IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Regularization for distributionally robust state estimation and prediction

Jean-Sébastien Brouillon, Florian Dörfler, and Giancarlo Ferrari-Trecate

Abstract— **The increasing availability of sensing techniques provides a great opportunity for engineers to design state estimation methods, which are optimal for the system under observation and the observed noise patterns. However, these patterns often do not fulfill the assumptions of existing approaches. We provide a direct method using samples of the noise to create a moving horizon observer for linear time-varying and nonlinear systems, which is optimal under the empirical noise distribution. Moreover, we show how to enhance the observer with distributional robustness properties in order to handle unmodeled components in the noise profile, as well as different noise realizations. We prove that, even though the design of distributionally robust estimators is a complex minmax problem over an infinite-dimensional space, it can be transformed into a regularized linear program using a system level synthesis approach. Numerical experiments with the Van der Pol oscillator show the benefits of not only using empirical samples of the noise to design the state estimator, but also of adding distributional robustness. We show that our method can significantly outperform state-of-the-art approaches under challenging noise distributions, including multi-modal and deterministic components.**

I. Introduction

Estimating and predicting the states of a system is a fundamental problem in many areas of science and engineering, ranging from control theory to signal processing and machine learning. The goal is to use a set of noisy and possibly incomplete observations of the system's output to infer the true internal state of the system with minimal error. The problem of state smoothing, filtering, and prediction (hereafter referred to as state estimation problem, for short) is challenging due to several factors, such as the presence of measurement noise, unmodeled dynamics, nonlinearities, and uncertainty. The recent advances in sensing and communications technologies and computation have allowed engineers to gather large amounts of data about the noise affecting systems of various nature.

The design of a high-performance state estimator for a given system follows three steps: (i) the accurate modeling of the system dynamics and the statistics of the process and measurement noises, (ii) the choice of an estimator that best fits the model and noise assumptions, and (iii) the optimization of the estimator parameters. This process can be difficult, especially if the noises follow an uncommon profile (e.g., including outliers or deterministic signals), or if the system is time-varying. In the latter case, the design process must be repeated online.

The most popular estimation method is the Kalman Filter (KF), which has a closed form solution that can be computed online. This is the backbone of the Extended Kalman Filter (EKF), which recomputes the filter parameters at each time step based on the linearization of a system at the current operating point [1]. KFs may not perform well when the variance is not accurately measured, even if the noise is Gaussian. To address this issue, [2] proposes an automatic method for learning KF parameters. Another popular estimation method is to stabilize the error dynamics and reject errors in the initial state estimate using a Luenberger Observer (LO). While the KF provides optimality guarantees for linear systems under Gaussian noise, the LO can be a better candidate for other noise distributions, even though its optimal design is challenging in real time.

When dealing with non-Gaussian disturbances, particle filters are a popular approach, but they are computationally expensive and do not exploit specific patterns in non-stochastic noise profiles. Other methods involve learning the non-stochastic part of the noise and assuming standard Gaussian or worst-case distributions for the stochastic component [3], [4]. However, these approaches still make strong assumptions about the noise, which can lead to poor performance if they are not verified. A more flexible method is Moving Horizon Estimation (MHE). It can model not only non-stochastic profiles by penalizing combinations of errors at different time steps, but also non-Gaussian noise distributions using non-quadratic cost functions[1] [6]. However, MHE requires significant computing power and can be sensitive to modeling errors in both the noise statistics and the system itself [7].

Distributionally Robust Optimization (DRO) is a powerful mathematical tool to mitigate errors in the statistical modeling of the noise, by considering the worst probability distribution within an uncertainty set around the empirical one [8]. Recent advances in this field have significantly simplified the computation of robust optimizers, by showing the equivalence between distributional robustness and regularization [9], [10]. DRO has recently been applied to Model Predictive Control (MPC) and Data-enabled Predictive Control (DeePC) to provide a direct method from noise samples to controller design [11], [12], [13]. This approach has only been applied to the field of state estimation under the assumption that the worst case distribution is Gaussian [14], [15], [16].

In this paper, we attempt to fill the gap and introduce a robust unconstrained MHE method that uses DRO to incorporate samples of the noise profile directly in the estimation process, hence eliminating the need for statistical modeling. To do so, we prove that the regularization-based relaxation proposed in [10] can be exact for ℓ_1 norm-based loss functions, which are relevant for MHE. This extends

[1] Although we focus on the unconstrained case in this paper, MHE can also implement constraints [5].

the results obtained for vector-valued parameters in [8], [9]. We show that our approach is capable of providing both predictions and filtered state estimates for discrete-time linear time-varying systems. Moreover, the final estimation problem is a combination of several small Linear Programs (LPs), which can be efficiently solved in real time. Finally, we provide a simulation example illustrating the performance of this new method for the observation of a linearized Van der Pol oscillator under challenging noise profiles.

A. Preliminaries and Notation

Time indices are denoted by the subscript t, and boldface letters denote the stacked vectors at all times in a window. Similarly, calligraphic letters denote linear operators applying to such stacked vectors. Underlined bold symbols are trajectory matrices, whose columns are bold symbol vectors. For example, for a state $x_t \in \mathbb{R}^n$, the trajectory over the window $[t-T, t]$ is $\boldsymbol{x} = [x_{t-T}^\top, \dots, x_t^\top]^\top \in \mathbb{R}^{n(T+1)}$, which can be affected by the operator $\mathcal{C}$ such that $\boldsymbol{y} = \mathcal{C}\boldsymbol{x}$. If N of these trajectories are available, they can be included in the matrix $\underline{\boldsymbol{x}} = [\boldsymbol{x}_0, \dots, \boldsymbol{x}_N] \in \mathbb{R}^{n(T+1)\times N}$.

The subscript i is used to denote the i^{th} row of a matrix. The matrix I denotes the identity and I_i is the i^{th} unit vector. The function blkdiag$([X_0, \dots, X_N])$ constructs a block-diagonal matrix from the blocks $X_0, \dots, X_N$.

