{x}} := \mathbb{Q}_{\mathbf{x}}$. In a general case when $M \neq N$ or they have different supports, to guarantee the existence of P^0_{ij}, we must let θ be strictly larger than $\min_{P_{ij}} \sum_{i=1}^{N}\sum_{j=1}^{M} \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \cdot P_{ij}$ over all P_{ij} such that $\sum_{j=1}^{M} P_{ij} = q_i,\ \forall i \in [N]$. Unlike Theorem 15, we additionally require the existence of P^0_{ij}, because the reference distribution $\mathbb{Q}_{\mathbf{x}}$ in this case is no longer guaranteed to be a relative interior point that satisfies the Slater's conditions. □

C.5 Proof of Theorem 18

If $\theta = 0$, the maximum entropy distribution solving (3.33) is $\boldsymbol{q}$ itself. Below discusses the case when $\theta > 0$. The Lagrange dual problem of (3.33) is

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

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

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

This gives (3.34). Substituting (3.34) back into the objective of the Lagrange dual problem, we have (3.35). Since (3.33) is concave, and $\boldsymbol{q}$ is a relative interior point in the feasible region of (3.33) such that the inequality is strictly satisfied (due to $\theta > 0$) and the equality is met, the strong duality holds due to the Slater's condition. Namely, if λ_0 and λ_1 solve (3.35), p_i in (3.34) solves (3.33). When $\theta = 0$, the gradient (3.36) vanishes if and only if $\boldsymbol{p} = \boldsymbol{q}$; i.e., (3.34) and (3.35) also work for the case when $\theta = 0$. In summary, this theorem works for all $\theta \geq 0$. □