Fairness in Forecasting and Learning Linear Dynamical Systems

Quan Zhou,[1] **Jakub Marecek,**[2] **Robert Shorten** [1,3]
[1] University College Dublin, Ireland
[2] Czech Technical University in Prague, the Czech Republic
[3] Imperial College London, UK
quan.zhou@ucdconnect.ie, jakub.marecek@fel.cvut.cz, r.shorten@imperial.ac.uk

Abstract

In machine learning, training data often capture the behaviour of multiple subgroups of some underlying human population. When the amounts of training data for the subgroups are not controlled carefully, under-representation bias arises. We introduce two natural notions of subgroup fairness and instantaneous fairness to address such under-representation bias in time-series forecasting problems. In particular, we consider the subgroup-fair and instant-fair learning of a linear dynamical system (LDS) from multiple trajectories of varying lengths and the associated forecasting problems. We provide globally convergent methods for the learning problems using hierarchies of convexifications of non-commutative polynomial optimisation problems. Our empirical results on a biased data set motivated by insurance applications and the well-known COMPAS data set demonstrate both the beneficial impact of fairness considerations on statistical performance and the encouraging effects of exploiting sparsity on run time.

Introduction

The identification of vector autoregressive processes with hidden components from time series of observations is a central problem across Machine Learning, Statistics, and Forecasting (West and Harrison 1997). This problem is also known as proper learning of linear dynamical systems (LDS) in System Identification (Ljung 1998). As a rather general approach to time-series analysis, it has applications ranging from learning population-growth models in actuarial science and mathematical biology to functional analysis in neuroscience. Indeed, one encounters either partially observable processes (Åström 1965) or questions of causality (Pearl 2009) that can be tied to proper learning of LDS (Geiger et al. 2015) in almost any application domain.

A discrete-time model of a linear dynamical system $\mathcal{L} = (G, F, V, W)$ (West and Harrison 1997) suggests that the random variable $Y_t \in \mathbb{R}^m$ capturing the observed component (output, observations, measurements) evolves over time $t \geq 1$ according to:

$$\phi_t = G\phi_{t-1} + w_t, \tag{1}$$

$$Y_t = F'\phi_t + v_t, \tag{2}$$

where $\phi_t \in \mathbb{R}^n$ is the hidden component (state) and $G \in \mathbb{R}^{n\times n}$ and $F \in \mathbb{R}^{n\times m}$ are compatible system matrices. Random variables w_t, v_t capture normally-distributed process noise and observation noise, with zero means and covariance matrices $W \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$, respectively. In this setting, proper learning refers to identifying the quadruple (G, F, V, W) given the observations $\{Y_t\}_{t\in\mathbb{N}}$ of $\mathcal{L}$. This also allows for the estimation of subsequent observations, in the so-called "prediction-error" approach to improper learning (Ljung 1998).

We consider a generalisation of the proper learning of LDS, where:

- There are a number of individuals $p \in \mathcal{P}$ within a population. The population $\mathcal{P}$ is partitioned into subgroups indexed by $\mathcal{S}$.
- For each subgroup $s \in \mathcal{S}$, there is a set $\mathcal{I}^{(s)}$ of trajectories of observations available and each trajectory $i \in \mathcal{I}^{(s)}$ has observations for periods $\mathcal{T}^{(i,s)}$, possibly of varying cardinality $|\mathcal{T}^{(i,s)}|$.
- Each subgroup $s \in \mathcal{S}$ is associated with a LDS, $\mathcal{L}^{(s)}$. For all $i \in \mathcal{I}^{(s)}$, $s \in \mathcal{S}$, the trajectory $\{Y_t\}^{(i,s)}$, for $t \in \mathcal{T}^{(i,s)}$, is hence generated by precisely one LDS $\mathcal{L}^{(s)}$.

Note that in our notation, the superscripts distinguish the trajectories and subgroups, while subscripts indicate the periods. In this setting, under-representation bias (Blum and Stangl 2019, cf. Section 2.2), where the trajectories of observations from one ("disadvantaged") subgroup are under-represented in the training data, harms both accuracy of the classifier overall and fairness in the sense of varying accuracy across the subgroups. This is particularly important if the problem is constrained to be subgroup-blind, i.e., constrained to consider only a single LDS as a model. This is the case when the use of attributes distinguishing each subgroup can be regarded as discriminatory (e.g., gender, race, cf. (Gajane and Pechenizkiy 2018)). Notice that such anti-discrimination measures are increasingly stipulated by legal systems, e.g., within product or insurance pricing, where the sex of the applicant cannot be used, despite being known.