The ℓ_2-norm of a vector is denoted by $\|\cdot\|_2$, which also denotes the spectral norm of a matrix (its largest singular value). The ℓ_1 norm of a vector is denoted by $\|\cdot\|_1$, and $\|\cdot\|_{F_1}$ is the ℓ_1 Frobenius norm of a matrix, i.e. the sum of the ℓ_1 norms of its rows.

II. System model

A. LTV dynamics and observer

We model a dynamical system using a discrete-time state-space representation, where the state $x_t \in \mathbb{R}^n$ is hidden, and only the output $y_t \in \mathbb{R}^p$ is observed. The state dynamics and output map are fully described by the equations

$$x_{t+1} = A_t x_t + w_t, \tag{1a}$$

$$y_t = C_t x_t + v_t, \tag{1b}$$

where w_t and v_t are generic process and measurement noises.

Assumption 1. *The system* (1) *is observable for all t.*

This assumption in very common and often necessary to estimate the states of a system [17].

We aim to compute the estimates $\hat{x}_\tau$ of the states in the window $[t-T_s, \dots, t+T_f]$ around the current time t. To do so, we use the following state estimator

$$\hat{x}_{\tau+1} = A_\tau \hat{x}_\tau - \sum_{k=t-T_s}^{t} L_{\tau,k}(C_k \hat{x}_k - y_k), \tag{2}$$

which uses the observations y_k for $k = t-T_s, \dots, t$, and design the gains $L_{\tau,k}$ for $\tau = t-T_s, \dots, t+T_f-1$. The observer gains must stabilize the dynamics of the error $e_\tau = \hat{x}_\tau - x_\tau$, given by

$$e_{\tau+1} = A_\tau e_\tau - w_\tau - \sum_{k=t-T_s}^{t} L_{\tau,k}(C_k e_k - v_k). \tag{3}$$

There are two main differences between (2) and the classical MHE problem [18]: (i) the presence of a forecasting horizon $[t+1, t+T_f]$ after the standard smoothing horizon $[t-T_s, t]$ and (ii) the optimization variables are matrix gains $L_{\tau,k}$, rather than the point estimates $\hat{x}_\tau$. This policy-based problem, similar to dynamic programming for control [6, Chapter 3.3], improves the estimate's robustness, while giving the same results as classic MHE in nominal conditions.

Remark 1. *Known system inputs are not included in the observer design since they cancel out when computing the error $e_\tau = \hat{x}_\tau - x_\tau$. If present, they can be added to* (2) *when computing the state estimate.*

Remark 2. *There are no assumptions on both v_t and w_t, which can also include modelling errors. For example, if* (1) *representes the linearization of the system $x_{t+1} = f(x_t, t) + \tilde{w}_t$, $y_t = h(x_t, t) + \tilde{v}_t$ around a state trajectory, the variables v_t and w_t can embed worst-case linearization errors.*

In the sequel, we consider the estimation problem for a single horizon with a fixed t. Hence, for simplicity t is omitted in the notation.

B. Error dynamics over the entire horizon

To design the LTV observer policy based on the gains $L_{\tau,k}$ for $\tau \in [t-T_s, t+T_f], k \in [t-T_s, t]$, we stack the dynamics of the state estimation error (3) as

$$\boldsymbol{e} = \mathcal{Z}\mathcal{A}\boldsymbol{e} - \mathcal{L}\mathcal{C}\mathcal{Z}\boldsymbol{e} + \mathcal{L}\boldsymbol{v} + \boldsymbol{w}, \tag{4}$$

$$\text{where } \mathcal{Z} = \begin{bmatrix} 0_{n\times n} & & & \\ I & \ddots & & \\ & \ddots & \ddots & \\ & & I & 0_{n\times n} \end{bmatrix}, \tag{5a}$$

$$\mathcal{A} = \begin{bmatrix} A_{t-T_s} & & & \\ & \ddots & & \\ & & A_{t+T_f-1} & \\ & & & 0_{n\times n} \end{bmatrix}, \mathcal{C} = \begin{bmatrix} 0_{n\times p} & & & \\ & C_{t-T_s} & & \\ & & \ddots & \\ & & & C_{t+T_f-1} \end{bmatrix}, \tag{5b}$$

$$\boldsymbol{v} = \begin{bmatrix} 0_{p\times 1} \\ v_{t-T_s} \\ \vdots \\ v_{t+T_f-1} \end{bmatrix}, \boldsymbol{w} = \begin{bmatrix} e_{t-T_s} \\ -w_{t-T_s} \\ \vdots \\ -w_{t+T_f-1} \end{bmatrix}, \boldsymbol{e} = \begin{bmatrix} e_{t-T_s} \\ \vdots \\ e_{t+T_f-1} \\ e_{t+T_f} \end{bmatrix}, \tag{5c}$$

and $\mathcal{L}$ is the observer policy to be designed, written as

$$\mathcal{L} = \begin{bmatrix} 0_{n\times p} & 0_{n\times p} & \dots & 0_{n\times p} & 0_{n\times p(T_f-1)} \\ 0_{n\times p} & L_{t-T_s,t-T_s} & \dots & L_{t-T_s,t} & 0_{n\times p(T_f-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_{n\times p} & L_{t+T_f-1,t-T_s} & \dots & L_{t+T_f-1,t} & 0_{n\times p(T_f-1)} \end{bmatrix}. \tag{5d}$$

The last block-columns in $\mathcal{L}$ are zero to ensure causality, meaning that the last $p(T_f-1)$ measurements in the window $[t-T_s, t+T_f]$, which are in the future, can not be used. The zero first block-column and -row allow one to ensure the equivalence between (2) and (4), as the first n equations of (4) only ensure that the initial error e_{t-T_s} is correctly propagated over time.

