On Regularized System Identification from a Martingale Distributional Robustness Perspective

Xianyu Li *, **Hao Ye** *, **Dexian Huang** *, **Chao Shang** *

* *Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: c-shang@tsinghua.edu.cn).*

Abstract: In this work, we propose a novel martingale-based distributionally robust regression (MDRR) approach to system identification of uncertain dynamical systems. Under data uncertainty, the ridge regression offers a useful remedy, which can be interpreted as a min-max problem through the lens of distributionally robust optimization. However, ignoring the specific structural properties, RR amounts to robustifying against unrealistic perturbations with evident dynamics and thus leads to over-conservatism. By considering the Hankel structure of uncertainty and incorporating martingale constraints into the Wasserstein ambiguity set, the realistic data perturbation pattern can be effectively captured, and this helps to considerably alleviate the conservatism. The induced min-max problem is solved by a subgradient-based algorithm. Empirical results on both simulation and real-world datasets validate the effectiveness of MDRR, showcasing its out-performance over generic regression models and ease of parameter calibration.

Keywords: Distributionally Robust Optimization, Martingale, System Identification, Regularization

1. INTRODUCTION

System identification (SID) seeks to build mathematical models of dynamic systems from observed input and output data (Ljung, 1999). It has been widely adopted as a data-driven strategy across various fields including control design (Huang and Kadali, 2008), fault detection (Ding, 2008), and so on. The success of SID lies in that the dynamic information of a system is carried by its input and output signals. However, the observed data are inevitably subject to uncertainty, primarily due to the noise or disturbance during the data collection process, and this can compromise the efficacy of SID. For example, it was evidenced in Lyandres et al. (2010) that conducting multiple identification experiments may yield rather different results.

As a foundational prototype of SID, the *least squares* (LS) method derives parameter estimates by minimizing the squared errors between measured and predicted outputs. It inherently assumes that the errors as a source of uncertainty are *independent and identically distributed* (i.i.d.) with constant variance, which is the prerequisite for unbiasedness and consistency of LS estimates. Despite its simplicity and asymptotic properties, LS has some flaws, especially when tackling noisy data or highly correlated input variables. In these scenarios, LS may suffer from overfitting, which compromises model performance on unseen data. Regularization techniques offer a useful remedy to alleviate these issues and improve model stability. A prevalent option is the *ridge regression* (RR), which extends LS by adding an ℓ_2-regularizer to the objective function.

Many recent works have established that the regularization amounts to robustification under certain conditions (Bertsimas and Copenhaver, 2018). The *distributionally robust optimization* (DRO) paradigm (Rahimian and Mehrotra, 2022; Boskos et al., 2020), a promising direction in operations research, has provided deep insights into conventional regularization techniques (Li et al., 2022; Chen and Paschalidis, 2018). That is, a regularized problem can be typically interpreted as robustifying against adversarial perturbations within the observed data. This can be formulated into a min-max optimization problem aimed at minimizing the loss function over model parameters while an adversary maximizes the negative impact of distributional shifts by probing worst-case distributions. Li et al. (2022) showcased that the generic RR enjoys distributional robustness in the sense of optimal transport with constraints on conditional expectations.

Unlike ordinary LS based on the i.i.d. assumption, the regressor matrix in SID exhibits a Hankel structure. Thus, unrealistic cases of perturbations are considered by traditional regularization methods, including those having evident dynamics, and this can induce conservatism. To address this, we propose a novel regularized regression method for the SID of uncertain dynamic systems from a new DRO viewpoint. The proposed approach, which is called martingale-based distributionally robust regression (MDRR), considers both the Hankel structure of uncertainty and admissible perturbations without evident dynamics, thereby yielding reduced conservatism of generic

⋆ This work was supported by National Natural Science Foundation of China (Nos. 62373211 and 62327807).

regularizers such as RR. To be specific, adding martingale constraints is helpful for excluding perturbations with evident dynamics and thus better characterizing the true data distribution. In this way, the out-of-sample generalization performance may be improved. Nevertheless, the induced min-max optimization problem is more complex than that of RR and no longer admits a convex reformulation. Thus, a tailored subgradient-based algorithm is developed. Using data collected from a simulated closed-loop system and a real-world glass furnace system, we showcase the superior performance of MDRR in handling complicated noise-corruptions in dynamic data and desirable generalization capability. Besides, as compared to the conventional RR, our MDRR shows better insensitivity to the choice of hyper-parameters.

The remainder of this article unfolds as follows. Section 2 revisits the foundational notions of conventional (regularized) LS identification methods and Wasserstein DRO. Section 3 presents the formulation of MDRR. Two case studies on a simulated closed-loop system and a glass furnace are investigated in Section 4, followed by final conclusions.

Notations: We consider the Euclidean space $\mathbb{R}^n$, with Euclidean norm $\|\cdot\|$. Let $\mathcal{P}(\mathbb{R}^n)$ be the space of (Borel) probability measures over $\mathbb{R}^n$. Given a sequence $\{\boldsymbol{u}(k)\}_{k=1}^N \in \mathbb{R}^d$, $\boldsymbol{u}_{(k:l)}$ denotes the restriction of $\boldsymbol{u}$ to the interval $[k,l]$ as $\mathbf{col}(\boldsymbol{u}(k),\boldsymbol{u}(k+1),\cdots,\boldsymbol{u}(l)) = [\boldsymbol{u}(k)^\top\ \boldsymbol{u}(k+1)^\top\ \cdots\ \boldsymbol{u}(l)^\top]^\top$, and $\boldsymbol{u}_{(l:-1:k)}$ denotes the restriction of $\boldsymbol{u}$ to the interval $[k,l]$ in reverse order as $\mathbf{col}(\boldsymbol{u}(l),\boldsymbol{u}(l-1),\cdots,\boldsymbol{u}(k)) = [\boldsymbol{u}(l)^\top\ \boldsymbol{u}(l-1)^\top\ \cdots\ \boldsymbol{u}(k)^\top]^\top$. For any data matrix or vector, we denote its uncertain version by adding a tilde symbol above it, i.e., $\tilde{\boldsymbol{u}}$, $\tilde{\boldsymbol{y}}$ or $\tilde{\boldsymbol{H}}$.