A natural notion of fairness in subgroup-blind learning of LDS involves estimating the system matrices or forecasting the next output of a single LDS that captures the overall behaviour across all subgroups, while taking into account the

varying amounts of training data for the individual subgroups. To formalise this, suppose that we learn one LDS $\mathcal{L}$ from the multiple trajectories and we define a loss function that measures the loss of accuracy for a certain observation $Y_t^{(i,s)}$, for $t \in \mathcal{T}^{(i,s)}$, $i \in \mathcal{I}^{(s)}$, $s \in \mathcal{S}$ when adopting the forecast f_t for the overall population. For $t \in \mathcal{T}^{(i,s)}$, $i \in \mathcal{I}^{(s)}$, $s \in \mathcal{S}$, we have

$$\overset{(i,s)}{\text{loss}}(f_t) := ||Y_t^{(i,s)} - f_t||. \tag{3}$$

Let $\mathcal{T}^+ = \cup_{i \in \mathcal{I}^{(s)}, s \in \mathcal{S}} \mathcal{T}^{(i,s)}$. We know that f_t is feasible only when $t \in \mathcal{T}^+$. Note that since each trajectory is of varying length, it is possible that for a certain triple (t, i, s), there is no observation $Y_t^{(i,s)}$.

We propose two objectives to address the under-representation bias, which extend group fairness (Feldman et al. 2015) to time series:

1. **Subgroup Fairness**. The objective seeks to equalise, across all subgroups, the sum of losses for the subgroup. Considering the number of trajectories in each subgroup and the number of observations across the trajectories may differ, we include $|\mathcal{I}^{(s)}|, |\mathcal{T}^{(i,s)}|$ as weights:

$$\min_{f_t, t \in \mathcal{T}^+} \max_{s \in \mathcal{S}} \left\{ \frac{1}{|\mathcal{I}^{(s)}|} \sum_{i \in \mathcal{I}^{(s)}} \frac{1}{|\mathcal{T}^{(i,s)}|} \sum_{t \in \mathcal{T}^{(i,s)}} \overset{(i,s)}{\text{loss}}(f_t) \right\} \tag{4}$$

2. **Instantaneous Fairness**. The objective seeks to equalise the instantaneous loss, by minimising the maximum of the losses across all subgroups and all times:

$$\min_{f_t, t \in \mathcal{T}^+} \left\{ \max_{t \in \mathcal{T}^{(i,s)}, i \in \mathcal{I}^{(s)}, s \in \mathcal{S}} \left\{ \overset{(i,s)}{\text{loss}}(f_t) \right\} \right\} \tag{5}$$

Following (Zhou and Marecek 2020), we also cast the proper and improper learning of a linear dynamical system with such fairness considerations as a non-commutative polynomial optimisation problem (NCPOP), which can be solved efficiently using a globally-convergent hierarchy of semidefinite programming (SDP) relaxations.

Related Work

This presents an algorithmic approach to addressing the under-representation bias studied by (Blum and Stangl 2019) and within the imbalanced learning literature (He and Ma 2013; Brabec et al. 2020, e.g.) and presents a step forward within the fairness in forecasting studied recently by (Gajane and Pechenizkiy 2018; Chouldechova 2017; Locatello et al. 2019), as outlined in the excellent survey of (Chouldechova and Roth 2020; Barocas, Hardt, and Narayanan 2019). It follows much work on fairness in classification, e.g., (Zliobaite 2015; Hardt, Price, and Srebro 2016; Kilbertus et al. 2017; Kusner et al. 2017; Chouldechova and Roth 2020; Aghaei, Azizi, and Vayanos 2019). It is complemented by several recent studies involving dynamics and fairness (Mouzannar, Ohannessian, and Srebro 2019; Paaßen et al. 2019; Jung et al. 2020), albeit not involving *learning* of dynamics. It relies crucially on tools developed in non-commutative polynomial optimisation (Pironio, Navascués, and Acín 2010; Wang, Magron, and Lasserre 2019, 2020) and non-commutative algebra (Gelfand and Neumark 1943; Segal 1947; McCullough 2001; Helton 2002), which have not seen much use in Statistics and Machine Learning, yet.

Fairness in Machine Learning

The last two years have seen an unprecedented explosion in attention of fairness in the field of artificial intelligence and machine learning (Chouldechova and Roth 2020). The widely used criminal risk assessment tool, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) was found to favour white defendants over black defendants by under-predicting recidivism for white and over-predicting recidivism for black defendants (Angwin et al. 2016; Dressel and Farid 2018).

There have been several candidate definitions of fairness. At the first stage, a naive notion "fairness under unawareness" is used. The shortcoming is that there are some features, used to predict but related to protected attributes. Therefore, the predictor from unawareness could still be related to protected attributes even though protected attributes are not given. "Demographic parity", proposed by (Calder, Malthouse, and Schaedel 2009), requires the proportion of each segment of a protected class (e.g. gender) should receive the positive outcome at equal rates. But it might be unfair in the case of unbalanced distributions of features between advantaged and disadvantaged subgroups even in the absent of biases. The notions of "equal opportunity" in (Hardt, Price, and Srebro 2016) and "counterfactual fairness " of (Kusner et al. 2017) require the predictor to be unrelated to protected attributes. In other words, they focus more on accurate prediction of the unbalanced distribution without discrimination as a predictor is very unlikely to be discriminatory if it only reflects the real outcomes.

