Wasserstein Distributionally Robust Linear-Quadratic Estimation under Martingale Constraints

Kyriakos Lotidis Stanford University
Nicholas Bambos Stanford University
Jose Blanchet Stanford University
Jiajin Li Stanford University

Abstract

We focus on robust estimation of the unobserved state of a discrete-time stochastic system with linear dynamics. A standard analysis of this estimation problem assumes a baseline innovation model; with Gaussian innovations we recover the Kalman filter. However, in many settings, there is insufficient or corrupted data to validate the baseline model. To cope with this problem, we minimize the worst-case mean-squared estimation error of adversarial models chosen within a Wasserstein neighborhood around the baseline. We also constrain the adversarial innovations to form a martingale difference sequence. The martingale constraint relaxes the i.i.d. assumptions which are often imposed on the baseline model. Moreover, we show that the martingale constraints guarantee that the adversarial dynamics remain adapted to the natural time-generated information. Therefore, adding the martingale constraint allows to improve upon over-conservative policies that also protect against unrealistic omniscient adversaries. We establish a strong duality result which we use to develop an efficient subgradient method to compute the distributionally robust estimation policy. If the baseline innovations are Gaussian, we show that the worst-case adversary remains Gaussian. Our numerical experiments indicate that the martingale constraint may also aid in adding a layer of robustness in the choice of the adversarial power.

1 INTRODUCTION

We propose and study a Wasserstein distributionally robust optimization (DRO) formulation with martingale constraints for the minimum mean-squared estimation of a linear state-space stochastic system. State estimation is a central problem that arises in many real world applications, involving control systems, such as motion tracking (Lee et al., 2013) and GPS navigation (Liu et al., 2018) in autonomous vehicles, real-time load tracking in the smart power grid (Ghahremani and Kamwa, 2011; Zhao et al., 2017), healthcare monitoring through wearable devices (Zhang et al., 2014) and heart-rate estimation (Prakash and Tucker, 2018), and wireless sensor networks (Ribeiro et al., 2010). However, estimating the hidden state from observation data requires either an accurate nominal model, or policies that are robust to model misspecification. DRO formulations have become increasingly popular in recent years because they provide a principled way to produce sound estimators which account for the impact in model misspecification, which may be caused, for example, by data corruption or non-stationarities (Rahimian and Mehrotra, 2019; Kuhn et al., 2019; Wiesemann et al., 2014; Lin et al., 2022; Delage and Ye, 2010; Lee and Raginsky, 2018; Yang, 2020; Chen and Paschalidis, 2018; Sinha et al., 2018). Not surprisingly, DRO has been applied to the types of mean-squared estimation problems that we consider.

There are two types of distributional perturbations often studied in DRO. The first type perturbs the likelihood of an outcome (Love and Bayraksan, 2016), while the second one perturbs the value of the outcome itself (Rahimian and Mehrotra, 2019). KL divergence is the canonical type of adversarial perturbation used to account for likelihood misspecification (Nguyen et al., 2020, 2019; Duchi and Namkoong, 2021; Namkoong and Duchi, 2017). The Wasserstein distance is typically used to account for perturbations in actual outcomes. In the context of KL divergence (for the types of models that we consider) Zorzi (2015); Lewis et al. (2017); Zorzi (2016); Yi and Zorzi (2021) study DRO formulations for mean-squared prediction of Gaussian systems. They show that the optimal mean-square estimation policy often remains invariant under adversarial KL-divergence perturbations, although the worst-case adversarial structure changes. On the other hand, when Wasserstein-based perturbations are allowed, Kuhn et al. (2019); Nguyen et al. (2021); Shafieezadeh Abadeh et al. (2018) have shown that the optimal mean-squared estimation policies can be sig-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

nificantly different. In both cases, when using a Gaussian model as th nominal model, the worst-case non-parametric estimation is possible, as the worst-case adversaries remain Gaussian. Our work contributes to this line of research by studying the impact of natural constraints in the adversarial Wasserstein-based perturbations case. We choose the Wasserstein distance because we are interested in comparing with adversarial strategies that affect the estimation policy.

In situations involving a stochastic process over time (as the one we study), a direct DRO formulation without constraints may result in over-conservative estimators (Li et al., 2022; Liu et al., 2021). This is because the unconstrained adversary has the ability to look into the future. Endowing the adversary with a degree of "clairvoyance" would be justified in some settings, like competitive markets where some players may have enhanced information sets. We propose to include martingale constraints for the system's innovations for adversarial distributions. These constraints, as we shall see, will ensure that the optimal worst-case adversary is "adapted" to the information generated by the stochastic process in time.

Moreover, martingale constraints for the innovations are often adopted in applications such as economics, signal processing, and engineering within the framework of linear state space dynamics, which is our focus here (Zheng et al., 2015; Oh and Lee, 2018; Kara et al., 1974; Chen and Caines, 1985). These constraints accommodate i.i.d. innovations, which lead to Markovian state-space models, often the models of choice in applications. Therefore, martingale constraints are natural in our setting, as the adversary has the ability to relax Markovianity while preserving information adaptivity.

We summarize our contributions as follows:

1. We provide a novel DRO formulation for linear state estimation with martingale constraints to overcome over-conservative solutions while preserving adaptivity.
2. We show that the DRO estimation problem is tractable and further develops a convex subgradient descent method.
3. If the nominal distribution is Gaussian, we find out that the worst-case adversary is Gaussian and thus DRO non-parametric estimation is possible even with martingale constraints.
4. We provide numerical experiments to showcase that martingale constraints can improve robustness in the uncertainty size and decrease estimation error when nominal models are fully misspecified but share the martingale innovations property with the out-of-sample distribution.

Paper Organization. The rest of the paper is organized as follows. In Section 2, we introduce some notions related to DRO that are necessary for our analysis and present the discrete-time linear model under consideration. In Section 3, we introduce the distributionally robust estimation problem with martingale constraints and develop our main results. Finally, in Section 4, we evaluate the performance of our model through simulation experiments and discuss the results, providing some useful insights.

Notation. Let $(\mathcal{Y}, d)$ be a metric space and $\mathcal{B}(\mathcal{Y})$ the associated Borel σ-algebra. We use $\mathcal{P}(\mathcal{Y})$ to denote the set of probability measures on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. For $\mathbb{P} \in \mathcal{P}(\mathcal{Y})$, $\mathbb{E}_{\mathbb{P}}$ denotes integration over $\mathbb{P}$, that is, $\mathbb{E}_{\mathbb{P}}[f(Z)] = \int_{\mathcal{Y}} f(z)\mathrm{d}\,\mathbb{P}(z)$. For $\pi \in \mathcal{P}(\mathcal{Y} \times \mathcal{Y})$, when $(Z, \bar{Z}) \sim \pi$, we denote $\mathbb{E}_{\pi}\big[c(Z, \bar{Z})\big] = \int_{\mathcal{Y}\times\mathcal{Y}} c(z, \bar{z})\mathrm{d}\pi(z, \bar{z})$ and $\mathbb{E}_{\pi}\big[f(\bar{Z})\big] = \int_{\mathcal{Y}\times\mathcal{Y}} f(\bar{z})\mathrm{d}\pi(z, \bar{z})$. Moreover, we use π_Z and $\pi_{\bar{Z}}$ to denote the marginal distribution of Z and $\bar{Z}$, respectively. Furthermore, we use $\mathcal{N}(\mu, \sigma^2)$ to denote the 1-dimensional normal distribution with mean μ and variance σ^2, and $\mathcal{N}_d(\mu, \Sigma)$ to denote the d-dimensional normal distribution with mean μ and covariance matrix Σ. We use I_d to denote the $d \times d$ identity matrix. Finally, we use $\mathcal{L}$ to denote the set of linear functions.

2 PRELIMINARIES

In this section, we present some notions and definitions that will be very useful for our analysis. We start by introducing the optimal transport cost, the distributionally robust optimization framework, and the linear state-space model under consideration.

2.1 Wasserstein Distance and DRO

First, we give the definition of the optimal transport cost between two probability measure μ and ν and relate it with the Wasserstein distance.

Definition 1 (Optimal Transport cost)**.** Let $c : \mathbb{R}^p \times \mathbb{R}^p \to [0, +\infty)$ be a lower semi-continuous function, satisfying $c(z, \bar{z}) = 0$ if and only if $z = \bar{z}$, and μ, ν probability measures on $\mathbb{R}^p$. The optimal transport cost between μ and ν under the cost function c is defined as:

$$\mathcal{W}_c(\mu, \nu) = \inf_{\pi \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R}^p)} \left\{ \mathbb{E}_{\pi}\big[c(Z, \bar{Z})\big] : \begin{matrix} \pi_Z = \mu \\ \pi_{\bar{Z}} = \nu \end{matrix} \right\}$$

where $\mathcal{P}(\mathbb{R}^p \times \mathbb{R}^p)$ is the set of probability distribution in $\mathbb{R}^p \times \mathbb{R}^p$, and $\pi_Z, \pi_{\bar{Z}}$ denote the marginal distributions of Z and $\bar{Z}$, respectively.

Intuitively, $\mathcal{W}_c(\mu, \nu)$ represents the minimum cost for transporting the mass from the source measure μ to the target one ν, while $\pi(A, B)$ indicates the amount of mass transported from A to B for $A, B \in \mathcal{B}(\mathbb{R}^p)$, where $\mathcal{B}(\mathbb{R}^p)$ is the Borel σ-algebra in $\mathbb{R}^p$.

Remark 1. When the cost function c is defined as $c(z, \bar{z}) := \|z - \bar{z}\|_2^2$, it gives rise to the type-2 Wasserstein squared distance, in which case $\mathcal{W}_c(\mu, \nu)$ is denoted by $\mathcal{W}_2^2(\mu, \nu)$.♦

Distributionally Robust Optimization (DRO) deals with the problem of optimizing the functional $\mathbb{E}_{\mathbb{P}}\big[f(\bar{Z})\big]$, with respect to $\mathbb{P}$, in the neighborhood of a misspecified model $\mathbb{P}_0$ (Blanchet and Murthy, 2019). The standard DRO formulation with transportation cost c, we consider, is given by the following problem:

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathbb{R}^p)} \big\{ \mathbb{E}_{\mathbb{P}}\big[f(\bar{Z})\big] : \mathcal{W}_c(\mathbb{P},\mathbb{P}_0) \leq \delta \big\} \tag{1}$$

for f an upper semi-continuous function and δ the confidence parameter or the radius of the uncertainty region. In other words, δ captures the amount of trust we have in the nominal model $\mathbb{P}_0$. Equivalently (Blanchet and Murthy, 2019), we can write Eq. (1) as

$$\sup_{\pi\in\mathcal{P}(\mathbb{R}^p\times\mathbb{R}^p)} \big\{ \mathbb{E}_{\pi}\big[f(\bar{Z})\big] : \mathbb{E}_{\pi}\big[c(Z,\bar{Z})\big] \leq \delta,\ \pi_Z = \mathbb{P}_0 \big\}$$

We proceed by defining the linear state-space model and the formulation of our distributionally robust estimation problem.