2. PRELIMINARIES

In this section, we recall some basics of LS identification of auto-regressive with extra inputs (ARX) models, regularized LS, and DRO.

2.1 LS identification of ARX models

We consider a discrete-time *linear time-invariant* (LTI) *single-input single-output* (SISO) dynamic system (Ljung, 1999):

$$A(z^{-1})y(k) = B(z^{-1})u(k) + v(k), \tag{1}$$

where $u(k), y(k)$ are observable input and output signals at time step k, respectively. $v(k)$ denotes the process noise. $A(z^{-1})$ and $B(z^{-1})$ can be described as follows:

$$\begin{cases} A(z^{-1}) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_{n_a} z^{-n_a} \\ B(z^{-1}) = b_1 z^{-1} + b_2 z^{-2} + \cdots + b_{n_b} z^{-n_b}, \end{cases} \tag{2}$$

where n_a and n_b denote the dynamic order of the system. When $v(k)$ is a white noise process, the system described by (1) becomes an *auto-regressive with exogenous inputs* (ARX) model. To identify (2) from data, we first rewrite (1) into an LS form:

$$y(k) = \boldsymbol{h}^\top(k)\boldsymbol{\theta} + v(k), \tag{3}$$

where $\boldsymbol{\theta} = [a_{(1:n_a)}^\top, b_{(1:n_b)}^\top]^\top \in \mathbb{R}^{n_a+n_b}$ represents the parameter vector, $\boldsymbol{h}(k) = [-y_{(k-1:-1:k-n_a)}^\top, u_{(k-1:-1:k-n_b)}^\top]^\top \in \mathbb{R}^{n_a+n_b}$ represents the regression vector including known data at time k. The vector form of (3) with data length L is given by:

$$\boldsymbol{y}_L = \boldsymbol{H}_L\boldsymbol{\theta} + \boldsymbol{v}_L, \tag{4}$$

where $\boldsymbol{y}_L = z_{(1:L)}$, $\boldsymbol{v}_L = v_{(1:L)}$, and

$$\begin{aligned} \boldsymbol{H}_L &= \begin{bmatrix} \boldsymbol{h}^\top(1) \\ \vdots \\ \boldsymbol{h}^\top(L) \end{bmatrix} = \left[\begin{array}{c:c} -y_{(0:-1:1-n_a)}^\top & u_{(0:-1:1-n_b)}^\top \\ \vdots & \vdots \\ -y_{(L-1:-1:L-n_a)}^\top & u_{(L-1:-1:L-n_b)}^\top \end{array}\right] \\ &=: [\boldsymbol{H}_{\boldsymbol{y}_L}, \boldsymbol{H}_{\boldsymbol{u}_L}] \end{aligned} \tag{5}$$

where both $\boldsymbol{H}_{\boldsymbol{y}_L}$ and $\boldsymbol{H}_{\boldsymbol{u}_L}$ have Hankel structures. The quadratic function is typically chosen as the cost function in the ordinary LS identification:

$$J(\boldsymbol{\theta}) = \|\boldsymbol{y}_L - \boldsymbol{H}_L\boldsymbol{\theta}\|_2^2. \tag{6}$$

Minimizing (6) then yields the LS solution:

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = (\boldsymbol{H}_L^\top\boldsymbol{H}_L)^{-1}\boldsymbol{H}_L^\top\boldsymbol{y}_L. \tag{7}$$

2.2 Regularized LS identification

In practice, both input and output data are subject to noise corruption, expressed as:

$$\begin{cases} \tilde{y}(k) = y(k) + \xi_y(k) \\ \tilde{u}(k) = u(k) + \xi_u(k), \end{cases} \tag{8}$$

where $\boldsymbol{\xi}(k) := \mathbf{col}(\xi_y(k), \xi_u(k)) \sim \mathbb{Q}_k$ is the joint observational error vector at time k. In the presence of noise corruption, it gives rise to noisy observations $\tilde{\boldsymbol{y}}_L$ and $\tilde{\boldsymbol{u}}_L$. The joint observational error vector $\boldsymbol{\xi}_L := \boldsymbol{\xi}_{(1:L)}$ is governed by the Cartesian product of the individual error distribution at each time step, described as:

$$\mathbb{Q} := \mathbb{Q}_1 \times \mathbb{Q}_2 \times \cdots \times \mathbb{Q}_L \subset \mathcal{P}(\mathbb{R}^{2L}). \tag{9}$$