Group fairness only provides an average guarantee for the individuals in a protected group (Awasthi et al. 2020) and is insufficient by itself. Sometimes even the notions of group fairness is maintained, from the view of an individual, the outcome is unfair. The individual definition asks for constraints that bind on specific pairs of individuals, rather than on a quantity that is averaged over groups (Chouldechova and Roth 2020), or in other words, it requires "similar individuals should be treated similarly" (Dwork et al. 2012). However, this notion requires a similarity metric capturing the ground truth, which requires general and task-specific assumption on its definition (Sharifi-Malvajerdi, Kearns, and Roth 2019).

As the increasingly many notions of model fairness arise, it is necessary to build up a comprehensive framework of multiple fairness criteria. For instance, the fairness resoultion model, proposed in (Awasthi et al. 2020), is guided by the unfairness complaints received by the system and it could be a more practical way to maintain both group and individual fairness.

Motivation

Insurance Pricing Let us consider two motivating examples. One important application arises in Actuarial Science.

In the European Union, a directive (implementing the principle of equal treatment between men and women in the access to and supply of goods and services), bars insurers from using gender as a factor in justifying differences in individuals' premiums. In contrast, insurers in many other territories classify insureds by gender, because females and males have different behavior patterns, which affects insurance payments. Take the annuity-benefit scheme for example. It is a well-known fact that females have a longer life expectancy than males (Huang and Salm 2020). The insurer will hence pay more to a female insured during her lifetime, compared to a male insured, on averag (Thiery and Van Schoubroeck 2006). Because of the directive, a unisex mortality table needs to be used. As a result, male insureds receive less benefits, while paying the same premium in total as the female subgroup (Thiery and Van Schoubroeck 2006). Consequently, male insureds might leave the annuity-benefit scheme (known as adverse selection), which makes the unisex mortality table more challenging to use in the estimation of the life expectancy of the "unisex" population, where female insureds become the advantaged subgroup.

Consider a simple actuarial pricing model of annuity insurance. Insureds enter an annuity-benefit scheme at time 0 and each insured can receive 1 euro at the end of each year for at most 10 years on the condition that it is still alive. Let p_t denotes how many insureds left in the scheme in the end of the t^{th} year. Suppose there are p_0 insureds in the beginning and the pricing interest rate is i ($i \leq 1$). The formula of calculating the pure premium is in (6), thus summing up the present values of payment in each year and then divided by the number of insureds in the beginning.

$$\text{premium} := \frac{\sum_{t=1}^{10} p_t \times (1+i)^{-t}}{p_0} \tag{6}$$

The most important quality p_t is derived from estimating insureds' life expectancy. Suppose the insureds can be divided into female and male subgroups. Each subgroup has one trajectory: $\{Y_t\}^{(\cdot,f)}$ for female subgroup, $\{Y_t\}^{(\cdot,m)}$ for male subgroup for $1 \leq t \leq 10$, where the superscript i is dropped. The two trajectories indicate how many female and male insureds are alive at the end of the t^{th} year, respectively. Both trajectories can be regarded as linear dynamic systems. We have

$$Y_t^{(\cdot,f)} = G^{(f)} Y_{t-1}^{(\cdot,f)} + \omega_t^{(f)}, \ 2 \leq t \leq 10, \tag{7}$$

$$Y_t^{(\cdot,m)} = G^{(m)} Y_{t-1}^{(\cdot,m)} + \omega_t^{(m)}, 2 \leq t \leq 10, \tag{8}$$

$$p_t = \ Y_t^{(\cdot,f)} + Y_t^{(\cdot,m)}, \quad 1 \leq t \leq 10, \tag{9}$$

where $\omega_t^{(f)}$ and $\omega_t^{(m)}$ are measurement noises while $G^{(f)}$ and $G^{(m)}$ are system matrices for female LDS $\mathcal{L}^{(f)}$ and male LDS $\mathcal{L}^{(m)}$ respectively. Note that these are state processes, without any observation process: the number of survivals can be precisely observed. To satisfy the directive, one needs to consider a unisex model:

$$f_t = G f_{t-1} + \omega_t, \ 2 \leq t \leq 10, \tag{10}$$

where $2 \leq t \leq 10$ and ω_t and G pertain to the unisex insureds LDS $\mathcal{L}$. Subsequently, the loss functions for female (f) and male (m) subgroups are:

$$\underset{\text{loss}}{^{(\cdot,f)}}(f_t) := ||Y_t^{(\cdot,f)} - f_t|| \ , 1 \leq t \leq 10, \tag{11}$$

$$\underset{\text{loss}}{^{(\cdot,m)}}(f_t) := ||Y_t^{(\cdot,m)} - f_t|| \ , 1 \leq t \leq 10, \tag{12}$$