2.2 Discrete-time Linear Dynamics and State Estimation

We consider a linear state-space model for the finite time horizon $n = 1,\dots,T$:

$$\begin{aligned} X_n &= D_n X_{n-1} + \eta_n \\ Y_n &= B_n X_n + \varepsilon_n \end{aligned} \tag{2}$$

where $\varepsilon_n \in \mathbb{R}^d$ are independent zero-mean random vectors with finite second moment, and $\eta_n \in \mathbb{R}^m$ are also independent zero-mean random vectors with finite second moment and independent of ε_n. Moreover, $D_n \in \mathbb{R}^{m\times m}$ and $B_n \in \mathbb{R}^{d\times m}$ are non-random matrices describing the underlying system. For simplicity, we assume $X_0 = 0$. Letting $\eta := (\eta_1,\dots,\eta_T)$ be the innovation process throughout the entire horizon, $\varepsilon := (\varepsilon_1,\dots,\varepsilon_T)$ be the noise process and $Z := (\eta,\varepsilon)$ the joint random vector, we define Ω to be the space where the random quantity Z lives, i.e., $\Omega := \mathbb{R}^{m\cdot T}\times\mathbb{R}^{d\cdot T}$. Finally, we denote its nominal distribution by $\mathbb{P}_0 \in \mathcal{P}(\Omega)$, and we assume that the covariance matrix of (X,Y) under $\mathbb{P}_0$ has full rank, where $X := (X_1,\dots,X_T)$ and $Y := (Y_1,\dots,Y_T)$.

Remark 2. The random vector Z, or equivalently, $\mathbb{P}_0$, captures all the randomness of the linear system (2). ♦

In Eq. (2), the variable X_n represents the unobservable state of the system, while Y_n the noisy observation at time n. Optimal filtering deals with the problem of finding the most accurate estimator of the hidden state X_n based on the observation history $(Y_1,\dots,Y_n)$. It is known (Anderson and Moore, 1979) that the minimum variance estimator of X_n given $(Y_1,\dots,Y_n)$ is the conditional expectation $\mathbb{E}[X_n \mid Y_1,\dots,Y_n]$, in the sense that it minimizes the mean-squared error $\mathbb{E}\big[\|X - g(Y)\|_2^2\big]$ among all functions g with $\int g^2(y)\mathrm{d}\,\mathbb{P} < \infty$. However, the conditional expectation is hard to compute efficiently in general. Instead, a class of estimators that balances between computational efficiency and estimation accuracy in practice is that of *linear estimators*, denoted by $\mathcal{L}$. Formally, the estimation problem under consideration can be written as:

$$\inf_{\phi_1,\dots,\phi_T\in\mathcal{L}} \left\{ \sum_{n=1}^{T} \mathbb{E}\big[\|X_n - \phi_n(Y_1,\dots,Y_n)\|_2^2\big] \right\} \tag{3}$$

where $\phi_n \in \mathcal{L}$ are linear functions.

For reasons that will become apparent next, it is convenient to unfold the linear equations (2) as follows: Using the linearity and the recursive nature of the system, X_n can be expressed as a linear combination of $\eta_1,\dots,\eta_n$, i.e., X_n can be written as:

$$X_n = \eta_n + \sum_{i=1}^{n-1}\left(\prod_{j=i+1}^{n} D_j\right)\eta_i = \sum_{i=1}^{n}\psi_{n,i}\eta_i \tag{4}$$

with $\psi_{n,i} = \prod_{j=i+1}^{n} D_j$ for $1 \leq i < n$ and $\psi_{n,n} = I_m$.Thus, Y_n can be written as:

$$Y_n = B_n\sum_{i=1}^{n}\psi_{n,i}\eta_i + \varepsilon_n = \sum_{i=1}^{n}\hat{\psi}_{n,i}\eta_i + \varepsilon_n \tag{5}$$

where $\hat{\psi}_{n,i} = B_n\psi_{n,i}$.

2.3 Distributionally Robust Estimation

Under model misspecification, the goal of distributionally robust (DR) estimation is to find the best estimator under the worst-case distribution within some uncertainty region. The type-2 Wasserstein distance has been widely used as a measure of divergence from the nominal model (Shafieezadeh Abadeh et al., 2018; Nguyen et al., 2021; Wang and Ye, 2022). In that case, the DR estimation problem would be defined as:

$$\inf_{\substack{\phi_n\in\mathcal{L}\\ n=1,..,T}} \sup_{\mathcal{W}_2^2(\mathbb{P},\mathbb{P}_0)\leq\delta} \left\{ \sum_{n=1}^{T} \mathbb{E}_{\mathbb{P}}\big[\|X_n - \phi_n(Y_1,\dots,Y_n)\|_2^2\big] \right\} \tag{6}$$

Intuitively, the estimation problem (6) can be viewed as a zero-sum game between a statistician and a powerful adversary, who tries to perturb the distribution in a way that results to the worst possible estimation with respect to the mean-squared error. Without further restrictions on the distributions $\mathbb{P}$ and $\mathbb{P}_0$, solving (6) might be intractable. Indeed, previous works (Shafieezadeh Abadeh et al., 2018; Nguyen et al., 2021; Wang and Ye, 2022) that study the problem of distributionally robust estimation of a linear state-space system, add the extra assumption that both $\mathbb{P}_0$ and $\mathbb{P}$ are Gaussian distributions.

In this work, we relax the Gaussian assumption by adding a set of martingale constraints, as described in Section 3.

3 DRO WITH MARTINGALE CONSTRAINTS

In this section, we define our distributionally robust estimation problem under martingale constraints for the linear state-space model discussed before.

3.1 Distributionally Robust Linear-quadratic Estimator

We consider the following distributionally perturbed linear state-space model:

$$\begin{aligned}\tilde{X}_n &= D_n\tilde{X}_{n-1} + \tilde{\eta}_n \\ \tilde{Y}_n &= B_n\tilde{X}_n + \tilde{\varepsilon}_n\end{aligned} \tag{7}$$

where $\tilde{\eta}_n$ and $\tilde{\varepsilon}_n$ denote the perturbed noise processes from the nominal ones, η_n and ε_n, respectively.

For $Z = (\eta, \varepsilon), \tilde{Z} = (\tilde{\eta}, \tilde{\varepsilon})$ random vectors in Ω, we consider the transportation cost c defined as:

$$c(Z, \tilde{Z}) := \|\varepsilon - \tilde{\varepsilon}\|_2^2 + \infty\|\eta - \tilde{\eta}\|_2^2 \tag{8}$$

with the convention that $0 \cdot \infty = 0$ (Blanchet et al., 2019; Blanchet and Murthy, 2019). Implicitly, this cost function enforces that $\eta = \tilde{\eta}$ almost surely under the worst-case distribution. Thus, from now on we neglect the second term of the cost function and set $\eta = \tilde{\eta}$, since the worst-case distribution chosen by the adversary under this transportation cost function automatically satisfies $\eta = \tilde{\eta}$ almost surely.

Remark 3. This choice of transportation cost function still allows the adversary to correlate $\tilde{\varepsilon}$ with η, in order to use the innovation process η in their favor, as we show in Theorem 1. So, this modeling choice does not severely limit the adversarial power. ♦

To obtain a more direct relation between the nominal model (2) and the perturbed one (7), we define $\Delta_n := \tilde{\varepsilon}_n - \varepsilon_n$. Thus, (7) becomes:

$$\begin{aligned}\tilde{X}_n &= D_n\tilde{X}_{n-1} + \eta_n \\ \tilde{Y}_n &= B_n\tilde{X}_n + \varepsilon_n + \Delta_n\end{aligned} \tag{9}$$

where, now, η_n, ε_n and Δ_n are *martingale differences* with finite second moment, as described below.

Formally, defining the underlying filtration $\mathcal{F} = (\mathcal{F}_n)_{n=1}^T$ such that $\mathcal{F}_n$ includes all the information known by a powerful adversary up to time n, i.e.,

$$\mathcal{F}_n = \sigma(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^n) \tag{10}$$

with $\mathcal{F}_0 = \{\emptyset, \Omega\}$. So, we restrict the adversary by introducing the following constraints:

(i) *Wasserstein-distance*:

$$\textstyle\sum_{n=1}^T \mathbb{E}_\pi\left[\|\Delta_n\|_2^2\right] \le \delta$$

(ii) *Martingale-difference*:

$$\mathbb{E}_\pi[(\eta_n, \varepsilon_n, \Delta_n) \mid \mathcal{F}_{n-1}] = 0$$

for all $n = 1, \dots, T$.

Note that ε_n and η_n are independent under $\mathbb{P}_0$, as defined in the nominal model (2).

The *Wasserstein-distance* constraint restricts the worst-case distribution, or, equivalently, the distribution of the perturbations $\Delta_1, \dots, \Delta_n$ within the type-2 squared Wasserstein ball of radius δ, centered at $\mathbb{P}_0$. The *Martingale-difference* constraints, combined with the fact that the objective function is quadratic, guarantee that the adversarial perturbations do not use future information, i.e., realization of the processes in the future, as we show next in Theorem 1. Therefore, this set of constraints makes the adversary to be *adapted*, which is a natural requirement in most time-dependent real-world applications.

Remark 4. If $\Delta_n = 0$ a.s. for all n, then model (9) collapses to the nominal one in (2). ♦

Now, we define the *Wasserstein with martingale constraints* distributional uncertainty region over $\mathcal{P}(\Omega \times \Omega)$ as:

$$\mathcal{D}_\delta = \left\{ \begin{array}{r} \textstyle\sum_{n=1}^T \mathbb{E}_\pi\left[\|\Delta_n\|_2^2\right] \le \delta \\ \pi \in \mathcal{P}(\Omega \times \Omega) : \ \mathbb{E}_\pi[(\eta_n, \varepsilon_n, \Delta_n) \mid \mathcal{F}_{n-1}] = 0 \\ \pi_{(\eta,\varepsilon)} = \mathbb{P}_0 \end{array} \right\}$$

Our goal is to find the best linear predictor of $(\tilde{X}_1, \dots, \tilde{X}_T)$ given the observations $(\tilde{Y}_1, \dots, \tilde{Y}_T)$ under the worst-case distribution in $\mathcal{D}_\delta$. Hence, we consider the following DR estimation problem:

$$\text{OPT} := \inf_{\phi \in \mathcal{L}} \sup_{\pi \in \mathcal{D}_\delta} \quad f(\phi, \pi) \tag{11}$$

where $f(\phi, \pi) := \sum_{n=1}^T \mathbb{E}_\pi\left[\|\tilde{X}_n - \phi_n(\tilde{Y}_1, \dots, \tilde{Y}_n)\|_2^2\right]$ and by ϕ we denote the linear functions $\phi_1, \dots, \phi_T$. Notice that when predicting $\tilde{X}_n$, we use only the observations $(\tilde{Y}_1, \dots, \tilde{Y}_n)$, i.e., our predictor does not make use of the future information $(\tilde{Y}_{n+1}, \dots, \tilde{Y}_T)$.

Remark 5. For each linear function $\phi_n \in \mathcal{L}$ we have $\phi_n(\tilde{Y}_1, \dots, \tilde{Y}_n) = \sum_{j=1}^n \phi_{n,j}\tilde{Y}_j$, where each $\phi_{n,j}$ is a $m \times d$ matrix. From now on, we will write the objective function $f(\phi, \pi)$ as:

$$f(\phi, \pi) = \mathbb{E}_\pi\left[\|\tilde{X}_n - \textstyle\sum_{j=1}^n \phi_{n,j}\tilde{Y}_j\|_2^2\right] \tag{12}$$

♦

Moreover, we define the value function V as:

$$V(\phi) := \sup_{\pi \in \mathcal{D}_\delta} f(\phi, \pi) \tag{13}$$

which we will show that it is convex and, ultimately, our goal will be to minimize, V over ϕ, i.e., to find the optimal linear estimator.