Hence, the regression matrix $\boldsymbol{H}_L$ involves uncertainty as well, giving rise to the noise-corrupted regression matrix $\tilde{\boldsymbol{H}}_L$ and thus unsatisfactory LS estimation performance with large variance. As an effective remedy, regularization has been widely adopted to robustify against observational errors. In SID, a prevalent option of regularized LS regression is RR (Hoerl and Kennard, 1970), which adds a regularizer in the form of squared ℓ_2-norm of $\boldsymbol{\theta}$ to shrink the coefficients and prevent overfitting:

$$J(\boldsymbol{\theta}) = \|\tilde{\boldsymbol{y}}_L - \tilde{\boldsymbol{H}}_L\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2. \tag{10}$$

Here, the ridge parameter $\lambda > 0$ controls the trade-off between minimizing the fitting error and penalizing large values of coefficients. Minimizing (10) yields a closed-form solution of RR:

$$\hat{\boldsymbol{\theta}}_{\mathrm{ridge}} = (\tilde{\boldsymbol{H}}_L^\top\tilde{\boldsymbol{H}}_L + \lambda\boldsymbol{I})^{-1}\tilde{\boldsymbol{H}}_L^\top\tilde{\boldsymbol{y}}_L. \tag{11}$$

2.3 Basics of Wasserstein DRO and distributionally robust interpretation of RR

Definition 1. (Wasserstein Distance, Kantorovich and Rubinshtein (1958)) For any $p \in [1, +\infty)$, the type-p Wasserstein distance between two probability distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$ supported on $\mathbb{R}^n$ is defined as:

$$\begin{aligned} &\mathcal{W}_p(\mathbb{Q}_1, \mathbb{Q}_2) \\ &= \left(\inf_{\pi\in\Pi(\mathbb{Q}_1,\mathbb{Q}_2)} \int_{\mathbb{R}^n\times\mathbb{R}^n} \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|^p \pi(\mathrm{d}\boldsymbol{\xi}_1, \mathrm{d}\boldsymbol{\xi}_2)\right)^{\frac{1}{p}}, \end{aligned} \tag{12}$$

where $\Pi(\mathbb{Q}_1, \mathbb{Q}_2)$ represents the set of all joint distributions of $\boldsymbol{\xi}_1 \in \mathbb{R}^n$ and $\boldsymbol{\xi}_2 \in \mathbb{R}^n$ whose marginals are $\mathbb{Q}_1$ and $\mathbb{Q}_2$, respectively.

Letting $p = 2$, the type-2 Wasserstein ambiguity set is defined as (Shafieezadeh Abadeh et al., 2018):

$$\mathcal{D}_\rho(\mathbb{P}_0) := \left\{\mathbb{Q} \in \mathcal{P}(\mathbb{R}^n) : \mathcal{W}_2^2(\mathbb{Q}, \mathbb{P}_0) \le \rho\right\}, \tag{13}$$

where $\mathbb{P}_0$ is a prescribed nominal distribution, and ρ is the Wasserstein radius characterizing the maximal level of adversarial perturbations.

Consider minimizing an objective $\mathbb{E}_{\mathbb{Q}}[l(\boldsymbol{x}, \boldsymbol{\xi})]$, where $\boldsymbol{x}$ is the decision variable, $\boldsymbol{\xi}$ is the uncertainty governed by an unknown distribution $\mathbb{Q}$, and $l(\cdot, \cdot)$ denotes a particular function. The Wasserstein ambiguity set $\mathcal{D}_\rho(\mathbb{P}_0)$ provides a description of the ambiguity in $\mathbb{Q}$ by imposing some perturbations on the nominal distribution $\mathbb{P}_0$. Based on this, a min-max optimization problem is formulated in DRO:

$$\inf_{\boldsymbol{\theta}} \sup_{\mathbb{Q} \in \mathcal{D}_\rho(\mathbb{P}_0)} \mathbb{E}_{\mathbb{Q}}[l(\boldsymbol{\theta}, \boldsymbol{\xi})]. \tag{14}$$

A key merit of DRO is that the resultant solution enjoys *distributional robustness* with the radius ρ suitably chosen. That is, a satisfactory performance is ensured even when there exists some deviations between the true distribution and $\mathbb{P}_0$.

Indeed, the conventional RR mentioned previously admits an alternative min-max reformulation through the lens of DRO. The input and output signals are assumed to be uncertain and governed by an independent and identical distribution, i.e., $\mathbb{Q}_1 = \mathbb{Q}_2 = \cdots = \mathbb{Q}_L =: \mathbb{P}$. The distribution $\mathbb{P}$ resembles the empirical distribution $\hat{\mathbb{P}} \triangleq \frac{1}{L}\sum_{k=1}^{L} \delta_{(y(k), \boldsymbol{h}^\top(k))}$, where $\delta_{(y(k), \boldsymbol{h}^\top(k))}$ denotes the Dirac measure at the kth sample $(y(k), \boldsymbol{h}^\top(k))$. On this basis, the regularized problem (10) is tantamount to the following min-max optimization problem (Li et al., 2022):

$$\begin{aligned} \inf_{\boldsymbol{\theta}} \sup_{\mathbb{P}} \quad & \mathbb{E}_{\mathbb{P}}\left[(\tilde{y} - \tilde{\boldsymbol{h}}^\top \boldsymbol{\theta})^2\right] \\ \text{s.t.} \quad & \mathbb{E}_{\mathbb{P}}[(\tilde{y}, \tilde{\boldsymbol{h}}^\top) | (y, \boldsymbol{h}^\top)] = (y, \boldsymbol{h}^\top), \\ & \mathbb{P} \in \mathcal{D}_\lambda(\hat{\mathbb{P}}). \end{aligned} \tag{15}$$

where the conditional expectation constraint ensures that the perturbed data have the same mean value as the empirical distribution but are likely to exhibit a different variance. However, (15) fails to take into account the Hankel structure as well as the underlying dynamics of uncertainty.