Note that $\mathcal{A}$ and $\mathcal{C}$ are block-diagonal matrices and that $\boldsymbol{v}$ and $\boldsymbol{w}$ include both the noise and the modelling errors. Moreover, the error on the initial state is embedded in the first block of the disturbance vector, i.e. $[\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n]^\top = x_{t-T_s} - \hat{x}_{t-T_s}$. In the sequel, we note the dimensions of $\boldsymbol{v}$ and $\boldsymbol{w}$ as $\boldsymbol{p} = p(T_s + T_f + 1)$ and $\boldsymbol{n} = n(T_s + T_f + 1)$, respectively.

III. Problem statement

We aim to design an optimal data-driven observer $\mathcal{L}$ from N noise samples $\tilde{\boldsymbol{v}}_i$ and $\tilde{\boldsymbol{w}}_i$ for $i = 1, \ldots, N$ collected offline, e.g. during tests prior to the deployment of the observer in the field. A naive approach is to maximize the likelihood based on the empirical distributions $\tilde{\mathbb{P}}_v(\boldsymbol{v}) = \frac{1}{N}\sum_{i=1}^N \delta(\boldsymbol{v} - \tilde{\boldsymbol{v}}_i)$ and $\tilde{\mathbb{P}}_w(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^N \delta(\boldsymbol{w} - \tilde{\boldsymbol{w}}_i)$, where $\delta(\cdot)$ is the Dirac distribution. This nominal method gives minimal errors for realizations of the noise that were in the training set, but can lead to a brittle estimator with poor out-of-sample performance. We introduce distributional robustness with respect to the worst case empirical risk in order to mitigate the impact of unforeseen noise realizations.

The worst-case empirical risk is given by the expected cost given by the worst possible probability distribution. For probability distributions in the sets $\mathbb{V}$ and $\mathbb{W}$ (i.e., $v \sim \mathbb{P}_v \in \mathbb{V}$ and $w \sim \mathbb{P}_w \in \mathbb{W}$), the worst-case empirical risk is defined by

$$\mathcal{R}(\boldsymbol{e}(\mathcal{L}, \boldsymbol{v}, \boldsymbol{w})) := \sup_{\substack{\mathbb{P}_v \in \mathbb{V} \\ \mathbb{P}_w \in \mathbb{W}}} \mathbb{E}_{\substack{v \sim \mathbb{P}_v \\ w \sim \mathbb{P}_w}} \operatorname{cost}(\boldsymbol{e}(\mathcal{L}, \boldsymbol{v}, \boldsymbol{w})). \tag{6}$$

Assumption 2. *The estimation cost is* $\operatorname{cost}(\boldsymbol{e}) = \|\mathcal{Q}\boldsymbol{e}\|_1$, *where* $\mathcal{Q} \in \mathbb{R}^{\boldsymbol{n} \times \boldsymbol{n}}$.

Although quadratic costs are more common in engineering applications, ℓ_1 costs are often used for their robustness to non-Gaussian noise [19].

The sets $\mathbb{V}$ and $\mathbb{W}$ are infinite-dimensional. In order to define them in a meaningful way, we introduce the following definition and assumption.

Definition 1. *The Wasserstein metric W_1 based on the ℓ_∞ norm is defined as*

$$W_1(\mathbb{P}_1, \mathbb{P}_2) = \inf_{\Pi} \int_{\Xi^2} \|\xi_1 - \xi_2\|_\infty \Pi(d\xi_1, d\xi_2),$$

where Ξ is the support of $\mathbb{P}_1$ and $\mathbb{P}_2$ and Π is a joint distribution of ξ_1 and ξ_2 with marginal distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively.

Assumption 3. *The sets $\mathbb{V}$ and $\mathbb{W}$ are Wasserstein-1 balls $\mathbb{B}_{\varepsilon_v}(\tilde{\mathbb{P}}_v)$ and $\mathbb{B}_{\varepsilon_w}(\tilde{\mathbb{P}}_w)$ given by*

$$\mathbb{V} = \mathbb{B}_{\varepsilon_v}(\tilde{\mathbb{P}}_v) = \{\mathbb{P}_v | W_1(\mathbb{P}_v, \tilde{\mathbb{P}}_v) \leq \varepsilon_v\},$$
$$\mathbb{W} = \mathbb{B}_{\varepsilon_w}(\tilde{\mathbb{P}}_w) = \{\mathbb{P}_w | W_1(\mathbb{P}_w, \tilde{\mathbb{P}}_w) \leq \varepsilon_w\}.$$

The support Ξ of the Wasserstein metric W_1 is the entire space $\mathbb{R}^{\boldsymbol{p}}$ or $\mathbb{R}^{\boldsymbol{n}}$ for $\mathbb{P}_v$ and $\mathbb{P}_w$, respectively.

In Definition 1, Wasserstein-1 balls[2] only require a norm, a center, and a radius to define a set in the infinite-dimensional space of probability distributions. We chose the ℓ_∞ norm because it treats each entry of the noise vectors separately[3], making it easier for a user to determine the radii ε_v and ε_w. These radii are given by the expected amount of noise in the worst sensor and the expected disturbance in the most perturbed state. The center is a distribution, which is a function with an unbounded support, and thus much harder to determine. We do away with this difficulty by centering the balls on empirical distributions.

Remark 3. *Robustness against worst-case bounded noise is more common and avoids infinite-dimensional problems. However, this often means one must choose between over-conservatism or lack of robustness if rare noise realizations are very large. Distributional robustness allows one to consider unbounded disturbances, while weighting them in accordance with their probability. This approach is therefore better suited to generic disturbance patterns.*

Under the assumptions 3 and 2, the worst-case empirical risk (6) becomes

$$\mathcal{R}(\boldsymbol{e}(\mathcal{L}, \boldsymbol{v}, \boldsymbol{w})) = \sup_{\substack{\mathbb{P}_v \in \mathbb{B}_{\varepsilon_v}(\tilde{\mathbb{P}}_v) \\ \mathbb{P}_w \in \mathbb{B}_{\varepsilon_w}(\tilde{\mathbb{P}}_w)}} \mathbb{E}_{\substack{v \sim \mathbb{P}_v \\ w \sim \mathbb{P}_w}} \|\mathcal{Q}\boldsymbol{e}(\mathcal{L}, \boldsymbol{v}, \boldsymbol{w})\|_1, \tag{8a}$$

and an optimal policy can be computed as

$$\mathcal{L}^\star = \underset{\mathcal{L} \text{ causal}}{\arg\inf} \inf_{\boldsymbol{e}} \mathcal{R}(\boldsymbol{e}(\mathcal{L}, \boldsymbol{v}, \boldsymbol{w})) \qquad \text{s.t. (4)}, \tag{8b}$$

which is coupled to (8a) through the constraint (4), and where the constraint "$\mathcal{L}$ causal" enforces the sparsity pattern given by the zero blocks in (5d).