Remark 6. Shafieezadeh Abadeh et al. (2018) study the mean-squared distributionally robust Kalman filter under a Wasserstein constraint involving a single time period, where the power of the adversary is replenished at every time step, and a time-by-time robustification and estimation are performed. This formulation is not designed to account for the impact of knock-on effects over a multi-period setting. Different from Shafieezadeh Abadeh et al. (2018), we impose a multi-period Wasserstein constraint, i.e., the time horizon is arbitrary but fixed. In our model, the total power δ of the adversary does not reset between different time steps. Because of this multi-temporal feature, martingale constraints arise naturally to enforce the adaptability both of the adversarial and the estimation policies. ♦

3.2 Impact of Martingale Constraints

To begin with, we present one crucial lemma which is the key to make the non-parameteric DRO estimator effective and tractable with martingale constraints. Technically, the following lemma can help us simplify the cross term when we expand the term $\mathbb{E}_\pi\Big[\|\tilde{X}_n - \sum_{j=1}^n \phi_{n,j}\tilde{Y}_j\|_2^2\Big]$.

Lemma 1. *For $\pi \in \mathcal{D}_\delta$ and $i \neq j$ it holds:*

(a) $\mathbb{E}_\pi\big[\tilde{\varepsilon}_j^\top A\tilde{\varepsilon}_i\big] = \mathbb{E}_\pi\big[\tilde{\varepsilon}_j^\top B\eta_i\big] = 0$ *for any* $A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times m}$.

(b) $\mathbb{E}_\pi\big[\tilde{\varepsilon}_i^\top B\eta_i\big] = \mathbb{E}_\pi\big[\Delta_i^\top B\eta_i\big]$ *for any* $B \in \mathbb{R}^{d\times m}$.

Proof. Let $i > j$.

(a) We have:

$$\begin{aligned}\mathbb{E}_\pi\big[\tilde{\varepsilon}_j^\top A\tilde{\varepsilon}_i\big] &= \mathbb{E}_\pi\big[\mathbb{E}_\pi\big[\tilde{\varepsilon}_j^\top A\tilde{\varepsilon}_i \mid \mathcal{F}_j\big]\big]\\ &= \mathbb{E}_\pi\big[\tilde{\varepsilon}_j^\top A\,\mathbb{E}_\pi[\tilde{\varepsilon}_i \mid \mathcal{F}_j]\big] = 0\end{aligned}$$

where we used the tower property and the fact that $\mathbb{E}_\pi[\mathbb{E}_\pi[\tilde{\varepsilon}_i \mid \mathcal{F}_{i-1}] \mid \mathcal{F}_j] = 0$. Similarly, we conclude that $\mathbb{E}_\pi\big[\tilde{\varepsilon}_j^\top B\eta_i\big] = 0$.

(b) It follows directly by writing $\tilde{\varepsilon}_i = \varepsilon_i + \Delta_i$ and using the independence of ε_i and η_i under $\mathbb{P}_0$.

■

Now, we are ready to state the following proposition.

Proposition 1. *The objective function $f(\phi, \pi)$ in (11), can be written as:*

$$\mathbb{E}_\pi\Bigg[\sum_{i=1}^T \eta_i^\top \kappa_i(\phi)\eta_i + \sum_{i=1}^T \tilde{\varepsilon}_i^\top c_i(\phi)\tilde{\varepsilon}_i - 2\sum_{i=1}^T \eta_i^\top a_i(\phi)\tilde{\varepsilon}_i\Bigg] \tag{14}$$

where we have $c_i(\phi) := \sum_{n=i}^T \phi_{n,i}^\top\phi_{n,i} \in \mathbb{R}^{d\times d}$, $\kappa_i(\phi) := \sum_{n=i}^T A_{n,i}(\phi)^\top A_{n,i}(\phi) \in \mathbb{R}^{m\times m}$, $a_i(\phi) := \sum_{n=i}^T A_{n,i}(\phi)^\top\phi_{n,i} \in \mathbb{R}^{m\times d}$, *and* $A_{n,i}(\phi) = \psi_{n,i} - \sum_{j=i}^n \phi_{n,j}\hat{\psi}_{j,i} \in \mathbb{R}^{m\times m}$.

Sketch of Proof. Expanding the quantity inside the expectation, and after some algebraic manipulations, we invoke Lemma 1 to cancel the cross-terms of the form $\mathbb{E}_\pi\big[\tilde{\varepsilon}_i^\top A\tilde{\varepsilon}_j\big]$ and $\mathbb{E}_\pi\big[\tilde{\varepsilon}_i^\top B\eta_j\big]$ for $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times m}$. The full proof can be found in the appendix. ■

Intuitively, the *Martingale-difference* constraints allow us to simplify the objective function in the more elegant expression of Proposition 1. At the same time, we see that this new expression does not involve inner products of different time-steps, allowing us to obtain a solution that respects the adaptability property.

3.3 Finding the Worst-case Distribution

We will now focus on the inner maximization problem of (11), i.e.,

$$\text{Val(P)} := \sup_{\pi\in\mathcal{D}_\delta} \sum_{n=1}^T \mathbb{E}_\pi\Bigg[\|\tilde{X}_n - \sum_{j=1}^n \phi_{n,j}\tilde{Y}_j\|_2^2\Bigg] \tag{15}$$

for some fixed $\phi \neq 0$. The case where $\phi = 0$ is trivial, since any $\pi \in \mathcal{D}_\delta$ gives the same objective value. In the following theorem, we show that problem (15) attains the supremum for $\tilde{\pi} \in \mathcal{D}_\delta$ and we fully characterize a optimal solution for any $\phi \neq 0$.

Theorem 1. *For $\phi \neq 0$, one of optimal solution of (15) is of the form:*

$$\tilde{\Delta}_i = \begin{cases} (c_i(\phi) - \tilde{\lambda}I)^{-1}\big(-c_i(\phi)\varepsilon_i + a_i(\phi)^\top\eta_i\big), & i \in \mathcal{C} \\ 0, \quad \textit{otherwise} \end{cases}$$

where $\mathcal{C} = \{i : c_i(\phi) \neq 0\}$ and $\tilde{\lambda} > 0$.

Sketch of Proof. The idea of our proof is based on duality theory; after, formulating the dual problem of (15), we will find a primal-dual optimal pair. The result will follow directly by verifying that strong duality holds.

Note that the primal problem has an *uncountable* set of constraints; each one of the three constraints

$$\mathbb{E}_\pi[(\eta_n, \varepsilon_n, \Delta_n) \mid \mathcal{F}_{n-1}] = 0$$

is a pathwise equality for any realization of $\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{n-1}$. Hence, the associated dual variables for the aforementioned constraints, namely $\xi_n(\cdot), \beta_n(\cdot), \zeta_n(\cdot)$ for $n = 1, \ldots, T$, can be chosen to be Lipschitz continuous functions of $\big(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{n-1}\big)$. Formally, the dual problem is defined:

$$\text{Val(D)} := \inf_{\substack{\lambda\geq 0\\ \xi_i,\beta_i,\zeta_i}} \lambda\delta + \mathbb{E}_{\mathbb{P}_0}\Big[\sup_\Delta h(\Delta)\Big] \tag{16}$$

where by $h(\Delta)$ we denote the quantity:

$$\sum_{i=1}^{T}\Big[\Delta_i^\top(c_i(\phi)-\lambda I)\Delta_i+2\Delta_i^\top\big(c_i(\phi)\varepsilon_i-a_i(\phi)^\top\eta_i\big)+\xi_i\big(\{\eta_t,\varepsilon_t,\Delta_t\}_{t=1}^{i-1}\big)^\top\eta_i+\beta_i\big(\{\eta_t,\varepsilon_t,\Delta_t\}_{t=1}^{i-1}\big)^\top\varepsilon_i+\zeta_i\big(\{\eta_t,\varepsilon_t,\Delta_t\}_{t=1}^{i-1}\big)^\top\Delta_i\Big] \tag{17}$$

As noted before, each one of the dual variables ξ_i, β_i and ζ_i is a Lipschitz continuous function of $\{\eta_t,\varepsilon_t,\Delta_t\}_{t=1}^{i-1}$ associated with the primal constraint $\mathbb{E}_\pi[(\eta_i,\varepsilon_i,\Delta_i)\mid\mathcal{F}_{i-1}]=0$, accordingly, and $\lambda\geq 0$ is the dual variable associated with the *Wasserstein* constraint $\sum_{n=1}^{T}\mathbb{E}_\pi\big[\|\Delta_n\|_2^2\big]\leq\delta$.

By weak duality, we automatically obtain that:

$$\text{Val(P)}\leq\text{Val(D)} \tag{18}$$

As a next step, we guess and verify that a primal-dual optimal pair is of the form:

a) $\tilde{\Delta}_i:=(c_i(\phi)-\tilde{\lambda}I)^{-1}\big(-c_i(\phi)\varepsilon_i+a_i(\phi)^\top\eta_i\big)$ for $i=1,\dots,T$ under $\tilde{\pi}$.

b) $\tilde{\lambda}$ is the greatest solution of equation:

$$\sum_{n=1}^{T}\mathbb{E}_\pi\Big[\|\tilde{\Delta}_n\|_2^2\Big]=\delta \tag{19}$$

and $\tilde{\xi}_i(\cdot)=\tilde{\beta}_i(\cdot)=\tilde{\zeta}_i(\cdot)=0$ for all $i=1,\dots,T$.

with the same primal and dual objective values.

Therefore, combining it with (18), we conclude that strong duality holds, i.e.,

$$\text{Val(P)}=\text{Val(D)}, \tag{20}$$

and the coupling $\tilde{\pi}$, whose second marginal is the law of $(\eta,\varepsilon+\tilde{\Delta})$, gives the worst-case distribution in $\mathcal{D}_\delta$.

The full proof can be found in the appendix. ■

Remark 7. We interchangeably use $\tilde{\Delta}$ or $\tilde{\pi}$ to denote the worst-case distribution, as appropriate. ♦

Remark 8. The solution $\tilde{\pi}$ is parametrized by ϕ, i.e., $\tilde{\pi}_\phi$. However, we omit it for notational convenience, when it is clear from context. Similarly for $\tilde{\Delta}$. ♦

Corollary 1. *For $\phi\neq 0$, if (η,ε) is normally distributed under $\mathbb{P}_0$, then there exist a worst-case coupling $\tilde{\pi}_\phi$ such that $\tilde{\Delta}_\phi$ is also normally distributed under $\tilde{\pi}_\phi$.*

Proof. The conclusion follows immediately, since $\tilde{\Delta}_\phi$ is a linear transformation of (η,ε) under $\tilde{\pi}_\phi$. ■

Remark 9. Note that even in the case that (η,ε) is normally distributed under $\mathbb{P}_0$, there might be other, non-Gaussian, solutions of (15). ♦

This result is remarkable because it enables non-parametric DRO estimation of the hidden state under both martingale and Wasserstein constraints. Precisely, when $\mathbb{P}_0$ is Gaussian, ϕ can be chosen to be a general L^2 function, but an optimal ϕ exists within the class of affine functions. This complements previous results by Zorzi (2016); Kuhn et al. (2019); Nguyen et al. (2021); Shafieezadeh Abadeh et al. (2018) which show that without martingale constraints (both with Wasserstein and KL divergence uncertainty) the worst case distribution remains Gaussian.