Since the trajectories $\{Y_t\}^{(\cdot,f)}$ and $\{Y_t\}^{(\cdot,m)}$ have the same length and there is only one trajectory in each subgroup, the two objective (4)-(5) has the form:

$$\min_{f_t, 1 \leq t \leq 10} \max \left\{ \sum_{t=1}^{10} \underset{\text{loss}}{^{(\cdot,f)}}(f_t), \sum_{t=1}^{10} \underset{\text{loss}}{^{(\cdot,m)}}(f_t) \right\} \tag{13}$$

$$\min_{f_t, 1 \leq t \leq 10} \left\{ \max_{1 \leq t \leq 10, s \in \{f,m\}} \left\{ \underset{\text{loss}}{^{(\cdot,s)}}(f_t) \right\} \right\} \tag{14}$$

Personalised Pricing Another application arises in personalised pricing (PP). For example, Amazon has been found (OECD 2018) to sell certain products to regular consumers at higher prices. This is legal, albeit questionable. In contrast, gender-based price discrimination (Abdou 2019) violates (OECD 2018) anti-discrimination laws in many jurisdictions.

Let us consider an idealised example of PP: Consider a soap retailer, whose customers contain female and male subgroups. Each gender has a specific dynamic system modelling its willing to pay ("demand price" of each subgroup), while the retailer should set a "unisex" price. As in the discussion of insurance pricing, we consider subgroups $S = \{\text{female}, \text{male}\}$ and use superscripts $(f), (m)$ to distinguish the related quantities. Unlike in insurance pricing, the demand price of each customer is regarded as a single trajectory. More importantly, since customers might start buying soap, quit buying the soap, or move to other substitutes at different time points, those trajectories of demand prices are assumed to be of varying lengths. For example, a customer starts to buy the soap at time 3 but decides to buy hand wash instead from time 7.

Let us assume there are $|\mathcal{I}^{(f)}|$ female customers and $|\mathcal{I}^{(m)}|$ customers in the overall time window $\mathcal{T}^+$. Let $Y_t^{(i,s)}$ denote the estimated demand price at time t of the i^{th} customer in subgroup s. These evolve as:

$$\phi_t^f = \ G^{(f)} \phi_{t-1}^{(f)} + \omega_t^{(f)} \quad , t \in \mathcal{T}^+, \tag{15}$$

$$Y_t^{(i,f)} = \ F^{(f)'} \phi_t^{(f)} + \nu_t^{(i,f)} \quad , t \in \mathcal{T}^{(i,f)}, i \in \mathcal{I}^{(f)}, \tag{16}$$

$$\phi_t^m = \ G^{(m)} \phi_{t-1}^{(m)} + \omega_t^{(m)} \quad , t \in \mathcal{T}^+, \tag{17}$$

$$Y_t^{(i,m)} = F^{(m)'} \phi_t^{(m)} + \nu_t^{(i,m)} \ , t \in \mathcal{T}^{(i,m)}, i \in \mathcal{I}^{(m)} \tag{18}$$

The unisex model for demand price considers the unisex state m_t, the unisex system matrices G, F, and unisex noises ω_t, ν_t:

$$m_t = G m_{t-1} + \omega_t \ , \ t \in \mathcal{T}^+, \tag{19}$$

$$f_t = \ F' m_t + \nu_t \quad , \ t \in \mathcal{T}^+. \tag{20}$$

For $\text{loss}^{(i,f)}(f_t) := ||Y_t^{(i,f)} - f_t||, t \in \mathcal{T}^{(i,f)}, i \in \mathcal{I}^{(f)}$ and $\text{loss}^{(i,m)}(f_t) := ||Y_t^{(i,m)} - f_t||, t \in \mathcal{T}^{(i,m)}, i \in \mathcal{I}^{(m)}$, the two objectives (4)-(5) have the form:

$$\min_{f_t, t\in\mathcal{T}^+} \max_{s\in\mathcal{S}} \left\{ \frac{1}{|\mathcal{I}^{(s)}|} \sum_{i=1}^{\mathcal{I}^{(s)}} \frac{1}{|\mathcal{T}^{(i,s)}|} \sum_{t\in\mathcal{T}^{(i,s)}} \overset{(i,s)}{\text{loss}}(f_t) \right\} \quad (21)$$

$$\min_{f_t, t\in\mathcal{T}^+} \left\{ \max_{t\in\mathcal{T}^{(i,s)}, i\in\mathcal{I}_s, s\in\mathcal{S}} \left\{ \overset{(i,s)}{\text{loss}}(f_t) \right\} \right\} \quad (22)$$

Our Models

We assume that the underlying LDS $\mathcal{L}^{(s)} = (G^{(s)}, F^{(s)}, V^{(s)}, W^{(s)})$ of each subgroup $s \in \mathcal{S}$ all have the form of (1)-(2), while only one subgroup-blind LDS $\mathcal{L}$ can be learned and used for prediction. The following model in (23)-(24) can be used to describe the subgroup-blind state evolution directly.

$$m_t = Gm_{t-1} + \omega_t, \quad (23)$$

$$f_t = F'm_t + \nu_t. \quad (24)$$

for $t \in \mathcal{T}^+$, where m_t represents the estimated subgroup-blind state and $\{f_t\}_{t\in\mathcal{T}^+}$ is the trajectory predicted by the subgroup-blind LDS $\mathcal{L}$.