IV. Tractable reformulation

At first glance, the problem (8) seems very challenging to solve. It is a non-convex, infinite-dimensional, and inf-sup problem. In this section, we first address the non-convexity, and then provide a closed-form solution for the risk $\mathcal{R}(\boldsymbol{e})$. In the end, the problem (8) is reduced to a simple LP.

A. Convexification

The first challenge is addressed by proposing a convex reformulation using a System Level Synthesis (SLS) representation of the estimation problem [20], which decouples (8a) and (8b). To do so, we use the disturbance-to-error and noise-to-error maps $\Phi_w \in \mathbb{R}^{\boldsymbol{n} \times \boldsymbol{n}}$ and $\Phi_v \in \mathbb{R}^{\boldsymbol{n} \times \boldsymbol{p}}$ defined in [21], i.e.,

$$\Phi_w = (I - \mathcal{Z}(\mathcal{A} - \mathcal{L}\mathcal{C}))^{-1},$$
$$\Phi_v = (I - \mathcal{Z}(\mathcal{A} - \mathcal{L}\mathcal{C}))^{-1}\mathcal{L}.$$

The estimation error is given by $\boldsymbol{e} = \Phi_v \boldsymbol{v} + \Phi_w \boldsymbol{w}$ and the risk (8a) can be rewritten as a function of Φ_v and Φ_w as

[2] The first order Wasserstein metric is the most common in the literature because it is one of the easiest to interpret and reformulate in a tractable way [8], [9].

[3] Other norms could be less conservative but the problem must be relaxed as in [10, Theorem 2.1] to become tractable.

$$\mathcal{R}(\Phi_v, \Phi_w)= \sup_{\substack{\mathbb{P}_v \in \mathbb{B}_{\varepsilon_v}(\tilde{\mathbb{P}}_v) \\ \mathbb{P}_w \in \mathbb{B}_{\varepsilon_w}(\tilde{\mathbb{P}}_w)}} \mathbb{E}_{\substack{v\sim\mathbb{P}_v \\ w\sim\mathbb{P}_w}} \|\mathcal{Q}(\Phi_{v,i}\boldsymbol{v} + \Phi_{w,i}\boldsymbol{w})\|_1. \quad (10)$$

Using (10), the problem (8) becomes

$$\underset{\substack{\Phi_v \text{ causal} \\ \Phi_w \text{ causal}}}{\arg\inf}\, \mathcal{R}(\Phi_v, \Phi_w) \quad (11)$$

$$\text{s.t. } [\Phi_v, \Phi_w] \begin{bmatrix} \mathcal{C}\mathcal{Z} \\ I - \mathcal{Z}\mathcal{A} \end{bmatrix} = I. \quad (12)$$

According to [20], (11) is convex in Φ_v and Φ_w if and only if $\mathcal{R}$ is convex. Moreover, because (12) must be satisfied, there exist a $\mathcal{L} = \Phi_w^{-1}\Phi_v$ solving (8)[4]. The causality constraints here need to result in a causal policy $\mathcal{L}$ (see Section II-B). On the one hand, the noise-to-error map Φ_v must have the same zero columns as $\mathcal{L}$ given in (5d), because this sparsity will be conserved in the multiplication with Φ^{-1}. On the other hand, Φ_w has the following structure.

$$\Phi_w = \begin{bmatrix} 0_{n\times n} & 0_{n\times n(T_s+1)} & 0_{n\times n(T_f-1)} \\ 0_{n(T_s+1)\times n} & \Phi_{w,11} & 0_{n(T_s+1)\times n(T_f-1)} \\ 0_{n(T_f-1)\times n} & \Phi_{w,21} & \Phi_{w,22} \end{bmatrix},$$

where $\Phi_{w,22}$ is lower-triangular to capture that future disturbances affect the prediction error in a causal way. In the sequel, we will only write "Φ_v causal" and "Φ_w causal" to refer to these sparsity patterns.

In the next section, we explain how to handle the challenge that (8) includes an infinite-dimensional inf-sup problem.

B. Risk closed-form solution

One of the most impactful results of DRO is its equivalence with regularization in regression problems [8]. The SLS reformulation allows us to use the DRO theory directly and express the risk in closed form.

Theorem 1. *The worst-case empirical risk* (10) *can be written in closed form as*

$$\mathcal{R}(\Phi_v, \Phi_w) = \left\| \mathcal{Q}[\Phi_v, \Phi_w] \begin{bmatrix} \underline{\tilde{\boldsymbol{v}}} \\ \underline{\tilde{\boldsymbol{w}}} \end{bmatrix} \right\|_{F_1} + \|\mathcal{Q}[\varepsilon_v\Phi_v, \varepsilon_w\Phi_w]\|_{F_1}, \quad (13)$$

where the N empirical measurements are stacked as

$$[\tilde{\boldsymbol{v}}_1, \ldots, \tilde{\boldsymbol{v}}_N] = \underline{\tilde{\boldsymbol{v}}} \in \mathbb{R}^{\boldsymbol{p}\times N},$$
$$[\tilde{\boldsymbol{w}}_1, \ldots, \tilde{\boldsymbol{w}}_N] = \underline{\tilde{\boldsymbol{w}}} \in \mathbb{R}^{\boldsymbol{n}\times N},$$