3.4 Finding an Optimal Estimator

In order to solve (11) and obtain an optimal estimator, we first argue that the value function is convex. Formally, we have the following lemma.

Lemma 2. *The value function $V(\phi)$ is a convex finite-valued function.*

Proof. For any π in the constraint set $\mathcal{D}_\delta$, the function $\phi\mapsto\sum_{n=1}^{T}\mathbb{E}_\pi\Big[\|\tilde{X}_n-\sum_{j=1}^{n}\phi_{n,j}\tilde{Y}_j\|_2^2\Big]$ is convex. Taking the supremum over $\pi\in\mathcal{D}_\delta$ we readily get that $V(\phi)$ is a convex function as the pointwise supremum of convex functions. The full proof can be found in the appendix. ■

Our strategy for obtaining an optimal estimator, will be to perform subgradient descent on the value function V. The following proposition implements a subgradient oracle for V at each ϕ, and its proof can be found in the appendix.

Proposition 2. *Let ϕ', and $\tilde{\pi}_{\phi'}$ the corresponding worst-case distribution in $\mathcal{D}_\delta$, as per Theorem 1. Then, defining*

$$g:=\nabla_\phi f(\phi,\tilde{\pi}_{\phi'})\mid_{\phi=\phi'} \tag{21}$$

it holds that $g\in\partial V(\phi')$.

The proof of Proposition 2 can be found in the appendix. Now, we claim that the value function V is coercive. For this, suppose that the function $\phi\mapsto\sum_{n=1}^{T}\mathbb{E}_{\mathbb{P}_0}\Big[\|X_n-\sum_{j=1}^{n}\phi_{n,j}Y_j\|_2^2\Big]$ is coercive. Then $V(\phi)$ is also coercive. To see this, by the definition of the V, we have that:

$$V(\phi)\geq\sum_{n=1}^{T}\mathbb{E}_{\mathbb{P}_0}\Big[\|X_n-\sum_{j=1}^{n}\phi_{n,j}Y_j\|_2^2\Big] \tag{22}$$

Taking $\|\phi\|_2\to\infty$, we obtain that $V(\phi)\to\infty$, as the RHS of (22) goes to infinity.

A sufficient condition for the function $\phi\mapsto\sum_{n=1}^{T}\mathbb{E}_{\mathbb{P}_0}\Big[\|X_n-\sum_{j=1}^{n}\phi_{n,j}Y_j\|_2^2\Big]$ to be coercive is that the covariance matrix of the random vector (X,Y) under $\mathbb{P}_0$ is positive definite, which is true by the assumptions imposed on $\mathbb{P}_0$ in Section 2.

By coercivity of $V(\phi)$, it is natural to assume that ϕ lives in a compact and convex set, denoted by Φ. By continuity

of V in the compact set Φ, we automatically get that V is Lipischitz continuous.

In Algorithm 1 we present our subgradient-based algorithm for computing Martingale Distributionally Robust Estimator (MaDRE), with $\tilde{\mathcal{O}}(k^{-1/2})$ convergence rate, where k is the iteration number (Nesterov, 2014), with each iteration performing spectral decomposition of symmetric matrices, where known efficient methods can be leveraged.

Algorithm 1: Martingale Distributionally Robust Estimator (MaDRE)

Initialize: $\phi_1 \in \Phi$, step-size schedule $\gamma_t \propto 1/\sqrt{t}$
for $t = 1, 2, \dots$ **do**
  Get $\tilde{\Delta}_{\phi_t}$, as per Theorem 1.
  Compute $g_t \in \partial V(\phi_t)$ as per Proposition 2.
  $\phi_{t+1} \leftarrow \text{proj}_{\Phi}(\phi_t - \gamma_t g_t / \|g_t\|_2)$
end for

4 PERFORMANCE EVALUATION

In this section, we describe our simulations setup and compare our model against that of Nguyen et al. (2021). Our goal is to highlight the impact of adding the martingale constraint both in terms of sensitivity to model misspecification and uncertainty size. To do this, we will perform two types of experiments on a simple, yet illustrative, set of models.

(E1) In the first type, we simulate data from a ground truth model. This data is used to fit a nominal model (i.e. $\mathbb{P}_0$) that does not coincide with the ground truth model, which is typically unknown to the modeler. Then, we use DRO both with martingale and without martingale constraints to estimate the hidden state based on the signal. We evaluate the performance of these estimators using the ground truth model out-of-sample via Monte Carlo simulation and study the impact of the uncertainty size power, δ.

(E2) In the second type, we fit a nominal model to data produced with a data generating mechanism which is *different* from the ground truth but still preserves martingale innovations. We again apply DRO with and without martingale constraints and also evaluate the out-of-sample performance of the estimators using the ground truth model. Our focus here is on out-of-sample performance when the data used to fit $\mathbb{P}_0$ is different from the ground truth.

4.1 Ground Truth and Nominal Models

Ground Truth Model. We now describe how do we generate the ground truth model. We focus on simple 1-dimensional models because the insights can already be studied in this simple setting. For $n = 1, \dots, T$:

$$\begin{aligned} X^*_{n+1} &= X^*_n + \eta^*_{n+1} \\ Y^*_{n+1} &= X^*_{n+1} + \varepsilon^*_{n+1}, \end{aligned} \tag{23}$$

with (i) η^*_n's i.i.d. with distribution $\mathcal{N}(0,1)$, and (ii) $\varepsilon^*_n = h_n(X^*_1, \dots, X^*_{n-1})U^*_n$ for U^*_n's which are i.i.d. $\text{unif}(-1,1)$, and $h_n(X^*_1, \dots, X^*_{n-1})$ is a history-dependent random function. Specifically, we consider h_n to be defined as $h_n(X^*_1, \dots, X^*_{n-1}) = (X^*_1 + \cdots + X^*_{n-1}) \cdot q^*_n$ with $q^*_n \sim \mathcal{N}(0,1)$ i.i.d. We assume that the sequences U^*_n, q^*_n and η^*_n are all mutually independent.

Nominal model. We assume that the nominal model, $\mathbb{P}_0$, is a linear Markovian model such that $n = 1, \dots, T$:

$$\begin{aligned} X_{n+1} &= X_n + \eta_{n+1} \\ Y_{n+1} &= X_{n+1} + \varepsilon_{n+1}, \end{aligned} \tag{24}$$

where the η_n's are i.i.d. with distribution $\mathcal{N}(0,1)$, the ε_n's are conditionally independent given the σ_n's with $\varepsilon_n \sim \mathcal{N}(0, \sigma^2_n)$. In turn, the σ_n's are computed as the sample variances of the generated data as we will explain in Section 4.3.

4.2 Equivalent Reformulation

In order to compare our model with that of Nguyen et al. (2021) , we need to bring the linear model (24) into the following formulation:

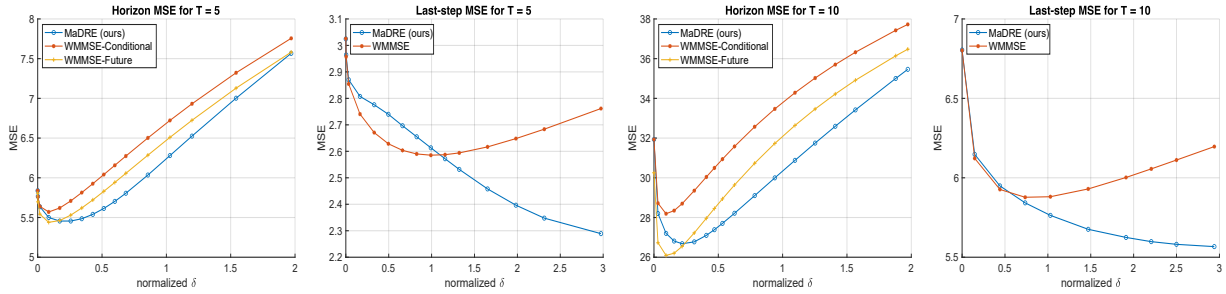
$$Y = HX + \varepsilon \tag{25}$$

where $Y \in \mathbb{R}^T$ is the observation vector, $\varepsilon \in \mathbb{R}^T$ is the additive noise, $X \in \mathbb{R}^T$ is the (unobserved) state of the system, and H the matrix of the system.

It is easy to see that for $\eta_i \sim \mathcal{N}(0,1)$ i.i.d., the state variable X_n in (24), is distributed as normal $\mathcal{N}(0,n)$. Therefore, if we consider the random vector $X = (X_1, \dots, X_T)$, we have that $X \sim \mathcal{N}_T(0, \Sigma_X)$ with $(\Sigma_X)_{ij} = \min\{i,j\}$. Moreover, for $\varepsilon \sim \mathcal{N}_T(0, \Sigma_\varepsilon)$ with $\Sigma_\varepsilon = \text{diag}(\sigma^2_1, \dots, \sigma^2_T)$, and ε independent of X, we get that model (25) is equivalent to (24) under $\mathbb{P}_0$, for H being the identity matrix.

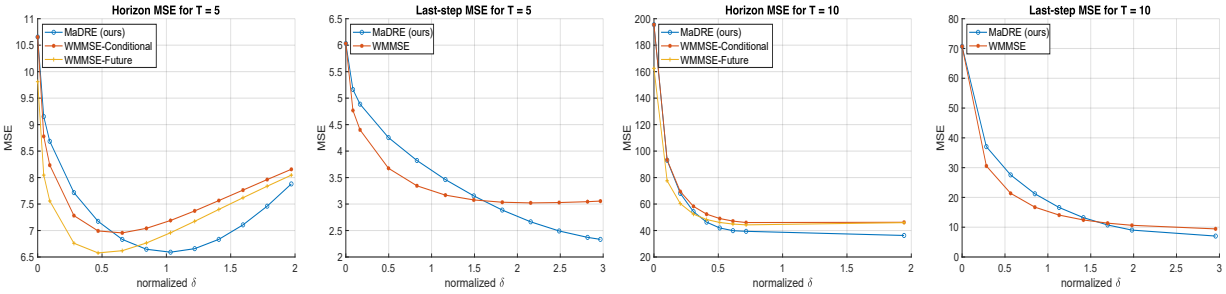
The optimal linear estimator of Nguyen et al. (2021) uses the whole observation profile $(Y_1, \dots, Y_T)$ for all prediction steps. In other words, their prediction of the state X_k at time-step k depends also on $Y_{k+1}, \dots, Y_T$, i.e., future values. So, we consider two interpretations of the estimator in Nguyen et al. (2021):

WMMSE-Future: The standard estimation of Nguyen et al. (2021) which uses also the future observations $Y_{k+1}, \dots, Y_T$ to predict state X_k. This estimator cannot be implemented at time k given only the observations $Y_1, \dots, Y_k$.