The objectives (4) and (5), subject to (23)-(24), yield two operator-valued optimisation problems. Their inputs are $Y_t^{(i,s)}, t \in \mathcal{T}^{(i,s)}, i \in \mathcal{I}^{(s)}, s \in \mathcal{S}$, i.e., the observations of multiple trajectories and the multiplier λ. The operator-valued decision variables $\mathcal{O}$ include operators proper F, G, vectors m_t, ω_t, and scalars f_t, ν_t, and z. Notice that t ranges over $t \in \mathcal{T}^+$, except for m_t, where $t \in \mathcal{T}^+ \cup \{0\}$. The auxiliary scalar variable z is used to reformulate "$\max$" in the objective (4) or (5). Since the observation noise is assumed to be a sample of mean-zero normally-distributed random variable, we add the sum of squares of ν_t to the objective with a multiplier λ, seeking a solution with ν_t close to zero. Overall, the subgroup-fair and instant-fair formulations read:

$$\begin{aligned} \min_{\mathcal{O}} \quad & z + \lambda \sum_{t\geq 1} \nu_t^2 && \text{Subgroup-Fair} \\ \text{s.t.} \quad & z \geq \frac{1}{|\mathcal{I}^{(s)}|} \sum_{i\in\mathcal{I}^{(s)}} \frac{1}{|\mathcal{T}^{(i,s)}|} \sum_{t\in\mathcal{T}^{(i,s)}} \overset{(i,s)}{\text{loss}}(f_t), s \in \mathcal{S}, \\ & m_t = Gm_{t-1} + \omega_t && , t \in \mathcal{T}^+, \\ & f_t = F'm_t + \nu_t && , t \in \mathcal{T}^+. \end{aligned} \quad (25)$$

$$\begin{aligned} \min_{\mathcal{O}} \quad & z + \lambda \sum_{t\geq 1} \nu_t^2 && \text{Instant-Fair} \\ \text{s.t.} \quad & z \geq \overset{(i,s)}{\text{loss}}(f_t) && , t \in \mathcal{T}^{(i,s)}, i \in \mathcal{I}^{(s)}, s \in \mathcal{S}, \\ & m_t = Gm_{t-1} + \omega_t, t \in \mathcal{T}^+, \\ & f_t = F'm_t + \nu_t && , t \in \mathcal{T}^+. \end{aligned} \quad (26)$$

For comparison, we use a traditional formulation that focuses on minimising the overall loss:

$$\begin{aligned} \min_{\mathcal{O}} \quad & \sum_{s\in\mathcal{S}} \sum_{i\in\mathcal{I}^{(s)}} \sum_{t\in\mathcal{T}^{(i,s)}} \overset{(i,s)}{\text{loss}}(f_t) + \lambda \sum_{t\geq 1} \nu_t^2 && \text{Unfair} \\ \text{s.t.} \quad & m_t = Gm_{t-1} + \omega_t, t \in \mathcal{T}^+, \\ & f_t = F'm_t + \nu_t && , t \in \mathcal{T}^+. \end{aligned} \quad (27)$$

To state our main result, we need a technical assumption related to the stability of the LDS, which suggests that the operator-valued decision variables (and hence estimates of states and observations) remain bounded. Let us define the quadratic module following (Pironio, Navascués, and Acín 2010). Let $Q = \{q_i\}$ be the set of polynomials determining the constraints. The positivity domain $\mathbf{S}_Q$ of Q are tuples $X = (X_1, \ldots, X_n)$ of bounded operators on a Hilbert space $\mathcal{H}$ making all $q_i(X)$ positive semidefinite. The quadratic module M_Q is the set of $\sum_i f_i^\dagger f_i + \sum_i \sum_j g_{ij}^\dagger q_i g_{ij}$ where f_i and g_{ij} are polynomials from the same ring. As in (Pironio, Navascués, and Acín 2010), we assume:

Assumption 1 (Archimedean)**.** *Quadratic module M_Q of* (25) *is Archimedean, i.e., there exists a real constant C such that $C^2 - (X_1^\dagger X_1 + \cdots + X_{2n}^\dagger X_{2n}) \in M_Q$.*

Our main result shows that it is possible to recover the quadruple (G, F, V, W) of the subgroup-blind $\mathcal{L}$ with guarantees of global convergence:

Theorem 2. *For any observable linear system $\mathcal{L} = (G, F, V, W)$, for any length $\mathcal{T}^+$ of a time window, and any error $\epsilon > 0$, under Assumption 1, there is a convex optimisation problem from whose solution one can extract the best possible estimate of system matrices of a system $\mathcal{L}$ based on the observations, with fairness subgroup-fair considerations* (25)*, up to an error of at most ϵ in Frobenius norm. Furthermore, with suitably modified assumptions, the result holds also for the instant-fair considerations* (26)*.*

The proof is available in the full version of the paper online (Zhou, Marecek, and Shorten 2020). It relies on the work of (Pironio, Navascués, and Acín 2010), which shows the existence of a sequence of convex optimisation problems, whose objective function approaches the optimum of the non-commutative polynomial optimisation problem, and on the work of Gelfand, Naimark, and Segal (Gelfand and Neumark 1943; Segal 1947; Klep, Povh, and Volcic 2018), which makes it possible to extract the minimiser of the non-commutative polynomial optimisation problem from the solution of the convex optimisation problem.

Numerical Illustrations

Generation of Biased Training Data

To illustrate the impact of our models on data with varying degrees of under-representation bias, we consider a method for generating data resembling the motivating applications in Section , with varying degrees of the bias. Suppose there is one advantaged subgroup and one disadvantaged subgroup, $S = \{\text{advantaged, disadvantaged}\}$ with trajectories $\mathcal{I}^{(a)}$ and $\mathcal{I}^{(d)}$ in each subgroup. Under-representation bias can enter the training set in the following steps:

1. Observations $Y_t^{(i,s)}$ are sampled from corresponding LDS $\mathcal{L}^{(s)}$. Thus each $Y_t^{(i,s)} \sim \mathcal{L}^{(s)}$.
2. Discard some trajectories in $\mathcal{I}^{(d)}$, if necessary, such that $|\mathcal{I}^{(a)}| \geq |\mathcal{I}^{(d)}|$.
3. Let $\beta^{(s)}, s \in \mathcal{S}$ denote the probability that an observation from subgroup s stays in the training data and $0 \leq \beta^{(s)} \leq 1$. Discard more observations of $\mathcal{I}^{(d)}$ than those of $\mathcal{I}^{(a)}$ so that $\beta^{(a)} \geq \beta^{(d)}$. If $\beta^{(a)}$ is fixed at 1, it can be seen as the ratio of the number of observations in disadvantaged subgroup to that of advantaged subgroup and the degree of under-representation bias can be controlled by simply adjusting $\beta^{(d)}$.

The last two steps discard more observations of the disadvantaged subgroup in the biased training data, so that the advantaged subgroup becomes over-represented. Note that for a small sample size, it is necessary to make sure there is at least one observation in each subgroup at each period.

Consider that the LDS for both subgroups $\mathcal{L}^{(s)}, s \in \mathcal{S}$ have the same system matrices:

$$G^{(s)} = \begin{bmatrix} 0.99 & 0 \\ 1.0 & 0.2 \end{bmatrix}, F^{(s)} = \begin{bmatrix} 1.1 \\ 0.8 \end{bmatrix},$$

while the covariance matrices $V^{(s)}, W^{(s)}, s \in \mathcal{S}$ are sampled randomly from a uniform distribution over $[0, 1)$ and $[0, 0.1)$, respectively. The initial states $m_0^{(s)}$ of each subgroups are 5 and 7. We set the time window to be 20 across 3 trajectories in the advantaged subgroup and 2 in disadvantaged one, i.e., $|\mathcal{T}^+| = 20$, $|\mathcal{I}^{(a)}| = 3$ and $|\mathcal{I}^{(d)}| = 2$. Then the bias is introduced according to the biased training data generalisation process described above, with random $\beta^{(s)}, s \in \mathcal{S}$.

Figure 1 shows the forecasts in 10 experiments on this example. For each experiment, the same set of observations $Y_t^{(i,s)}, t \in \mathcal{T}^{(i,s)}, i \in \mathcal{I}^{(s)}, s \in \mathcal{S}$ is reused and the trajectories of advantaged and disadvantaged subgroups are denoted by dotted lines and dashed lines, respectively. However, the observations that are discarded vary across the experiments. Thus, a new biased training set is generated in each experiment, albeit based on the same "ground set" of observations. The three models (25)-(27) are applied in each experiment with λ of 1, 3, and 5, respectively, as chosen by iterating over integers 1 to 10. The mean of forecast f and its standard deviation are displayed as solid curves with error bands.