3. MARTINGALE-BASED DISTRIBUTIONALLY ROBUST REGULARIZER

3.1 Problem formulation

Indeed, (8) indicates that in the presence of observational errors in inputs and outputs, the noise term $\boldsymbol{\xi}_{\boldsymbol{H}_L}$ added to $\boldsymbol{H}_L$ has a Hankel structure:

$$\begin{aligned} \tilde{\boldsymbol{H}}_L &= \boldsymbol{H}_L + \boldsymbol{\xi}_{\boldsymbol{H}_L} \\ &= \boldsymbol{H}_L + \begin{bmatrix} -\xi_{y(0:-1:1-n_a)}^\top & \xi_{u(0:-1:1-n_b)}^\top \\ \vdots & \vdots \\ -\xi_{y(L-1:-1:L-n_a)}^\top & \xi_{u(L-1:-1:L-n_b)}^\top \end{bmatrix}. \end{aligned} \tag{16}$$

The Hankel structure results in a lower freedom of uncertainty than that is assumed by RR in (15). Taking the Hankel structure into consideration, we formulate the following min-max optimization problem to robustify the identified systems against noise corruptions in $\boldsymbol{H}_L$:

$$\begin{aligned} &\inf_{\boldsymbol{\theta}} \sup_{\mathbb{Q} \in \mathcal{M}} \mathbb{E}_{\mathbb{Q}}\left[\|\boldsymbol{y}_L - \boldsymbol{H}_L \boldsymbol{\theta}\|_2^2\right] \\ =&\inf_{\boldsymbol{\theta}} \sup_{\mathbb{Q} \in \mathcal{M}} \mathbb{E}_{\mathbb{Q}}\left[\|(\tilde{\boldsymbol{y}}_L - \boldsymbol{\xi}_{\boldsymbol{y}_L}) - (\tilde{\boldsymbol{H}}_L - \boldsymbol{\xi}_{\boldsymbol{H}_L})\boldsymbol{\theta}\|_2^2\right], \end{aligned} \tag{17}$$

where $\mathcal{M}$ is an ambiguity set used to describe the inexact distribution $\mathbb{Q}$ of $\boldsymbol{\xi}_L$. Based on the Hankel structure, we further incorporate martingale constraints into a Wasserstein ambiguity set, ensuring that the imposed perturbations do not carry evident dynamic patterns. In the following, we first provide a detailed definition of the ambiguity set $\mathcal{M}$, which serves as the core of MDRR, and then proceed to address the solution to (17).

3.2 Wasserstein with martingale ambiguity set

To construct the ambiguity set $\mathcal{M}$, we first introduce the definition of martingale difference processes.

Definition 2. (Martingale Difference Process) A sequence of random variables $\{v_k\}_{k\ge 0}$ is said to be a martingale difference process if $\mathbb{E}[v_k | \mathcal{F}_{k-1}] = 0$, where $\mathcal{F}_{k-1}$ is the filtration at time $k-1$. A filtration $\{\mathcal{F}_k\}_{k\ge 0}$ is an increasing sequence of σ-algebras, representing the accumulated information up to time k.

The following result is a consequence of the tower property and the fundamental property of martingale difference processes.

Lemma 3. Given two martingale difference processes $\{v_1(k)\}$ and $\{v_2(k)\}$, it holds that $\mathbb{E}[v_1(i)v_1(j)] = 0$ and $\mathbb{E}[v_1(i)v_2(j)] = 0$, $i \neq j$.

By definition, a martingale difference process accounts for temporal dependencies, but its conditional expectation being zero ensures the non-existence of predictable patterns in the first-order moment; in other words, there is no systematic trend on average. The martingale difference process with a finite second moment can be regarded as a relaxation of the white noise process, which itself is a specific type of martingale difference.

In this work, we adopt the martingale difference assumption to relax the conventional i.i.d. assumption implicitly made by RR. More precisely, in (15), the perturbations imposed across all $\mathbb{Q}_k$'s are identical. Thus, the perturbation added to $\mathbb{Q} = \mathbb{Q}_1 \times \cdots \times \mathbb{Q}_L$ embodies evident dynamics. This is indeed unrealistic, because such dynamics may not be effectively described by the parameterization $\boldsymbol{\theta}$ and thus cause over-conservatism. In contrast, our approach allows $\mathbb{Q}_k$ to be dependent on $\mathbb{Q}_1, \cdots, \mathbb{Q}_{k-1}$. Therefore, in stark contrast to (13), we propose to construct the ambiguity set by considering temporal dependency through additional martingale constraints:

$$\mathcal{M}_\rho(\mathbb{P}_0) := \left\{ \mathbb{Q} \in \mathcal{P}(\mathbb{R}^{2L}) : \begin{array}{c} \mathbb{E}_{\mathbb{Q}}[\boldsymbol{\xi}(k) \mid \mathcal{F}_{k-1}] = 0 \\ \sum_{k=1}^{L} \mathcal{W}_2^2(\mathbb{Q}_k, \mathbb{P}_0) \le \rho \\ \mathbb{Q} = \mathbb{Q}_1 \times \cdots \times \mathbb{Q}_L \end{array} \right\}. \tag{18}$$

Considering the Hankel structure and the martingale constraints altogether imposes more assumptions on the dynamics of perturbations, thereby effectively excluding those exhibiting evident dynamics.