(a) Horizon MSE, $T = 5$ **(b)** Last-step MSE, $T = 5$ **(c)** Horizon MSE, $T = 10$ **(d)** Last-step MSE, $T = 10$

Figure 1: $\mathbb{P}_0$ fitted in the data generated by the true model (M1)



(a) Horizon MSE, $T = 5$ **(b)** Last-step MSE, $T = 5$ **(c)** Horizon MSE, $T = 10$ **(d)** Last-step MSE, $T = 10$

Figure 2: $\mathbb{P}_0$ fitted in the data generated by the different model (M2)

WMMSE-Conditional: The estimation that replaces $Y_{k+1}, \ldots, Y_T$ with their conditional expectations under the corresponding worst-case distribution in Nguyen et al. (2021) given the observations $Y_1, \ldots, Y_k$, i.e. with $\bar{Y}_{k+1} := \mathbb{E}[Y_{k+1} \,|\, Y_1, \ldots, Y_k], \ldots, \bar{Y}_T := \mathbb{E}[Y_T \,|\, Y_1, \ldots, Y_k]$.

Remark 10. It is important to note that while WMMSE-Future cannot be implemented with current information, it is useful to keep as a benchmark to quantify the future information value. However, WMMSE-Future can be implemented for X_T. In this case, both WMMSE-Future and WMMSE-Conditional coincide and this special case is studied separately in our numerical experiments. ♦

4.3 Data Generating Mechanisms

Now, we describe the two data generating mechanisms we consider for fitting the nominal model $\mathbb{P}_0$, as described in (24). We simulate 20 independent batches of samples for $T = 5$ and $T = 10$ generated

(M1) Mechanism 1 is used in the first type of experiments (E1). The data is generated directly from the true model. (23),

(M2) Mechanism 2 applies to the second type of experiments (E2). The data is generated from a different model from ground truth. We assume that the noise process is generated as $\bar{\varepsilon}_n \sim \text{unif}(-1, 1)$.

4.4 Evaluation & Discussion

To probe the behavior of MaDRE and evaluate its performance we run a number of experiments, as follows.

For each δ (Wasserstein radius) considered, we generate 1000 batches from the true model (23) and compute the mean-squared error (MSE) of the predictions from our model, and the two variants: WMMSE-Future and WMMSE-Conditional, as discussed before. In each plot, the vertical axis corresponds to the MSE and the horizontal to the *normalized* radius of the uncertainty region δ. We normalize the absolute values of δ in each plot with the respective MSE estimated under $\mathbb{P}_0$, that is, for $\delta = 0$.

Fig. 1 corresponds to the first experiment (E1), where the data for fitting the nominal model $\mathbb{P}_0$ come from the true mechanism (M1), and Fig. 2 corresponds to the experiment (E2), where the data come from mechanism (M2).

In Fig. 1a, Fig. 1c, Fig. 2a, Fig. 2c we compare the out-of-sample MSE of MaDRE to that of WMMSE-Conditional

and WMMSE-Future, throughout the entire horizon T, for $T = 5, 10$. In Fig. 1b, Fig. 1d, Fig. 2b, Fig. 2d, we compare the out-of-sample MSE of the estimators only at the last time-step T. As explained in Remark 10, it is meaningful to compare the error of the last time-step T too since the WMMSE-Conditional and WMMSE-Future coincide and this could be interpreted as a way to directly apply the WMMSE strategy. We use the label WMMSE to denote the estimator in the plots.

In Fig. 1a and Fig. 1c, we see that for any given uncertainty budget MaDRE is uniformly better than WMMSE-Conditional and outperforms (in most cases) even WMMSE-Future which uses future information (both for the adversary but also for the policy). This also is apparent not only comparing fixed uncertainty budgets, but also the best global achievable estimated error over δ. Moreover, it is clear that the MSE of MaDRE changes more smoothly than that of the others as the δ parameter varies, and its error appears consistently lower as the value of δ increases. From this, we see that introducing the martingale constraints can add an extra layer of robustness in the choice of the uncertainty parameter δ. We find this property particularly useful since the process of tuning the hyperparameter δ is not typically easy. This "extra layer of robustness" and its effect on performance becomes even more visible in Fig. 1b and Fig. 1d, where the estimators use the same amount of information.

In Fig. 2a and Fig. 2c, where the data for fitting $\mathbb{P}_0$ are generated by the different mechanism (M2), we observe that the WMMSE-Conditional and WMMSE-Future have greater estimation improvements for smaller values of δ. However, as δ increases, there is a broad region of uncertainty where MaDRE outperforms both WMSSE-Conditional and WMMSE-Future. Overall, the martingale constraints achieve the best global achievable estimated error, for suitable δ. The intuitive reason, we believe, is the following. Because $\mathbb{P}_0$ is fitted to data generated by a different model from the testing model, the model misspecification gap is larger. Therefore, the over-conservative policies produced by WMMSE-Conditional and WMMSE-Future lead to a steepest decrease in the estimation error for smaller values of δ, since the adversary, being less restrictive, has a stronger ability to efficiently help hedge the misspecification gap per marginal unit of uncertainty budget. However, as δ increases, the adversary becomes too powerful, and the martingale constraints start playing a key role since this is a common feature shared both by the data generating process and the out-of-sample distribution. This allows MaDRE to outperform for a much broader parameter space. Similar behavior is observed in Fig. 2b and Fig. 2d.

5 CONCLUDING REMARKS

In this work, we study the impact of martingale constraints on the estimation of the unobserved state of a discrete-time linear state-space model. We show that the worst-case adversary is adapted to the natural time-generated information and develop an efficient subgradient-based algorithm for computing an optimal linear estimator. Our experiments demonstrate that the martingale constraints broadly avoid over-conservative policies, adding an extra layer of robustness in the choice of the size of the uncertainty region.

Acknowledgements

The material in this paper is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397. Additional support is gratefully acknowledged from NSF 1915967 and 2118199. Finally, this work was supported by the Onassis Foundation (F ZR 033-1/2021-2022), and by the Koret Foundation via the "Digital Living 2030" project.

References

Anderson, B. and Moore, J. (1979). *Optimal Filtering*. Prentice-Hall.

Blanchet, J., Kang, Y., and Murthy, K. (2019). Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857.

Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600.

Chen, H. and Caines, P. (1985). On the adaptive control of a class of systems with random parameters and disturbances. *Automatica*, 21(6):737–741.

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).

Delage, E. and Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612.

Duchi, J. C. and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406.

Ghahremani, E. and Kamwa, I. (2011). Dynamic state estimation in power system by applying the extended Kalman filter with unknown inputs to phasor measurements. *IEEE Transactions on Power Systems*, 26(4):2556–2566.

Kara, H., Mandrekar, V., and Park, G. L. (1974). Wide-sense martingale approach to linear optimal estimation. *SIAM Journal on Applied Mathematics*, 27(2):293–302.

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

Lee, G. H., Faundorfer, F., and Pollefeys, M. (2013). Motion estimation for self-driving cars with a generalized camera. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2746–2753.

Lee, J. and Raginsky, M. (2018). Minimax statistical learning with Wasserstein distances. *Advances in Neural Information Processing Systems*, 31.

Lewis, F. L., Xie, L., and Popa, D. (2017). *Optimal and robust estimation: with an introduction to stochastic control theory*. CRC press.

Li, J., Lin, S., Blanchet, J., and Nguyen, V. A. (2022). Tikhonov regularization is optimal transport robust under martingale constraints. *arXiv preprint arXiv:2210.01413*.

Lin, F., Fang, X., and Gao, Z. (2022). Distributionally robust optimization: A review on theory and applications. *Numerical Algebra, Control & Optimization*, 12(1):159.

Liu, J., Shen, Z., Cui, P., Zhou, L., Kuang, K., and Li, B. (2021). Distributionally robust learning with stable adversarial training. *arXiv preprint arXiv:2106.15791*.

Liu, Y., Fan, X., Lv, C., Wu, J., Li, L., and Ding, D. (2018). An innovative information fusion method with adaptive Kalman filter for integrated INS/GPS navigation of autonomous vehicles. *Mechanical Systems and Signal Processing*, 100:605–616.

Love, D. and Bayraksan, G. (2016). Phi-divergence constrained ambiguous stochastic programs for data-driven optimization.

Namkoong, H. and Duchi, J. C. (2017). Variance-based regularization with convex objectives. *Advances in neural information processing systems*, 30.

Nesterov, Y. (2014). *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition.

Nguyen, V. A., Shafieezadeh-Abadeh, S., Kuhn, D., and Mohajerin Esfahani, P. (2021). Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization. *Mathematics of Operations Research*.

Nguyen, V. A., Shafieezadeh Abadeh, S., Yue, M.-C., Kuhn, D., and Wiesemann, W. (2019). Optimistic distributionally robust optimization for nonparametric likelihood approximation. *Advances in Neural Information Processing Systems*, 32.

Nguyen, V. A., Zhang, X., Blanchet, J., and Georghiou, A. (2020). Distributionally robust parametric maximum likelihood estimation. *Advances in Neural Information Processing Systems*, 33:7922–7932.

Oh, H. and Lee, S. (2018). On parameter change test for ARMA models with martingale difference errors. In *Predictive Econometrics and Big Data*, pages 246–254, Cham. Springer International Publishing.

Prakash, S. K. A. and Tucker, C. S. (2018). Bounded Kalman filter method for motion-robust, non-contact heart rate estimation. *Biomed. Opt. Express*, 9(2):873–897.

Rahimian, H. and Mehrotra, S. (2019). Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*.

Ribeiro, A., Schizas, I. D., Roumeliotis, S. I., and Giannakis, G. (2010). Kalman filtering in wireless sensor networks. *IEEE Control Systems Magazine*, 30(2):66–86.

Rudin, W. (1987). *Real and Complex Analysis, 3rd Ed.* McGraw-Hill, Inc., USA.

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.

Sinha, A., Namkoong, H., and Duchi, J. (2018). Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*.

Wang, S. and Ye, Z. S. (2022). Distributionally robust state estimation for linear systems subject to uncertainty and outlier. *IEEE Transactions on Signal Processing*, 70:452–467.

Wiesemann, W., Kuhn, D., and Sim, M. (2014). Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376.

Yang, I. (2020). Wasserstein distributionally robust stochastic control: A data-driven approach. *IEEE Transactions on Automatic Control*, 66(8):3863–3870.

Yi, S. and Zorzi, M. (2021). Robust Kalman Filtering under model uncertainty: the case of degenerate densities. *IEEE Transactions on Automatic Control*.

Zhang, Z., Silva, I., Wu, D., Zheng, J., Wu, H., and Wang, W. (2014). Adaptive motion artefact reduction in respiration and ECG signals for wearable healthcare monitoring systems. *Medical & Biological Engineering & Computing*, 52(12):1019–1030.

Zhao, J., Netto, M., and Mili, L. (2017). A robust iterated extended Kalman filter for power system dynamic state estimation. *IEEE Transactions on Power Systems*, 32(4):3205–3216.

Zheng, T., Xiao, H., and Chen, R. (2015). Generalized ARMA models with martingale difference errors. *Journal of Econometrics*, 189(2):492–506. Frontiers in Time Series and Financial Econometrics.

Zorzi, M. (2015). On the robustness of the Bayes and Wiener estimators under model uncertainty.