Effects of Under-Representation Bias on Accuracy

Figure 2 suggests how the degree of bias affects accuracy with and without considering fairness. With the number of trajectories in both subgroups set to 2, i.e., $|I_a| = |I_d| = 2$ and $\beta^{(a)} = 1$, we vary the degree of bias by adjusting $\beta^{(d)}$ within the range of $[0.5, 0.9]$. To measure the effect of the degree on accuracy, we introduce the normalised root mean square error (nrmse) fitness value for each subgroup $s \in \mathcal{S}$:

$$\overset{(s)}{\text{nrmse}} := \sqrt{\frac{\sum_{i \in \mathcal{I}^{(s)}} \sum_{t \in \mathcal{T}^{(i,s)}} \left(Y_t^{(i,s)} - f_t\right)^2}{\sum_{i \in \mathcal{I}^{(s)}} \sum_{t \in \mathcal{T}^{(i,s)}} \left(Y_t^{(i,s)} - \text{mean}^{(s)}\right)^2}}, \tag{28}$$

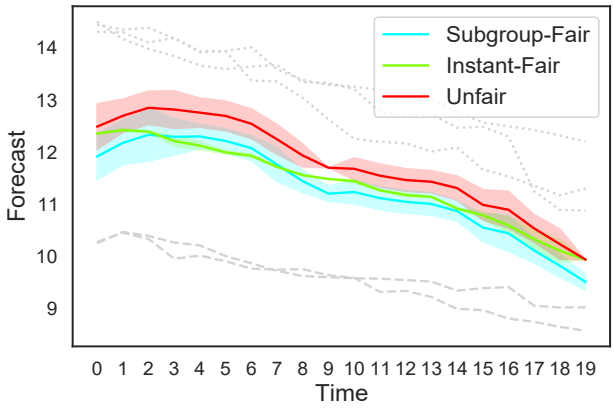

Figure 1: Forecast obtained using (25)-(27): the solid lines in primary colours with error bands display the mean and standard deviation of the forecasts over 10 experiments. For reference, dotted lines and dashed lines in grey denote the trajectories of observations of advantaged and disadvantaged subgroups, respectively, before discarding any observations.

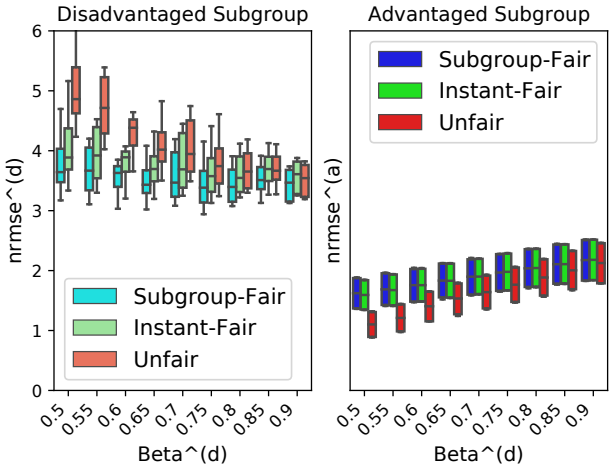

Figure 2: Accuracy as a function of the degree of under-representation bias: The boxplot of $\text{nrmse}^{(s)}, s \in \mathcal{S}$ against $\beta^{(d)}$, where $\beta^{(d)} = [0.5, 0.55, \ldots, 0.9]$, with boxes for the quartiles of $\text{nrmse}^{(s)}$ obtained from 5 experiments, using the observations in Figure 1.

where $\text{mean}^{(s)} := \frac{1}{|\mathcal{I}^{(s)}|} \sum_{i \in \mathcal{I}^{(s)}} \frac{1}{|\mathcal{T}^{(i,s)}|} \sum_{t \in \mathcal{T}^{(i,s)}} Y_t^{(i,s)}$.

Higher $\text{nrmse}^{(s)}$ indicates lower accuracy for subgroup s, i.e., the predicted trajectory of subgroup-blind $\mathcal{L}$ is further away from the subgroup.

For the training data, the same set of observations $Y_t^{(i,s)}, t \in \mathcal{T}^{(i,s)}, i \in \mathcal{I}^{(s)}, s \in \mathcal{S}$ in Figure 1 is reused but $|I_a| = |I_d| = 2$. Thus, one trajectory in the advantaged subgroup is discarded. Then, the biased training data generalisation process (described above) is applied in each experiment with $\beta^{(a)} = 1$ and the values for $\beta^{(d)}$ selecting from 0.5 to 0.9 at the step of 0.05. For each value of $\beta^{(d)}$, three models (25)-(27) are run with new biased training data and

the experiment is repeated for 5 times. Hence, the quartiles of $\mathrm{nrmse}^{(s)}$ for each subgroup are shown as boxes in Figure 2.

One could expect that nrmse fitness values of the advantaged subgroup in Figure 2 to be generally lower than those of the disadvantaged subgroup ($\mathrm{nrmse}^{(d)} \geq \mathrm{nrmse}^{(a)}$), leaving a gap. Those gaps narrow down as $\beta^{(d)}$ increases, simply because more observations of disadvantaged subgroup remain in the training data. Compared the to "Unfair", models with fairness constraints, i.e., "Subgroup-Fair" and "Instant-Fair", show narrower gaps and higher fairness between two subgroups. More surprisingly, when $\mathrm{nrmse}^{(a)}$ decreases as $\beta^{(d)}$ gets close to 0.5, "Subgroup-Fair" model still can keep the $\mathrm{nrmse}^{(d)}$ at almost the same level, indicating a rise in overall accuracy. This is in contrast with the results of (Zliobaite 2015; Dutta et al. 2020) in classification.