Proof. Let $\kappa = \frac{\varepsilon_v}{\varepsilon_w}$ be the ratio between the Wasserstein balls radii, and let $\tilde{\mathbb{P}}'_w$ be the rescaled $\tilde{\mathbb{P}}_w$ distribution defined by

$$\tilde{\mathbb{P}}'_w(\kappa w) = \frac{1}{N}\sum_{i=1}^{N} \delta(\kappa(\boldsymbol{w} - \tilde{\boldsymbol{w}}_i)).$$

Hence, with $\boldsymbol{w}' = \kappa\boldsymbol{w}$, we have

$$\mathcal{R}(\Phi_v, \Phi_w) = \sup_{\substack{\mathbb{P}_v \in \mathbb{B}_{\varepsilon_v}(\tilde{\mathbb{P}}_v) \\ \mathbb{P}_w \in \mathbb{B}_{\varepsilon_v}(\tilde{\mathbb{P}}'_w)}} \mathbb{E}_{\substack{v\sim\mathbb{P}_v \\ w'\sim\mathbb{P}_w}} \|\mathcal{Q}(\Phi_v\boldsymbol{v} + \Phi_w\kappa^{-1}\boldsymbol{w}')\|_1.$$

Theorem 10 in [8] states that with a Lipschitz cost $\ell(z) = \|\mathcal{Q}[\Phi_v, \kappa^{-1}\Phi_w]z\|_1$ and an unbounded support Ξ in Definition 1, we have

$$\mathcal{R}(\Phi_v, \Phi_w) = \mathbb{E}_{\substack{v\sim\tilde{\mathbb{P}}_v \\ w'\sim\tilde{\mathbb{P}}'_w}} \|\mathcal{Q}(\Phi_v\boldsymbol{v} + \kappa^{-1}\Phi_w\boldsymbol{w}')\|_1 + \varepsilon_v \sup_{z|\ell^\star(z)<+\infty} \|z\|_\star, \quad (15)$$

where $\ell^\star$ is the convex conjugate function of ℓ and $\|\cdot\|_\star = \|\cdot\|_1$ is the dual to the ℓ_∞ norm used in Definition 1. The function $\ell^\star$ is given by

$$\begin{aligned} \ell^\star(z) &= \sup_{x\in\mathbb{R}^{\boldsymbol{n}+\boldsymbol{p}}} z^\top x - \ell(x), \\ &= \sum_i^{\boldsymbol{n}+\boldsymbol{p}} \sup_{x_i} z_i x_i - |x_i| \left\| \left([\Phi_v, \kappa^{-1}\Phi_w]^\top \mathcal{Q}^\top\right)_i \right\|_1. \end{aligned}$$

Each of the supremums is either zero or infinite, depending on which of the two terms is larger in absolute value. This means that to obtain $\ell^\star(z) < +\infty$, each $|z_i|$ must not be greater than $\left\|\left([\Phi_v, \kappa^{-1}\Phi_w]^\top \mathcal{Q}^\top\right)_i\right\|_1$. Hence, the supremum in (15) is given by $\|\mathcal{Q}[\Phi_v, \kappa^{-1}\Phi_w]\|_{F1}$. To conclude the proof, we substitute this closed-form solution in (15) and compute explicitly the empirical expectation to obtain (13). □

Theorem 1 gives a closed-form solution for the worst-case empirical risk $\mathcal{R}$. This removes the inner supremum in (11). Moreover, the resulting regularized cost is convex, which means that the infimum (11) is equal to a unique, global, and achievable minimum of the risk (13).

Corollary 2. *If $\mathcal{Q}$ is diagonal, then the problem* (11) *can be split into n separate optimization problems. The final solution of the full problem is given by*

$$\Phi_v = \mathcal{Q}^{-1} \begin{bmatrix} 0,\ldots,0, & 0,\ldots,0 \\ (\arg\min_\phi \|\Psi\phi - \mu_2\|_1)^\top, & 0,\ldots,0 \\ \vdots & \vdots \\ (\arg\min_\phi \|\Psi\phi - \mu_{\boldsymbol{n}}\|_1)^\top, & \underbrace{0,\ldots,0}_{\boldsymbol{p}_0 \text{ times}} \end{bmatrix}, \quad (16)$$

$$\Phi_w = (I - \mathcal{Z}\mathcal{A})^{-1} - \Phi_v\mathcal{C}\mathcal{Z}(I - \mathcal{Z}\mathcal{A})^{-1}, \quad (17)$$

where $\boldsymbol{p}_0 = p(T_f - 1)$,
$\mu_i = \left(\left[\mathcal{Q}(I - \mathcal{Z}\mathcal{A})^{-1}\right]_i \left[0\cdot\mathcal{C}^\top, \varepsilon_w I, \underline{\tilde{\boldsymbol{w}}}\right]\right)^\top \quad \forall i = 2, \ldots, \boldsymbol{n},$
and Ψ is the matrix formed by the $\boldsymbol{p} - \boldsymbol{p}_0$ first columns of

$$\Psi_{nc} = \begin{bmatrix} -\varepsilon_v I \\ \varepsilon_w(I - \mathcal{Z}\mathcal{A})^{-\top}\mathcal{Z}^\top\mathcal{C}^\top \\ \underline{\tilde{\boldsymbol{w}}}^\top(I - \mathcal{Z}\mathcal{A})^{-\top}\mathcal{Z}^\top\mathcal{C}^\top - \underline{\tilde{\boldsymbol{v}}}^\top \end{bmatrix}.$$

Proof. First, we note that the constraint (12) is equivalent to (17). Moreover, if Φ_v is causal, then so is Φ_w because both $(I - \mathcal{Z}\mathcal{A})$ and $\mathcal{C}\mathcal{Z}$ are block lower-triangular. Plugging (17) into (11) yields

$$\underset{\Phi_v \text{ causal}}{\arg\min} \left\| \mathcal{Q}[\Phi_v, I_{\mathcal{Z}\mathcal{A}} - \Phi_v\mathcal{C}\mathcal{Z}I_{\mathcal{Z}\mathcal{A}}] \begin{bmatrix} \underline{\tilde{\boldsymbol{v}}} \\ \underline{\tilde{\boldsymbol{w}}} \end{bmatrix} \right\|_{F_1} + \|\mathcal{Q}[\varepsilon_v\Phi_v, \varepsilon_w I_{\mathcal{Z}\mathcal{A}} - \varepsilon_w\Phi_v\mathcal{C}\mathcal{Z}I_{\mathcal{Z}\mathcal{A}}]\|_{F_1},$$