Zorzi, M. (2016). Robust Kalman filtering under model perturbations. *IEEE Transactions on Automatic Control*, 62(6):2902–2907.

A MISSING PROOFS

A.1 Proof of Proposition 1

Proposition 1. *The objective function $f(\phi, \pi)$ in (11), can be written as:*

$$\mathbb{E}_\pi \left[\sum_{i=1}^{T} \eta_i^\top \kappa_i(\phi) \eta_i + \sum_{i=1}^{T} \tilde{\varepsilon}_i^\top c_i(\phi) \tilde{\varepsilon}_i - 2 \sum_{i=1}^{T} \eta_i^\top a_i(\phi) \tilde{\varepsilon}_i \right] \tag{14}$$

where we have $c_i(\phi) := \sum_{n=i}^{T} \phi_{n,i}^\top \phi_{n,i} \in \mathbb{R}^{d\times d}$, $\kappa_i(\phi) := \sum_{n=i}^{T} A_{n,i}(\phi)^\top A_{n,i}(\phi) \in \mathbb{R}^{m\times m}$, $a_i(\phi) := \sum_{n=i}^{T} A_{n,i}(\phi)^\top \phi_{n,i} \in \mathbb{R}^{m\times d}$, *and* $A_{n,i}(\phi) = \psi_{n,i} - \sum_{j=i}^{n} \phi_{n,j} \hat{\psi}_{j,i} \in \mathbb{R}^{m\times m}$.

Proof. The general term in (11) can be written as:

$$\begin{aligned}
\mathbb{E}_\pi \left[\|\tilde{X}_n - \sum_{j=1}^{n} \phi_{n,j} \tilde{Y}_j\|_2^2 \right] &= \mathbb{E}_\pi \left[\|\sum_{i=1}^{n} \psi_{n,i}\eta_i - \sum_{j=1}^{n} \phi_{n,j}\tilde{\varepsilon}_j - \sum_{j=1}^{n} \phi_{n,j} \sum_{i=1}^{j} \hat{\psi}_{j,i}\eta_i \|_2^2 \right] \\
&= \mathbb{E}_\pi \left[\|\sum_{i=1}^{n} \psi_{n,i}\eta_i - \sum_{j=1}^{n} \phi_{n,j}\tilde{\varepsilon}_j - \sum_{i=1}^{n} \left(\sum_{j=i}^{n} \phi_{n,j}\hat{\psi}_{j,i} \right) \eta_i \|_2^2 \right] \\
&= \mathbb{E}_\pi \left[\|\sum_{i=1}^{n} \left(\psi_{n,i} - \sum_{j=i}^{n} \phi_{n,j}\hat{\psi}_{j,i} \right) \eta_i - \sum_{i=1}^{n} \phi_{n,i}\tilde{\varepsilon}_i \|_2^2 \right] \\
&= \mathbb{E}_\pi \left[\|\sum_{i=1}^{n} A_{n,i}(\phi)\eta_i - \sum_{i=1}^{n} \phi_{n,i}\tilde{\varepsilon}_i \|_2^2 \right] \\
&= \mathbb{E}_\pi \left[\|\sum_{i=1}^{n} A_{n,i}(\phi)\eta_i\|_2^2 + \|\sum_{i=1}^{n} \phi_{n,i}\tilde{\varepsilon}_i\|_2^2 - 2\left\langle \sum_{i=1}^{n} A_{n,i}(\phi)\eta_i, \sum_{i=1}^{n} \phi_{n,i}\tilde{\varepsilon}_i \right\rangle \right] \\
&= \mathbb{E}_\pi \left[\sum_{i=1}^{n} \|A_{n,i}(\phi)\eta_i\|_2^2 + \sum_{i=1}^{n} \|\phi_{n,i}\tilde{\varepsilon}_i\|_2^2 - 2\sum_{i=1}^{n} \langle A_{n,i}(\phi)\eta_i, \phi_{n,i}\tilde{\varepsilon}_i \rangle \right]
\end{aligned} \tag{A.1}$$

where $A_{n,i}(\phi) = \psi_{n,i} - \sum_{j=i}^{n} \phi_{n,j}\hat{\psi}_{j,i}$ for $1 \le i \le n$. The last equality in (A.1) follows from Lemma 1 and independence of η_n's under $\mathbb{P}_0$. Therefore, the objective function of (11) becomes:

$$\begin{aligned}
f(\phi, \pi) &= \sum_{n=1}^{T} \mathbb{E}_\pi \left[\|\tilde{X}_n - \sum_{j=1}^{n} \phi_{n,j}\tilde{Y}_j\|_2^2 \right] \\
&= \mathbb{E}_\pi \left[\sum_{n=1}^{T} \left(\sum_{i=1}^{n} \|A_{n,i}(\phi)_i\|_2^2 + \sum_{i=1}^{n} \|\phi_{n,i}\tilde{\varepsilon}_i\|_2^2 - 2\sum_{i=1}^{n} \langle A_{n,i}(\phi)\eta_i, \phi_{n,i}\tilde{\varepsilon}_i \rangle \right) \right] \\
&= \mathbb{E}_\pi \left[\sum_{i=1}^{T} \sum_{n=i}^{T} \|A_{n,i}(\phi)\eta_i\|_2^2 + \sum_{i=1}^{T} \sum_{n=i}^{T} \|\phi_{n,i}\tilde{\varepsilon}_i\|_2^2 - 2\sum_{i=1}^{T} \sum_{n=i}^{T} \langle A_{n,i}(\phi)\eta_i, \phi_{n,i}\tilde{\varepsilon}_i \rangle \right] \\
&= \mathbb{E}_\pi \left[\sum_{i=1}^{T} \eta_i^\top \left(\sum_{n=i}^{T} A_{n,i}(\phi)^\top A_{n,i}(\phi) \right) \eta_i + \sum_{i=1}^{T} \tilde{\varepsilon}_i^\top \left(\sum_{n=i}^{T} \phi_{n,i}^\top \phi_{n,i} \right) \tilde{\varepsilon}_i - 2\sum_{i=1}^{T} \eta_i^\top \left(\sum_{n=i}^{T} A_{n,i}(\phi)^\top \phi_{n,i} \right) \tilde{\varepsilon}_i \right]
\end{aligned} \tag{A.2}$$

Setting $c_i(\phi) := \sum_{n=i}^{T} \phi_{n,i}^\top \phi_{n,i}$, $\kappa_i(\phi) := \sum_{n=i}^{T} A_{n,i}(\phi)^\top A_{n,i}(\phi)$ and $a_i(\phi) := \sum_{n=i}^{T} A_{n,i}(\phi)^\top \phi_{n,i}$, we obtain that

$$f(\phi, \pi) = \mathbb{E}_\pi \left[\sum_{i=1}^{T} \eta_i^\top \kappa_i(\phi) \eta_i + \sum_{i=1}^{T} \tilde{\varepsilon}_i^\top c_i(\phi) \tilde{\varepsilon}_i - 2\sum_{i=1}^{T} \eta_i^\top a_i(\phi) \tilde{\varepsilon}_i \right] \tag{A.3}$$

■

A.2 Proof of Theorem 1

Theorem 1. *For $\phi \neq 0$, one of optimal solution of (15) is of the form:*

$$\tilde{\Delta}_i = \begin{cases} (c_i(\phi) - \tilde{\lambda} I)^{-1}\big(-c_i(\phi)\varepsilon_i + a_i(\phi)^\top \eta_i\big), & i \in \mathcal{C} \\ 0, \quad \textit{otherwise} \end{cases}$$

where $\mathcal{C} = \{i : c_i(\phi) \neq 0\}$ and $\tilde{\lambda} > 0$.

Proof. By the reformulation of the objective function as per Proposition 1, we can write the inner maximization problem of (11) as:

$$\begin{aligned}
\mathrm{Val}(P) &= \sup_{\pi \in \mathcal{D}_\delta} \left\{ \mathbb{E}_\pi \left[\sum_{i=1}^T \eta_i^\top \kappa_i(\phi) \eta_i + \sum_{i=1}^T \tilde{\varepsilon}_i^\top c_i(\phi) \tilde{\varepsilon}_i - 2 \sum_{i=1}^T \eta_i^\top a_i(\phi) \tilde{\varepsilon}_i \right] \right\} \\
&= \sup_{\pi \in \mathcal{D}_\delta} \left\{ \mathbb{E}_\pi \left[\sum_{i=1}^T \eta_i^\top \kappa_i(\phi) \eta_i \right] + \mathbb{E}_\pi \left[\sum_{i=1}^T \tilde{\varepsilon}_i^\top c_i(\phi) \tilde{\varepsilon}_i - 2 \sum_{i=1}^T \eta_i^\top a_i(\phi) \tilde{\varepsilon}_i \right] \right\} \\
&= \sup_{\pi \in \mathcal{D}_\delta} \left\{ \mathbb{E}_{\mathbb{P}_0} \left[\sum_{i=1}^T \eta_i^\top \kappa_i(\phi) \eta_i \right] + \mathbb{E}_\pi \left[\sum_{i=1}^T \tilde{\varepsilon}_i^\top c_i(\phi) \tilde{\varepsilon}_i - 2 \sum_{i=1}^T \eta_i^\top a_i(\phi) \tilde{\varepsilon}_i \right] \right\} \\
&= \mathbb{E}_{\mathbb{P}_0} \left[\sum_{i=1}^T \eta_i^\top \kappa_i(\phi) \eta_i \right] + \sup_{\pi \in \mathcal{D}_\delta} \left\{ \mathbb{E}_\pi \left[\sum_{i=1}^T \tilde{\varepsilon}_i^\top c_i(\phi) \tilde{\varepsilon}_i - 2 \sum_{i=1}^T \eta_i^\top a_i(\phi) \tilde{\varepsilon}_i \right] \right\} \\
&= \mathbb{E}_{\mathbb{P}_0} \left[\sum_{i=1}^T \eta_i^\top \kappa_i(\phi) \eta_i \right] + \sup_{\pi \in \mathcal{D}_\delta} \left\{ \mathbb{E}_\pi \left[\sum_{i=1}^T \tilde{\varepsilon}_i^\top c_i(\phi) \tilde{\varepsilon}_i - 2 \sum_{i=1}^T \eta_i^\top a_i(\phi) \tilde{\varepsilon}_i \right] \right\}
\end{aligned} \tag{A.4}$$

Replacing $\tilde{\varepsilon}_i$ by $\Delta_i + \varepsilon_i$ and using the independence of η_i and ε_i under $\mathbb{P}_0$ in (A.4), we get:

$$\mathrm{Val}(P) = \mathbb{E}_{\mathbb{P}_0} \left[\sum_{i=1}^T \eta_i^\top \kappa_i(\phi) \eta_i + \sum_{i=1}^T \varepsilon_i^\top c_i(\phi) \varepsilon_i \right] + \sup_{\pi \in \mathcal{D}_\delta} \left\{ \mathbb{E}_\pi \left[\sum_{i=1}^T \Delta_i^\top c_i(\phi) \Delta_i + 2 \sum_{i=1}^T \Delta_i^\top \big(c_i(\phi) \varepsilon_i - a_i(\phi)^\top \eta_i \big) \right] \right\} \tag{A.5}$$

Hence, we will focus on the problem:

$$\mathrm{Val}(\tilde{P}) := \sup_{\pi \in \mathcal{D}_\delta} \left\{ \mathbb{E}_\pi \left[\sum_{i=1}^T \Delta_i^\top c_i(\phi) \Delta_i + 2 \sum_{i=1}^T \Delta_i^\top \big(c_i(\phi) \varepsilon_i - a_i(\phi)^\top \eta_i \big) \right] \right\} \tag{A.6}$$

Without loss of generality, we can assume that $c_i(\phi) \neq 0$ for all $i = 1, \ldots, T$. Indeed, if $c_i(\phi) = 0$ for some i, this would mean that $\sum_{n=i}^T \phi_{n,i}^\top \phi_{n,i} = 0$, and since each one of the summands is positive semi-definite, we would get that $\phi_{n,i} = 0$ for all $n = i, \ldots, T$. This would also imply that $a_i(\phi) = 0$, since $a_i(\phi) = \sum_{n=i}^T A_{n,i}(\phi)^\top \phi_{n,i}$, and therefore, Δ_i would not be present in the objective function of (A.6).