Owing to the presence of martingale constraints $\mathbb{E}_{\mathbb{Q}}[\boldsymbol{\xi}(k) \mid \mathcal{F}_{k-1}] = 0$, the min-max problem (17) based on (18) is challenging to solve. Thus, we further make some simplifying assumptions and present a decomposition strategy to reformulate $\mathcal{M}_\rho(\mathbb{P}_0)$. It is assumed that both $\{\xi_y(k)\}$ is a martingale difference processes resembling a Gaussian white noise process $\{\epsilon(k)\}$, and $\{\xi_u(k)\}$ coincides with another white noise process $\zeta(k)$. Defining $\eta(k) := \xi_y(k) - \epsilon(k)$, we have:

$$\begin{cases} \xi_y(k) := \epsilon(k) + \eta(k) \\ \xi_u(k) := \zeta(k). \end{cases} \tag{19}$$

where $\epsilon(k) \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is independent of $\zeta(k) \sim \mathcal{N}(0, \sigma_\zeta^2)$, and the variances σ_ϵ^2 and σ_ζ^2 serve as tunable hyperparameters in our proposal. Since the sum or difference of two martingale difference processes remains a martingale difference sequence, $\{\eta(k)\}$ is also a martingale difference. It then suffices to characterize the maximal level of perturbations using the variance information of $\{\eta(k)\}$, which leads to the following reformulation of (18):

$$\mathcal{M}'_\rho(\mathbb{P}_0) = \left\{ \begin{array}{l} \pi \in \mathcal{P}(\mathbb{R}^{2L}) \times \mathcal{P}(\mathbb{R}^{2L}) : \\ \quad \mathbb{E}_\pi[(\epsilon(k), \eta(k), \zeta(k)) \mid \mathcal{F}_{k-1}] = 0 \\ \quad \sum_{k=1}^{L} \mathbb{E}_\pi[\|\eta(k)\|_2^2] \le \rho,\ \pi_{(\boldsymbol{\epsilon},\boldsymbol{\zeta})} = \mathbb{P}_0 \end{array} \right\}. \tag{20}$$

3.3 Worst-case distribution

Having defined the ambiguity set, we then formulate the min-max problem (17) of MDRR as:

$$\begin{aligned} &\inf_{\boldsymbol{\theta}} \sup_{\pi \in \mathcal{M}'_\rho(\mathbb{P}_0)} \mathbb{E}_\pi \left[\|(\tilde{\boldsymbol{y}}_L - \boldsymbol{\xi}_{\boldsymbol{y}_L}) - (\tilde{\boldsymbol{H}}_L - \boldsymbol{\xi}_{\boldsymbol{H}_L})\boldsymbol{\theta}\|_2^2 \right] \\ =&\inf_{\boldsymbol{\theta}} \left\{ \|\tilde{\boldsymbol{y}}_L - \tilde{\boldsymbol{H}}_L\boldsymbol{\theta}\|_2^2 + \sup_{\pi \in \mathcal{M}'_\rho(\mathbb{P}_0)} \mathbb{E}_\pi \left[\|\boldsymbol{\xi}_{\boldsymbol{y}_L} - \boldsymbol{\xi}_{\boldsymbol{H}_L}\boldsymbol{\theta}\|_2^2 \right] \right\}. \end{aligned} \tag{21}$$

When expanding the expectation, all cross terms vanish because both $\mathbb{E}_\pi[\boldsymbol{\xi}_{\boldsymbol{y}_L}]$ and $\mathbb{E}_\pi[\boldsymbol{\xi}_{\boldsymbol{H}_L}]$ are zero, due to the martingale property and the tower property of conditional expectations.

We proceed by discussing how to effectively solve (21) for the parameter $\boldsymbol{\theta}$ of MDRR. For a given $\boldsymbol{\theta}$, the related worst-case distribution can be identified by solving the inner maximization problem:

$$\sup_{\pi \in \mathcal{M}'_\rho(\mathbb{P}_0)} h(\pi, \boldsymbol{\theta}) := \sup_{\pi \in \mathcal{M}'_\rho(\mathbb{P}_0)} \mathbb{E}_\pi \left[\|\boldsymbol{\xi}_{\boldsymbol{y}_L} - \boldsymbol{\xi}_{\boldsymbol{H}_L}\boldsymbol{\theta}\|_2^2 \right]. \tag{22}$$

Theorem 4. (Worst-Case Distribution) For $\boldsymbol{\theta} \neq 0$, one of the optimal solutions to (22) has the form:

$$\hat{\eta}_{\boldsymbol{\theta}}(k) = \begin{cases} \left(c_k(\boldsymbol{\theta}) - \hat{\lambda}\right)^{-1} \left(d_k(\boldsymbol{\theta})\zeta(k) - c_k(\boldsymbol{\theta})\epsilon(k)\right), k \in \mathcal{K} \\ 0, \quad \text{otherwise} \end{cases} \tag{23}$$

where $\mathcal{K} = \{1 - \min(n_a, n_b), \cdots, L-1\}$ is a set of indices, and

$$\begin{cases} c_k(\boldsymbol{\theta}) := 1 + \displaystyle\sum_{i=\max(1,1-k)}^{n_a} \theta_i^2 \\ d_k(\boldsymbol{\theta}) := \displaystyle\sum_{i=\max(1,1-k)}^{\min(n_a,n_b)} \theta_i \theta_{n_a+i}. \end{cases} \tag{24}$$

For $\hat{\lambda} > 1 + \sum_{i=1}^{n_a} \theta_i^2$, $\hat{\lambda}$ is the unique solution of the equality

$$\sum_{k=1}^{L} \mathbb{E}_\pi[\|\hat{\eta}_{\boldsymbol{\theta}}(k)\|_2^2] = \rho. \tag{25}$$

Sketch of Proof. Thanks to the similarity in the structural form of the squared ℓ_2-norm of the problem to that in Lotidis et al. (2023), we borrow the idea of the latter to make the proof, which mainly builds upon the duality theory. We begin by expanding (22) and simplify it to isolate the terms involving $\eta(k)$. Next, we derive its dual problem and find a primal-dual optimal pair. Finally, strong duality can be readily verified, which completes the proof.