[4] By contradiction, assume that $\Phi_w^{-1} = (I - \mathcal{Z}(\mathcal{A} - \mathcal{L}\mathcal{C}))$ is not invertible. Under Assumption 3, the Wasserstein ball $\mathbb{B}_{\varepsilon_w}$ is supported by $\mathbb{R}^{\boldsymbol{n}}$. Hence, $\mathbb{V}$ and $\mathbb{W}$ always contain realizations of the noises satisfying $\boldsymbol{w} + \mathcal{L}\boldsymbol{v} \notin \text{span}(I - \mathcal{Z}(\mathcal{A} - \mathcal{L}\mathcal{C})) \subset \mathbb{R}^{\boldsymbol{n}}$, which would invalidate the dynamics (4). This contradiction proves that Φ_w must have maximal rank.

where $I_{\mathcal{Z}\mathcal{A}} = (I - \mathcal{Z}\mathcal{A})^{-1}$. Rearranging the terms and inverting the sign inside the norm gives

$$\begin{aligned}\underset{\Phi_v \text{ causal}}{\arg\min} &\, \|\mathcal{Q}\Phi_v(\mathcal{C}\mathcal{Z}I_{\mathcal{Z}\mathcal{A}}\underline{\tilde{\boldsymbol{w}}} - \underline{\tilde{\boldsymbol{v}}}) - \mathcal{Q}I_{\mathcal{Z}\mathcal{A}}\underline{\tilde{\boldsymbol{w}}}\|_{F_1} \\ &+ \|\mathcal{Q}\Phi_v[-\varepsilon_v I, \varepsilon_w \mathcal{C}\mathcal{Z}I_{\mathcal{Z}\mathcal{A}}] - \mathcal{Q}[0\cdot\mathcal{C}^\top, \varepsilon_w I_{\mathcal{Z}\mathcal{A}}]\|_{F_1},\end{aligned}$$

where $0 \cdot \mathcal{C}^\top$ is used to construct a zero matrix of the same shape as Φ_v. By replacing the sum of norms by the norm of an augmented matrix, we obtain

$$\begin{aligned}\underset{\Phi_v \text{ causal}}{\arg\min} &\, \|\mathcal{Q}\,\Phi_v[-\varepsilon_v I, \varepsilon_w \mathcal{C}\mathcal{Z}I_{\mathcal{Z}\mathcal{A}}, \mathcal{C}\mathcal{Z}I_{\mathcal{Z}\mathcal{A}}\underline{\tilde{\boldsymbol{w}}} - \underline{\tilde{\boldsymbol{v}}}] \\ &- \mathcal{Q}I_{\mathcal{Z}\mathcal{A}}[0\cdot\mathcal{C}^\top, \varepsilon_w I, \underline{\tilde{\boldsymbol{w}}}]\|_{F_1}.\end{aligned} \tag{18}$$

Let $\mathcal{Q} = \text{diag}([q_1, \dots, q_{\boldsymbol{n}}])$, for $i = 1, \dots, \boldsymbol{n}$ the i^{th} term of the Frobenius norm in (18) is written as

$$\left\| \begin{bmatrix} -\varepsilon_v I \\ \varepsilon_w I_{\mathcal{Z}\mathcal{A}}^\top \mathcal{Z}^\top \mathcal{C}^\top \\ \underline{\tilde{\boldsymbol{w}}}^\top I_{\mathcal{Z}\mathcal{A}}^\top \mathcal{Z}^\top \mathcal{C}^\top - \underline{\tilde{\boldsymbol{v}}}^\top \end{bmatrix} (q_i \Phi_{v,i})^\top - \begin{bmatrix} 0\cdot\mathcal{C} \\ \varepsilon_w I \\ \underline{\tilde{\boldsymbol{w}}}^\top \end{bmatrix} (\mathcal{Q}I_{\mathcal{Z}\mathcal{A}})_i^\top \right\|_1 . \tag{19}$$

Note that each term only depends on the corresponding row of Φ_v. Hence, the minimization problem can be separated into $\boldsymbol{n}$ independent sub-problems. Finally, one can solve for $\phi = q_i \Phi_{v,i}$ and plugging the constraint that Φ_v is causal simply removes the columns corresponding to the desired zeros in the matrix that pre-multiplies $(q_i \Phi_{v,i})^\top$ in (19). Due to the sparisty pattern of $\mathcal{Z}^\top$ in (19), the first p entries of the optimizer ϕ are zero for all $i = 1, \dots, \boldsymbol{n}$, so we do not need to remove the first p columns of Φ_v. □

Theorem 1 and Corollary 2 are the main results of this paper, as they allows to write (11) as several small LPs. This allows one to first compute Φ_v row by row, and then obtain the observer policy $\mathcal{L}$ using

$$\mathcal{L} = \Phi_w^{-1}\Phi_v = (I - \mathcal{Z}\mathcal{A})(I - \Phi_v\mathcal{C}\mathcal{Z})^{-1}\Phi_v,$$

which is implemented using (2).

V. Numerical results

To highlight the ability of our method to handle time-varying and even nonlinear systems, we will perform experiments on a Van der Pol oscillator under complex disturbance patterns. The dynamics are given by

$$\begin{aligned}\dot{x}(t) &= \left[x_2(t), \left(1 - x_1(t)^2\right)x_2(t) - x_1(t)\right]^\top + w(t), \\ y &= x_1(t) + v(t).\end{aligned}$$

Note that this system is both continuous and non linear, so we cannot use (16) directly. Hence, we discretize the system with a sampling frequency of 10Hz using the forward Euler method, and linearize it at each point of its trajectory (see Remark 2 in Section II-A) resulting in an LTV system. The time horizon considered is 1s (or 10 samples) and we are interested in the one step ahead prediction (i.e., $T_s = 8$ and $T_f = 1$). Additional details can be found in [22].