Obtaining the weak dual. Let Λ be the collection of $\big(\rho, \lambda, \{\xi_i, \beta_i, \zeta_i\}_{i=1}^T\big)$ with the following properties: (i) $\rho(\cdot)$ is a Borel measurable function of (η, ε) in $\bar{\mathbb{R}} := [-\infty, +\infty]$, (ii) $\lambda \geq 0$ is a non-negative real number, (iii) $\xi_i(\cdot), \beta_i(\cdot), \zeta_i(\cdot)$ are Lipschitz-continuous (and, hence, Borel measurable) functions of $\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}$ for $i = 1, \ldots, T$, and (vi) for all $\eta, \varepsilon, \Delta$, it holds:

$$\begin{aligned}
\rho(\eta, \varepsilon) \geq & \sum_{i=1}^T \Delta_i^\top c_i(\phi) \Delta_i + 2 \sum_{i=1}^T \Delta_i^\top \big(c_i(\phi) \varepsilon_i - a_i(\phi)^\top \eta_i \big) - \lambda \sum_{i=1}^T \Delta_i^\top \Delta_i \\
& + \sum_{i=1}^T \xi_i\big(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\big)^\top \eta_i + \sum_{i=1}^T \beta_i\big(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\big)^\top \varepsilon_i \\
& + \sum_{i=1}^T \zeta_i\big(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\big)^\top \Delta_i
\end{aligned} \tag{A.7}$$

Clearly, the set Λ is non-empty, as (A.7) can be satisfied trivially for the function $\rho = \infty$ everywhere. Moreover, it is easy to see that the right-hand side of (A.7) is integrable under any $\pi \in \mathcal{D}_\delta$, since η, ε and Δ have finite second-moments and $\xi_i(\cdot), \beta_i(\cdot), \zeta_i(\cdot)$ are Lipschitz-continuous functions of $\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}$. Then, the integral $\mathbb{E}_\pi[\rho(\eta, \varepsilon)]$ is well-defined and we claim that

$$\begin{aligned} \text{Val}(\tilde{D}) \coloneqq \inf \quad & \lambda\delta + \mathbb{E}_\pi[\rho(\eta, \varepsilon)] \\ \text{s.t.} \quad & \left(\rho, \lambda, \{\xi_i, \beta_i, \zeta_i\}_{i=1}^T\right) \in \Lambda \end{aligned} \tag{A.8}$$

is a weak dual of (A.6). Indeed, let $\left(\rho, \lambda, \{\xi_i, \beta_i, \zeta_i\}_{i=1}^T\right) \in \Lambda$ and $\pi \in \mathcal{D}_\delta$. We have:

$$\begin{aligned} \lambda\delta + \mathbb{E}_\pi[\rho(\eta, \varepsilon)] &\overset{\text{(A.7)}}{\geq} \lambda\delta + \mathbb{E}_\pi\Bigg[\sum_{i=1}^T \Delta_i^\top c_i(\phi)\Delta_i + 2\sum_{i=1}^T \Delta_i^\top \left(c_i(\phi)\varepsilon_i - a_i(\phi)^\top \eta_i\right) - \lambda\sum_{i=1}^T \Delta_i^\top \Delta_i \\ &\qquad + \sum_{i=1}^T \xi_i\left(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\right)^\top \eta_i + \sum_{i=1}^T \beta_i\left(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\right)^\top \varepsilon_i \\ &\qquad + \sum_{i=1}^T \zeta_i\left(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\right)^\top \Delta_i\Bigg] \\ &= \mathbb{E}_\pi\left[\sum_{i=1}^T \Delta_i^\top c_i(\phi)\Delta_i + 2\sum_{i=1}^T \Delta_i^\top \left(c_i(\phi)\varepsilon_i - a_i(\phi)^\top \eta_i\right)\right] + \lambda\left(\delta - \mathbb{E}_\pi\left[\sum_{i=1}^T \Delta_i^\top \Delta_i\right]\right) \\ &\quad + \mathbb{E}_\pi\Bigg[\sum_{i=1}^T \xi_i\left(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\right)^\top \eta_i + \sum_{i=1}^T \beta_i\left(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\right)^\top \varepsilon_i \\ &\qquad + \sum_{i=1}^T \zeta_i\left(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\right)^\top \Delta_i\Bigg] \\ &\geq \mathbb{E}_\pi\left[\sum_{i=1}^T \Delta_i^\top c_i(\phi)\Delta_i + 2\sum_{i=1}^T \Delta_i^\top \left(c_i(\phi)\varepsilon_i - a_i(\phi)^\top \eta_i\right)\right] \end{aligned} \tag{A.9}$$

where the last inequality holds because for $\pi \in \mathcal{D}_\delta$, we have:

- $\delta \geq \mathbb{E}_\pi\left[\sum_{i=1}^T \Delta_i^\top \Delta_i\right]$
- $\mathbb{E}_\pi\left[\xi_i\left(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\right)^\top \eta_i\right] = \mathbb{E}_\pi\left[\beta_i\left(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\right)^\top \varepsilon_i\right] = \mathbb{E}_\pi\left[\zeta_i\left(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\right)^\top \Delta_i\right] = 0$, due to the martingale constraints.

Therefore, taking the infimum over $\left(\rho, \lambda, \{\xi_i, \beta_i, \zeta_i\}_{i=1}^T\right) \in \Lambda$ and the supremum over $\pi \in \mathcal{D}_\delta$ in (A.9), we readily get:

$$\text{Val}(\tilde{D}) \geq \text{Val}(\tilde{P}) \tag{A.10}$$

Hence, problem (A.8) is a weak dual of (A.6).

Finally, since (A.7) holds for all $\eta, \varepsilon, \Delta$, and since the pointwise supremum of a family of continuous functions is lower semi-continuous, and hence, Borel measurable (Rudin, 1987), the dual problem (A.8) can be rewritten as:

$$\begin{aligned} \text{Val}(\tilde{D}) = \inf_{\substack{\lambda \geq 0 \\ \xi_i, \beta_i, \zeta_i}} \quad & \lambda\delta + \mathbb{E}_{\mathbb{P}_0}\Bigg[\sup_{\Delta}\Bigg\{\sum_{i=1}^T \Delta_i^\top c_i(\phi)\Delta_i + 2\sum_{i=1}^T \Delta_i^\top \left(c_i(\phi)\varepsilon_i - a_i(\phi)^\top \eta_i\right) - \lambda\sum_{i=1}^T \Delta_i^\top \Delta_i \\ & + \sum_{i=1}^T \xi_i\left(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\right)^\top \eta_i + \sum_{i=1}^T \beta_i\left(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\right)^\top \varepsilon_i \\ & + \sum_{i=1}^T \zeta_i\left(\{\eta_t, \varepsilon_t, \Delta_t\}_{t=1}^{i-1}\right)^\top \Delta_i\Bigg\}\Bigg] \end{aligned} \tag{A.11}$$

From now on, we will refer to (A.11) as the dual.

Finding an optimal primal-dual pair. Now, we will find an optimal primal-dual solution pair for $\tilde{P}$ and $\tilde{D}$. Since $c_i(\phi) \in \mathbb{R}^{d\times d}$ is positive semi-definite, all its eigenvalues are non-negative and it admits a spectral decomposition $c_i(\phi) = Q_i\Sigma_i Q_i^\top$, where Q_i orthonormal, and $\Sigma_i \coloneqq \operatorname{diag}(\sigma_{i,1},\dots,\sigma_{i,d})$. Then, setting $\sigma_* \coloneqq \max_{i,j}\sigma_{i,j}$ to be the maximum eigenvalue of all $c_i(\phi)$'s for $i=1,\dots,T$, we have that for $\lambda > \sigma_*$, the matrices $c_i(\phi) - \lambda I$ are negative definite, $i=1,\dots,T$, and, hence, invertible. Therefore,

$$(c_i(\phi)-\lambda I)^{-1} = Q_i(\Sigma_i-\lambda I)^{-1}Q_i^\top = Q_i\operatorname{diag}\big((\sigma_{i,1}-\lambda)^{-1},\dots,(\sigma_{i,d}-\lambda)^{-1}\big)Q_i^\top \tag{A.12}$$

Setting $\tilde{\Delta}_{i,\phi} \coloneqq (c_i(\phi)-\lambda I)^{-1}\big(-c_i(\phi)\varepsilon_i + a_i(\phi)^\top\eta_i\big)$, we obtain that

$$\begin{aligned}\|\tilde{\Delta}_{i,\phi}\|_2^2 = \tilde{\Delta}_{i,\phi}^\top\tilde{\Delta}_{i,\phi} &= \big(-c_i(\phi)\varepsilon_i + a_i(\phi)^\top\eta_i\big)^\top Q_i(\Sigma_i-\lambda I)^{-1}Q_i^\top Q_i(\Sigma_i-\lambda I)^{-1}Q_i^\top\big(-c_i(\phi)\varepsilon_i + a_i(\phi)^\top\eta_i\big)\\ &= \big(-c_i(\phi)\varepsilon_i + a_i(\phi)^\top\eta_i\big)^\top Q_i(\Sigma_i-\lambda I)^{-2}Q_i^\top\big(-c_i(\phi)\varepsilon_i + a_i(\phi)^\top\eta_i\big)\\ &= \big(-c_i(\phi)\varepsilon_i + a_i(\phi)^\top\eta_i\big)^\top Q_i\operatorname{diag}\big((\sigma_{i,1}-\lambda)^{-2},\dots,(\sigma_{i,d}-\lambda)^{-2}\big)Q_i^\top\big(-c_i(\phi)\varepsilon_i + a_i(\phi)^\top\eta_i\big)\end{aligned} \tag{A.13}$$

where we used that $Q_i^\top Q_i = I_d$, since Q_i orthonormal. Letting $w_i = (w_{i,1},\dots,w_{i,d})^\top$ be defined as:

$$w_i \coloneqq Q_i^\top\big(-c_i(\phi)\varepsilon_i + a_i(\phi)^\top\eta_i\big) \tag{A.14}$$

we obtain

$$\|\tilde{\Delta}_{i,\phi}\|_2^2 = \sum_{j=1}^d w_{i,j}^2(\sigma_{i,j}-\lambda)^{-2} \tag{A.15}$$

Claim 1. *For each j such that $\sigma_{i,j} > 0$, we have*

$$\mathbb{E}_{\mathbb{P}_0}[w_{i,j}^2] \neq 0$$

where w_i is defined in (A.14).