The worst-case distribution $\hat{\pi}_{\boldsymbol{\theta}}$, parameterized by $\boldsymbol{\theta}$, has its first marginal distribution corresponding to the law of $(\boldsymbol{\epsilon}, \boldsymbol{\zeta})$, i.e., $\mathbb{P}_0$, and its second marginal corresponding to the law of $(\boldsymbol{\epsilon} + \hat{\boldsymbol{\eta}}, \boldsymbol{\zeta})$. Thus, the worst-case distribution can be characterized using the distribution of $\hat{\boldsymbol{\eta}}$ as well. A key observation from Theorem 4 is that $\hat{\boldsymbol{\eta}}$ depends on $\boldsymbol{\epsilon}$ and $\boldsymbol{\zeta}$. At time step k, $\hat{\xi}_y(k)$ can be expressed as:

$$\begin{aligned} \hat{\xi}_y(k) &= \epsilon(k) + \hat{\eta}_{\boldsymbol{\theta}}(k) \\ &= \left(c_k(\boldsymbol{\theta}) - \hat{\lambda}\right)^{-1} \left[d_k(\boldsymbol{\theta})\zeta(k) - \hat{\lambda}\epsilon(k) \right]. \end{aligned} \tag{26}$$

The worst-case distribution is governed by three tuning parameters ρ, σ_ϵ^2, and σ_ζ^2. The parameter ρ influences $\hat{\lambda}$, thereby adjusting the proportion of $\hat{\boldsymbol{\xi}}_{\boldsymbol{y}_L}$ affected by $\boldsymbol{\epsilon}$ and $\boldsymbol{\zeta}$, which are characterized by σ_ϵ^2, and σ_ζ^2, respectively. In some cases, the inputs are known to be certain, so σ_ζ^2 can be set to zero. For σ_ϵ^2, a rough estimation is given by the sample variance of the fitting error of LS, i.e., $\sigma_\epsilon^2 = \|\tilde{\boldsymbol{y}}_L - \tilde{\boldsymbol{H}}_L\hat{\boldsymbol{\theta}}_{\text{LS}}\|_2^2/L$.

3.4 Solving the min-max problem

Due to the martingale constraints in (20), the problem under MDRR admits neither a convex reformulation nor a closed-form solution like (11). Hence, a subgradient descent algorithm (Nesterov, 2013) is developed. The crux to handle this min-max problem lies in computing the subgradient for a given $\boldsymbol{\theta}'$. This turns out to be related to the worst-case distribution $\hat{\pi}_{\boldsymbol{\theta}'}$ given by Theorem 4.

For simplicity, we rewrite the objective function of the outer minimization problem as:

$$f(\boldsymbol{\theta}) := \|\tilde{\boldsymbol{y}}_L - \tilde{\boldsymbol{H}}_L\boldsymbol{\theta}\|_2^2 + \mathbb{E}_{\hat{\pi}_{\boldsymbol{\theta}'}} \left[\|\boldsymbol{\xi}_{\boldsymbol{y}_L} - \boldsymbol{\xi}_{\boldsymbol{H}_L}\boldsymbol{\theta}\|_2^2 \right]. \tag{27}$$

The function $f(\boldsymbol{\theta})$ is a convex, finite-valued, yet non-smooth function (Lotidis et al., 2023). The following proposition provides a subgradient oracle for $f(\cdot)$.

Proposition 5. (Nesterov, 2013) Given $\boldsymbol{\theta}'$, and the corresponding worst-case distribution $\hat{\pi}_{\boldsymbol{\theta}'} \in \mathcal{M}'_\rho(\mathbb{P}_0)$ and defining

$$g := 2(\tilde{\boldsymbol{H}}_L^\top \tilde{\boldsymbol{H}}_L)\boldsymbol{\theta}' - 2\tilde{\boldsymbol{H}}_L^\top \tilde{\boldsymbol{y}}_L + \nabla h(\hat{\pi}_{\boldsymbol{\theta}'}, \boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}, \tag{28}$$

it holds that $g \in \partial f(\boldsymbol{\theta}')$.

The whole procedure for SID with MDRR is summarized in Algorithm 1. For non-smooth problems, the norm of the subgradient, $\|g\|_2$, only carries a limited amount of information. Thus, the subgradient oracle uses the normalized direction $g/\|g\|_2$ instead. In addition, the rate

of convergence of the subgradient method is $\mathcal{O}(t^{-1/2})$, where t denotes the iteration number. This rate depends on the optimal step size selection strategy, with step size proportional to $1/\sqrt{t}$ (Nesterov, 2013).