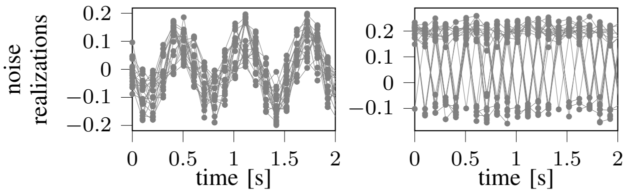
A. Noise profiles

We consider two different cases. First, we apply a noise profile following a sinusoidal pattern plus uniformly distributed noise of the same amplitude as the sine wave (see Fig. 1a). Second, we apply noise following a bi-modal noise distribution based on a mixture of two Gaussians. The precise distributions are:

$$[w_{t,s}^\top, v_{t,s}] \sim \mathcal{U}(sin(10t)[0.1, 0.1, -0.1], 0.1), \tag{20}$$

$$\begin{aligned}[w_{t,b}^\top, v_{t,b}] &\sim 0.25\mathcal{N}(0.05[1,1,-1], \text{diag}(0.025[1,1,2])) \\ &\quad + 0.75\mathcal{N}(0.05[-1,-1,2], \text{diag}(0.025[1,1,2])),\end{aligned} \tag{21}$$

where $\mathcal{U}(\mu, \sigma)$ is the uniform density supported by $[\mu - \sigma, \mu + \sigma]$ and $\mathcal{N}(\mu, \Sigma)$ is the Gaussian density with mean μ and variance Σ. In the sequel, we refer to (20) and (21) as the sine and bimodal noises, respectively. We generate 70 realizations at each time step and (i) use 20 of them as training data (shown in Fig. 1) to build $\underline{\tilde{\boldsymbol{v}}}$ and $\underline{\tilde{\boldsymbol{w}}}$ and (ii) the other 50 to validate the methods.



(a) Sine (b) Bimodal

Fig. 1: First two seconds of the 20 measurement noise realizations used to build the empirical distribution $\tilde{\mathbb{P}}_v$.

The non-stochastic components in (20) and (21) are a sine wave and a bias, respectively. Although the noise is bounded in (20), the linearization error may exceed the bound measured on the training samples. Hence, distributional robustness is well motivated for both types of noises.

B. Results

This section shows the prediction error (i.e. the last n elements of $\boldsymbol{e}$) given by The EKF [1], unconstrained MHE with a quadratic cost [6], and distributionally robust MHE (16) with $\mathcal{Q} = \varepsilon I$ and $\varepsilon_w = \varepsilon_v = \varepsilon = 0.2$ (corresponding to the upper bound or the 95th percentile of the error), denoted in what follows by DRO[5]. The error is $\|\hat{x}_{t+1} - x_{t+1}\|_1$, where $\hat{x}_{t+1}$ is the prediction made by each method and x_{t+1} is the exact state of the oscillator at time $t + 1$.

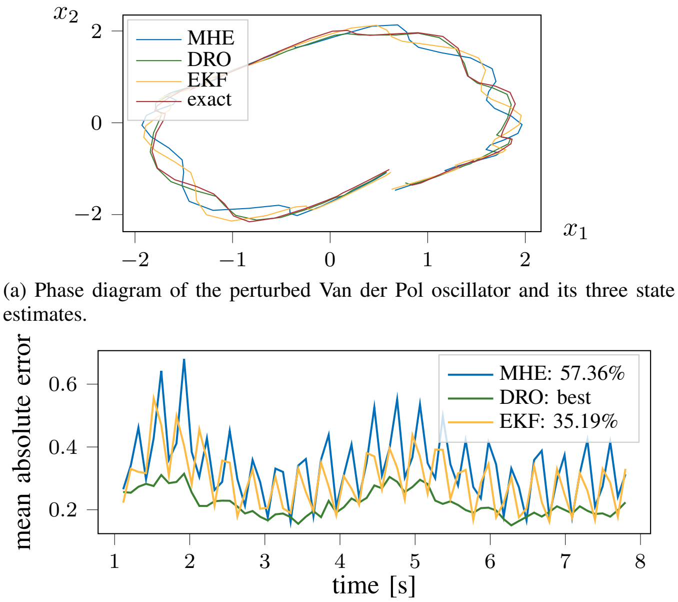
Figures 2 and 3 show a significantly better estimation performance provided by distributionally robust MHE compared to classical MHE and the EKF. In particular, one can see in the error plot that the use of the empirical distribution mitigates the oscillations caused by the sine wave (Fig. 2) and the drift cause by the swings between the modes of the bimodal distribution (Fig. 3). Indeed, one can observe that in both cases, both MHE and EKF generate around 25% to 60% more error than our method.

C. Wasserstein radius

In Section V-B, we tuned the Wasserstein radius to be approximately equal to the sum of the magnitudes of the linearization error and the stochastic component in the noise. Fig. 4 also analyzes the performance of the naive data-driven estimator (i.e., when $\varepsilon = 0$) and shows that in both cases,

[5] Solving MHE and DRO at each timestep takes about 0.04s and 0.1s, respectively, using Python on one core of a RaspberryPi.

while this approach still outperforms methods relying on the assumption of Gaussian noise, the mean and the variance of the error are much larger than with distributional robustness.



(a) Phase diagram of the perturbed Van der Pol oscillator and its three state estimates.

(b) Average error of three estimator at each time step and over all 50 test realizations. The legend shows the total relative error increments.

Fig. 2: Performance analysis of the EKF, MHE and distributionally robust MHE (DRO) under sine noise.