Proof of Claim 1: Let j such that $\sigma_{i,j} > 0$. By the definition of w_i, we have that

$$\begin{aligned}w_i &= Q_i^\top\big(-c_i(\phi)\varepsilon_i + a_i(\phi)^\top\eta_i\big)\\ &= -Q_i^\top c_i(\phi)\varepsilon_i + Q_i^\top a_i(\phi)^\top\eta_i\\ &= -Q_i^\top Q_i\Sigma_i Q_i^\top\varepsilon_i + Q_i^\top a_i(\phi)^\top\eta_i\\ &= -\Sigma_i Q_i^\top\varepsilon_i + Q_i^\top a_i(\phi)^\top\eta_i\end{aligned} \tag{A.16}$$

Hence, $w_{i,j} = -\sigma_{i,j}Q_{i,j}^\top\varepsilon_i + Q_{i,j}^\top a_i(\phi)^\top\eta_i$, where $Q_{i,j}$ is the j-th row of the matrix $Q_i^\top$. Since Q_i orthonormal, we readily get that $Q_{i,j}$ is not the zero vector. If $\mathbb{E}_{\mathbb{P}_0}[w_{i,j}^2] = 0$, it would hold that $w_{i,j} = 0$, $\mathbb{P}_0$-a.s., and so:

$$\sigma_{i,j}Q_{i,j}^\top\varepsilon_i = Q_{i,j}^\top a_i(\phi)^\top\eta_i \quad \mathbb{P}_0\text{ - a.s.} \tag{A.17}$$

However, ε_i cannot be perpendicular to $Q_{i,j}$ a.s. under $\mathbb{P}_0$, since this would mean that ε_i lives in a lower-dimensional subspace, but the covariance matrix of ε_i is assumed to have full rank. Therefore, (A.17) would imply that there is linear relation between ε_i and η_i. But, since the covariance matrix of (X,Y) under $\mathbb{P}_0$ is assumed to have full rank, and $[X^\top\ Y^\top]^\top = \mathbb{A}[\eta^\top\ \varepsilon^\top]^\top$ for some $\mathbb{A}$ square matrix, this implies that the covariance matrix of (η,ε) is also full rank. Hence, (A.17) leads to a contradiction, and, so, we reach the conclusion. ♦

Then, we have that:

$$\sum_{i=1}^T\mathbb{E}_\pi\Big[\|\tilde{\Delta}_{i,\phi}\|_2^2\Big] = \sum_{i=1}^T\sum_{j=1}^d\mathbb{E}_{\mathbb{P}_0}\big[w_{i,j}^2\big](\sigma_{i,j}-\lambda)^{-2} \tag{A.18}$$

Setting:

$$G(\lambda) := \sum_{i=1}^{T}\sum_{j=1}^{d} \mathbb{E}_{\mathbb{P}_0}\left[w_{i,j}^2\right](\sigma_{i,j} - \lambda)^{-2} \tag{A.19}$$

and, since ε_i, η_i have finite second-moment for all i, we get that:

- $G(\lambda) \to +\infty$, as $\lambda \to \sigma_*$
- For $\lambda > \sigma_*$, $G(\lambda)$ is continuously decreasing with $G(\lambda) \to 0$ as $\lambda \to +\infty$

Hence, for $\lambda > \sigma_*$, there is a unique solution $\tilde{\lambda}$ of the equation $G(\lambda) = \delta$, i.e.,

$$\sum_{i=1}^{T}\sum_{j=1}^{d} \mathbb{E}_{\mathbb{P}_0}\left[w_{i,j}^2\right](\sigma_{i,j} - \tilde{\lambda})^{-2} = \delta \tag{A.20}$$

Then, we claim that the primal-dual solution pair:

(I) $\tilde{\Delta}_{i,\phi} := (c_i(\phi) - \tilde{\lambda} I)^{-1}\big(-c_i(\phi)\varepsilon_i + a_i(\phi)^\top \eta_i\big)$ for $i = 1, \dots, T$.

(II) $\tilde{\lambda} :=$ the unique solution of (A.20) greater than σ_*, and $\tilde{\xi}_i(\cdot) = \tilde{\beta}_i(\cdot) = \tilde{\zeta}_i(\cdot) = 0$ for $i = 1, \dots, T$.

with primal-dual objective values $\tilde{p}$ and $\tilde{d}$ of $\tilde{P}$ and $\tilde{D}$, respectively, is optimal. Indeed, the solution pair is feasible for the primal and dual problem, respectively. Moreover, the maximization problem inside the expectation in (A.11) for $\lambda = \tilde{\lambda}$, and $\tilde{\xi}_i(\cdot) = \tilde{\beta}_i(\cdot) = \tilde{\zeta}_i(\cdot) = 0$ for $i = 1, \dots, T$ becomes:

$$\sup_{\Delta} \left\{ \sum_{i=1}^{T} \Delta_i^\top (c_i(\phi) - \tilde{\lambda} I)\Delta_i + 2\sum_{i=1}^{T} \Delta_i^\top \big(c_i(\phi)\varepsilon_i - a_i(\phi)^\top \eta_i\big) \right\} \tag{A.21}$$

Since $(c_i(\phi) - \tilde{\lambda} I)$ is negative definite for all $i = 1, \dots, T$ (because $\tilde{\lambda} > \sigma_*$, as argued before), we get that the objective function in (A.21) is strictly concave. Solving the first-order optimality condition, we readily get that it is maximized for $\Delta_i = \tilde{\Delta}_{i,\phi}$, where $\tilde{\Delta}_{i,\phi}$ is defined in (I).

Therefore, the objective value, $\tilde{d}$, of the dual problem (A.11) for the assignment of the dual variables as per (II), becomes:

$$\tilde{d} = \tilde{\lambda}\delta + \mathbb{E}_{\mathbb{P}_0}\left[\sum_{i=1}^{T} \tilde{\Delta}_{i,\phi}^\top c_i(\phi)\tilde{\Delta}_{i,\phi} + 2\sum_{i=1}^{T} \tilde{\Delta}_{i,\phi}\big(c_i(\phi)\varepsilon_i - a_i(\phi)^\top \eta_i\big) - \tilde{\lambda}\sum_{i=1}^{T} \|\tilde{\Delta}_{i,\phi}\|_2^2\right] \tag{A.22}$$

$$= \mathbb{E}_{\mathbb{P}_0}\left[\sum_{i=1}^{T} \tilde{\Delta}_{i,\phi}^\top c_i(\phi)\tilde{\Delta}_{i,\phi} + 2\sum_{i=1}^{T} \tilde{\Delta}_{i,\phi}\big(c_i(\phi)\varepsilon_i - a_i(\phi)^\top \eta_i\big)\right] \tag{A.23}$$

$$= \tilde{p} \tag{A.24}$$

where in (A.23) we used that $\sum_{i=1}^{T} \mathbb{E}_\pi[\|\tilde{\Delta}_{i,\phi}\|_2^2] = \delta$ by (A.20), and in (A.24) we invoked that this quantity equals to the objective value, $\tilde{p}$, of the primal problem (A.6) for the assignment of the primal variables as per (I). Hence, combining (A.24) with (A.10), we conclude that strong duality holds, i.e.,

$$\mathrm{Val}(\tilde{P}) = \mathrm{Val}(\tilde{D}) \tag{A.25}$$

and the coupling $\tilde{\pi}$, whose second marginal is the law of $(\eta, \varepsilon + \tilde{\Delta})$, is an optimal solution of (15) in the constraint set $\mathcal{D}_\delta$. This concludes our proof.

■

A.3 Proof of Lemma 2

Lemma 2. *The value function $V(\phi)$ is a convex finite-valued function.*

Proof. Let ϕ with real-valued entries. Then, we have for $\pi \in \mathcal{D}_\delta$:

$$\begin{aligned}\mathbb{E}_\pi\left[\|\tilde{X} - \phi\tilde{Y}\|_2^2\right] &\leq 2\,\mathbb{E}_\pi\left[\|\tilde{X}\|_2^2\right] + 2\,\mathbb{E}_\pi\left[\|\phi\tilde{Y}\|_2^2\right] \\ &\leq 2\,\mathbb{E}_\pi\left[\|\tilde{X}\|_2^2\right] + 2\|\phi\|_2^2\,\mathbb{E}_\pi\left[\|\tilde{Y}\|_2^2\right] \end{aligned} \tag{A.26}$$

Then, taking the supremum over $\pi \in \mathcal{D}_\delta$, we get

$$V(\phi) \leq \sup_{\pi\in\mathcal{D}_\delta} 2\,\mathbb{E}_\pi\left[\|\tilde{X}\|_2^2\right] + 2\|\phi\|_2^2\,\mathbb{E}_\pi\left[\|\tilde{Y}\|_2^2\right] < \infty \tag{A.27}$$

since, the second norms of both $\tilde{X}$ and $\tilde{Y}$ are uniformly bounded within $\mathcal{D}_\delta$ due to the Wasserstein constraint.

Now, for any π in the constraint set $\mathcal{D}_\delta$, the function

$$\phi \mapsto \sum_{n=1}^{T} \mathbb{E}_\pi\left[\|\tilde{X}_n - \sum_{j=1}^{n} \phi_{n,j}\tilde{Y}_j\|_2^2\right]$$

is convex. Taking the supremum over $\pi \in \mathcal{D}_\delta$ we readily get that $V(\phi)$ is a convex function as the pointwise supremum of convex functions Nesterov (2014). ■

A.4 Proof of Proposition 2

Proposition 2. *Let ϕ', and $\tilde{\pi}_{\phi'}$ the corresponding worst-case distribution in $\mathcal{D}_\delta$, as per Theorem 1. Then, defining*

$$g \coloneqq \nabla_\phi f(\phi, \tilde{\pi}_{\phi'})\,|_{\phi=\phi'} \tag{21}$$

it holds that $g \in \partial V(\phi')$.

Proof. By the definition of of $\tilde{\pi}_{\phi'}$, we obtain that

$$V(\phi') = \sup_{\pi\in\mathcal{D}_\delta} f(\phi', \pi) = f(\phi', \tilde{\pi}_{\phi'}) \tag{A.28}$$

Moreover, by the definition f, it is easy to see that $f(\cdot, \tilde{\pi}_{\phi'})$ is convex and differentiable, and letting $g \coloneqq \nabla_\phi f(\phi, \tilde{\pi}_{\phi'})\,|_{\phi=\phi'}$, we obtain:

$$f(\phi, \tilde{\pi}_{\phi'}) \geq f(\phi', \tilde{\pi}_{\phi'}) + \langle g, \phi - \phi'\rangle \tag{A.29}$$

Then,

$$\begin{aligned} V(\phi) \geq f(\phi, \tilde{\pi}_{\phi'}) &\overset{(A.29)}{\geq} f(\phi', \tilde{\pi}_{\phi'}) + \langle g, \phi - \phi'\rangle \\ &\overset{(A.28)}{=} V(\phi') + \langle g, \phi - \phi'\rangle \end{aligned} \tag{A.30}$$

i.e., $g \in \partial V(\phi')$. ■