Algorithm 1 MDRR for SID of ARX Models

Input: The uncertain data $\tilde{\boldsymbol{y}}_L$ and $\tilde{\boldsymbol{H}}_L$, iteration step size $\gamma_t \propto 1/\sqrt{t}$, Wasserstein distance $\rho > 0$, variance σ_ζ^2 of $\zeta(k)$, variance σ_ϵ^2 of $\epsilon(k)$:

1: Obtain the LS estimate $\hat{\boldsymbol{\theta}}_{\text{LS}}$ as per (7) and use $\hat{\boldsymbol{\theta}}_{\text{LS}}$ as the initial values in the iterative process, i.e., $\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_{\text{LS}}$;
2: **while** Not convergent **do**
3: Get $\hat{\eta}_{\boldsymbol{\theta}_t}$, according to Theorem 4;
4: Compute g_t, as stated by Proposition 5;
5: Set $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \gamma_t g_t / \|g_t\|_2$;
6: $t = t + 1$;
7: **end while**
8: **Return** $\hat{\boldsymbol{\theta}}_{\text{MDRR}}$

4. CASE STUDIES

In this section, we validate our MDRR method using data collected from a simulated closed-loop system and a realistic glass furnace system. Its performance is compared against RR in (11) and standard LS in (7) as references.

4.1 A simulated closed-loop system

Consider the following closed-loop system, in which the open-loop system is described by:

$$\begin{aligned} &z(k) - 0.35z(k-1) + 0.65z(k-2) \\ &\qquad = 1.10u(k-1) - 0.70u(k-2) + e_1(k), \end{aligned} \tag{29}$$

and the feedback controller is given by

$$u(k) = 1.00u(k-1) - 0.50z(k) + 0.20z(k-1) + e_2(k), \tag{30}$$

where $\{e_1(k), e_2(k)\}$ are set to be i.i.d. and uniformly distributed over the interval $[-0.1, 0.1]$ and $[-0.05, 0.05]$, respectively. Based on the above settings, an input-output trajectory of length $L = 492$ is generated as the observed data. This dataset is used to fit an ARX model, with the order selected as $(n_a, n_b) = (2, 2)$. To evaluate the prediction performance of the three methods, we generate a test trajectory of length $L_{\text{test}} = 300$. In this simulation case, we assume that there is no error in input variables $u(k)$ by setting $\sigma_\zeta^2 = 0$. Fig. 1 depicts the test errors of different methods using varying parameter values in this simulated closed-loop system.

Due to feedback control, there exist dependencies between inputs, outputs, and past input variables, complicating the identification process. As shown in Fig. 1, by tuning ρ, RR can achieve performance improvement compared to LS. It enforces the coefficients of highly correlated variables toward smaller values, thereby distributing their influence more evenly and reducing interdependence among them. However, overly large ρ values cause RR to over-shrink the estimate, masking meaningful relationships and degrading performance. In contrast, MDRR consistently outperforms RR. For any σ_ϵ^2, the optimal estimate obtained by MDRR achieves a lower global prediction error than RR. The intuitive reason, we believe, is that RR ignores the specific structure of the regression matrix, leading to the

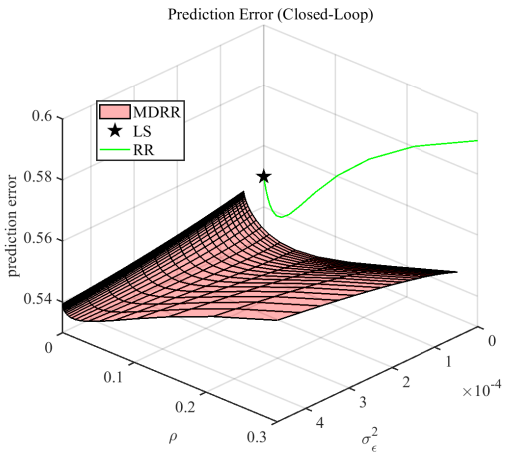


Fig. 1. Profiles of prediction error on test data with increasing ρ and σ_ϵ^2 in closed-loop system

inclusion of unrealistic distributions. By introducing martingale constraints, MDRR considers perturbations without evident dynamics, effectively mitigating conservatism. Another important observation from Fig. 1 is that the prediction error of MDRR varies more smoothly than that of RR, suggesting that our proposal is less sensitive to the choice of tuning, and thus enjoys ease of hyper-parameter calibration.

4.2 Robust identification of a real-world glass furnace

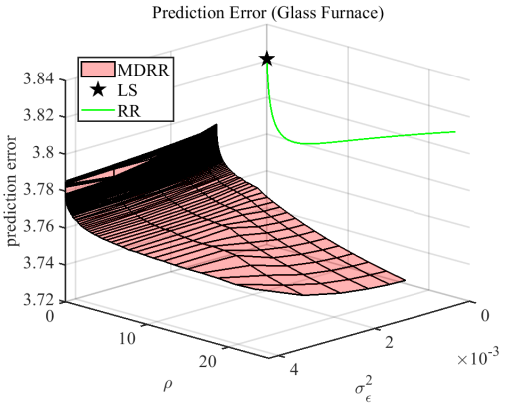
We further assess the performance of MDRR on a real-world dataset collected from a glass furnace system. This system consists of three inputs (i_1, i_2, i_3) —two burners and one ventilator—and six outputs $(o_1, o_2, o_3, o_4, o_5, o_6)$, which are temperature measurements from sensors positioned across a furnace cross-section (De Moor et al., 1997; Van Overschee and De Moor, 1994). In this case, we seek to predict the temperature of one position (o_3). Given that the temperature at different positions in this glass furnace system exhibits correlation, we use not only all three inputs but also the remaining five outputs as input features. The associated choice of model orders is detailed in Table 1.