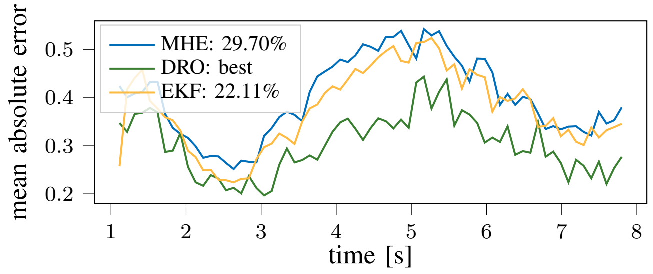
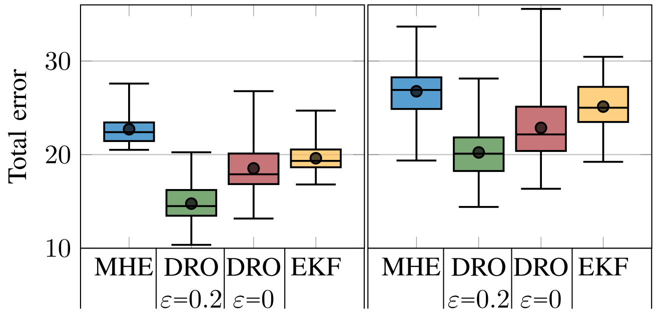


Fig. 3: Average error of the EKF, MHE and distributionally robust MHE (DRO) at each time step and over all 50 test realizations of bimodal noise. The legend shows the total relative error increments.



(a) Sine (b) Bimodal

Fig. 4: Statistics of the error of the EKF, MHE and distributionally robust MHE (DRO), and DRO with zero-radii Wasserstein balls over all 50 test realizations. The mean is shown with a dot marker. The total error is the sum of the errors at each time step.

VI. Conclusions

In this paper, we present a novel MHE method based on the empirical distribution of a system's noise and distributional robustness theory. We prove that our approach can be implemented as computationally-inexpensive LPs.

Future work will focus on studying quadratic costs, as they relate to energy or covariance. Moreover, we will study how to include constraints in our formulation, as it is frequently done in MHE.

References

[1] S. J. Julier and J. K. Uhlmann, "New extension of the kalman filter to nonlinear systems," in *Signal processing, sensor fusion, and target recognition VI*, vol. 3068. Spie, 1997, pp. 182–193.

[2] K. Lee and E. N. Johnson, "State estimation using gaussian process regression for colored noise systems," in *2017 IEEE Aerospace Conference*. IEEE, 2017, pp. 1–8.

[3] J. Huang, D. Shi, and T. Chen, "Event-triggered robust state estimation for systems with unknown exogenous inputs," *Automatica*, vol. 122, p. 109248, 2020.

[4] G. N. Nair, "A nonstochastic information theory for communication and state estimation," *IEEE Transactions on automatic control*, vol. 58, no. 6, pp. 1497–1510, 2013.

[5] A. Alessandri, M. Baglietto, and G. Battistelli, "Moving-horizon state estimation for nonlinear discrete-time systems: New stability results and approximation schemes," *Automatica*, vol. 44, no. 7, pp. 1753–1765, 2008.

[6] J. Rawlings, D. Mayne, and M. Diehl, *Model Predictive Control: Theory, Computation, and Design*. Nob Hill Publishing, 2017.

[7] T. M. Wolff, V. G. Lopez, and M. A. Müller, "Robust data-driven moving horizon estimation for linear discrete-time systems," *arXiv preprint arXiv:2210.09017*, 2022.

[8] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *Operations research & management science in the age of analytics*. Informs, 2019, pp. 130–166.

[9] 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.

[10] R. Chen and I. C. Paschalidis, "Robustified multivariate regression and classification using distributionally robust optimization under the wasserstein metric," *arXiv preprint arXiv:2006.06090*, 2020.

[11] S. Lu, J. H. Lee, and F. You, "Soft-constrained model predictive control based on data-driven distributionally robust optimization," *AIChE Journal*, vol. 66, no. 10, p. e16546, 2020.

[12] L. Aolaritei, M. Fochesato, J. Lygeros, and F. Dörfler, "Wasserstein tube mpc with exact uncertainty propagation," *arXiv preprint arXiv:2304.12093*, 2023.

[13] J. Coulson, J. Lygeros, and F. Dörfler, "Regularized and distributionally robust data-enabled predictive control," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 2696–2701.

[14] S. Wang and Z.-S. Ye, "Distributionally robust state estimation for linear systems subject to uncertainty and outlier," *IEEE Transactions on Signal Processing*, vol. 70, pp. 452–467, 2021.

[15] A. Hakobyan and I. Yang, "Wasserstein distributionally robust control of partially observable linear systems: Tractable approximation and performance guarantee," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 4800–4807.

[16] S. Shafieezadeh Abadeh, V. A. Nguyen, D. Kuhn, and P. M. Mohajerin Esfahani, "Wasserstein distributionally robust kalman filtering," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[17] D. Luenberger, "An introduction to observers," *IEEE Transactions on automatic control*, vol. 16, no. 6, pp. 596–602, 1971.

[18] C. V. Rao, J. B. Rawlings, and J. H. Lee, "Constrained linear state estimation—a moving horizon approach," *Automatica*, vol. 37, no. 10, pp. 1619–1628, 2001.

[19] P. Bloomfield and W. Steiger, "Least absolute deviations curve-fitting," *SIAM Journal on scientific and statistical computing*, vol. 1, no. 2, pp. 290–301, 1980.

[20] J. Anderson, J. C. Doyle, S. H. Low, and N. Matni, "System level synthesis," *Annual Reviews in Control*, vol. 47, pp. 364–393, 2019.

[21] J.-S. Brouillon, G. Ferrari-Trecate, and F. Dörfler, "Minimal regret state estimation of time-varying systems," *arXiv preprint arXiv:2211.14033*, 2022.

[22] J.-S. Brouillon, "Source code of the simulation script for distributionally robust estimation," https://zenodo.org/record/7912963, 2023, DOI:10.5281/zenodo.7912963.