Table 1. Parameter Setup

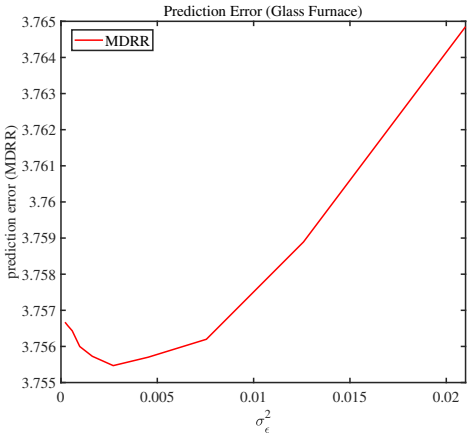
	Input Features	Target Variable
Var.	$[i_1, i_2, i_3, o_1, o_2, o_4, o_5, o_6]$	o_3
Order	[2,2,2,1,1,1,1,1]	2

The train trajectory consists of 1000 data points, while the other 247 are reserved as the test trajectory to evaluate the prediction performance of three estimates. Considering that five temperature sensor readings are input features as well, we set σ_ζ^2 equals to σ_ϵ^2. Fig. 2 illustrates how prediction error varies with increasing ρ and varying σ_ϵ^2 (σ_ζ^2) across different methods in the glass furnace system.

In this real-world case, observational errors are unavoidable. Additionally, the temperature readings from any sensor in a glass furnace can be inferred from the other five due to their relative positions, resulting in collinearity that indicates evident dynamics in uncertainty. Fig. 2 exhibits



(a) $\rho \in [0, 25],\ \sigma_\epsilon^2 \in [1.1, 4.2] \times 10^{-3}$



(b) $\rho = 1000,\ \sigma_\epsilon^2 \in [0.21, 21] \times 10^{-3}$

Fig. 2. Profiles of prediction error on test data with increasing ρ and σ_ϵ^2 in a glass furnace system

the superior performance of MDRR as compared to LS and RR. Using a carefully calibrated ρ, RR outperforms LS. Nevertheless, MDRR attains enhanced performance across a range of ρ and σ_ϵ^2. The experimental results demonstrate that martingale constraints enable MDRR to hedge against only perturbations without evident dynamics. Moreover, it is noteworthy that, as shown in Fig. 2b, even for sufficiently large ρ, (e.g. 1000), MDRR continues to achieve performance improvement. This highlights its insensitivity against the choice of hyper-parameters.

5. CONCLUSION

In this work, we proposed a new regression approach to SID of dynamic systems from a martingale distributional robustness perspective. Our proposed MDRR relaxes i.i.d. assumption and integrates martingale constraints into a Wasserstein ambiguity set, ensuring that the perturbations imposed do not encode evident patterns and alleviating conservatism. Due to the complexity of the induced ambiguity set, we also presented a subgradient-based algorithm addressing the non-smoothness of the DRO problem. Experiments on simulations and a real-world glass furnace system highlighted the performance improvement of our proposed approach over (regularized) LS methods, particularly in the presence of data collinearity. Besides, MDRR shows lower sensitivity against the choice of parameters and thus enjoys ease of manipulating the robustness.

REFERENCES

Bertsimas, D. and Copenhaver, M.S. (2018). Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, 270(3), 931–942.

Boskos, D., Cortés, J., and Martínez, S. (2020). Data-driven ambiguity sets for linear systems under disturbances and noisy observations. In *2020 American Control Conference (ACC)*, 4491–4496.

Chen, R. and Paschalidis, I.C. (2018). A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13), 1–48.

De Moor, B., De Gersem, P., De Schutter, B., Favoreel, W., et al. (1997). Daisy: A database for identification of systems. *Journal A*, 38(4), 5. URL `http://homes.esat.kuleuven.be/~smc/daisy/`.

Ding, S.X. (2008). *Model-Based Fault Diagnosis Techniques: Design schemes, Algorithms, and Tools*. Springer Science & Business Media.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.

Huang, B. and Kadali, R. (2008). *Dynamic Modeling, Predictive Control and Performance Monitoring: A Data-Driven Subspace Approach*. Springer.

Kantorovich, L.V. and Rubinshtein, S. (1958). On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7), 52–59.

Li, J., Lin, S., Blanchet, J., and Nguyen, V.A. (2022). Tikhonov regularization is optimal transport robust under martingale constraints. In *Advances in Neural Information Processing Systems*, volume 35, 17677–17689.

Ljung, L. (1999). *System Identification: Theory for the User*. Prentice Hall, 2nd edition.

Lotidis, K., Bambos, N., Blanchet, J., and Li, J. (2023). Wasserstein distributionally robust linear-quadratic estimation under martingale constraints. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, 8629–8644. PMLR.

Lyandres, O., Van Duyne, R.P., Walsh, J.T., Glucksberg, M.R., and Mehrotra, S. (2010). Prediction range estimation from noisy Raman spectra with robust optimization. *The Analyst*, 135(8), 2111—2118.

Nesterov, Y. (2013). *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media.

Rahimian, H. and Mehrotra, S. (2022). Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3, 1–85.

Shafieezadeh Abadeh, S., Nguyen, V.A., Kuhn, D., and Mohajerin Esfahani, P.M. (2018). Wasserstein distributionally robust Kalman filtering. In *Advances in Neural Information Processing Systems*, volume 31.

Van Overschee, P. and De Moor, B. (1994). N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*, 30(1), 75–93.