Run-Time

Notice that minimising multivariate operator-valued polynomial optimization problems (25)-(27) is non-trivial, but that there exists sparsity-exploiting variants (TSSOS) of the globally convergent Navascués-Pironio-Acín (NPA) hierarchy used in the proof of Theorem 2. See (Klep, Magron, and Povh 2019; Wang, Magron, and Lasserre 2019, 2020; Wang and Magron 2020). The SDP of a given order in the respective hierarchy can be constructed using `ncpol2sdpa` of (Wittek 2015) or the tools of (Wang and Magron 2020) and solved by `sdpa` of (Yamashita, Fujisawa, and Kojima 2003). Our implementation is available on-line at https://github.com/Quan-Zhou/Fairness-in-Learning-of-LDS.

In Figure 4, we illustrate the run-time and size of the relaxations as a function of the length of the time window with the same data set as above (i.e., Figure 1). The grey curve displays the number of variables in the first-order SDP relaxation of "Subgroup-Fair" and "Instant-Fair" models against the length of time window. The deep-pink and cornflower-blue curves show the run-time of the first-order SDP relaxation of NPA and the second-order SDP relaxation of TSSOS hierarchy, respectively, on a laptop equipped by Intel Core i7 8550U at 1.80 Ghz. The results of "Subgroup-Fair" and "Instant-Fair" models are presented by solid and dashed curves, respectively. Since each experiment is repeated for three times, the mean and mean $\pm$ 1 standard deviation of run-time are presented by curves with shaded error bands. It is clear that the run-time of TSSOS exhibits a modest growth with the length of time window, while that of the plain-vanilla NPA hierarchy grows much faster.

An Alternative Approach to COMPAS Dataset

Finally, we wish to suggest the broader applicability of the two notions of subgroup fairness and instantaneous fairness. We use the well-known dataset (Angwin et al. 2016) of estimates of the likelihood of recidivism made by the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), as used by courts in the United States. The dataset comprises of defendants' gender, race, age, charge degree, COMPAS recidivism scores, two-year recidivism label, as well as information on prior incidents. The two-year recidivism label denotes whether a person got rearrested

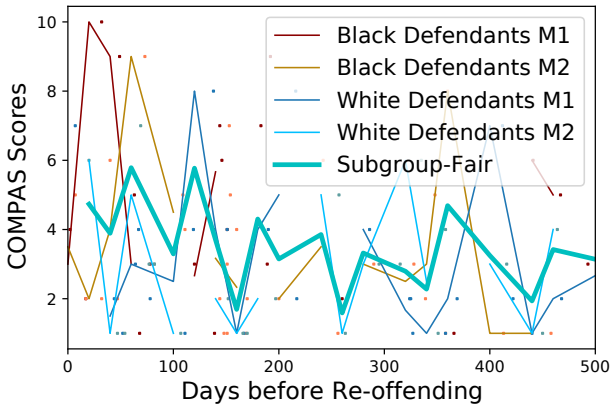

Figure 3: COMPAS recidivism scores of black and white defendants against the actual days before their re-offending. The sample of defendants' scores are divided into 4 sub-samples based on race and type of re-offending, distinguished by colours. Dots and curves with the same colour denote the scores of one sub-sample and the trajectory extracted from the scores respectively. The cyan curve displays the result of "Subgroup-Fair" model with 4 trajectories.

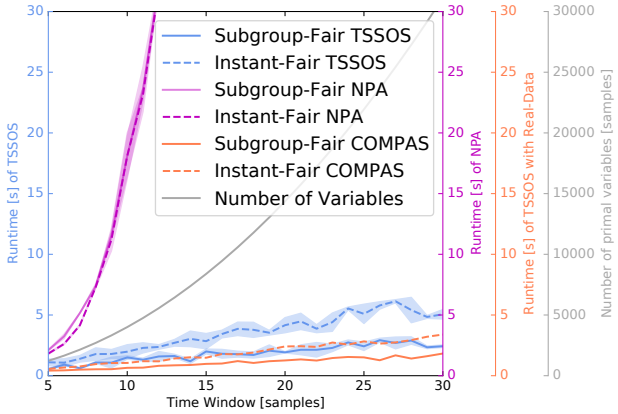

Figure 4: The dimensions of relaxations and the run-time of SDPA thereupon as a function of the length of time window. Run-time of TSSOS and NPA is displayed in cornflower-blue and deep-pink curves, respectively, while the grey curve shows the number of variables in relaxations. Additionally, the run-time of the COMPAS dataset of Figure 3 using TSSOS is also displayed as coral-coloured curves. For run-time, the mean and mean $\pm$ 1 standard deviations across 3 runs are presented by curves with shaded error bands.

within two years (label 1) or not (label 0). If the two-year recidivism label is 1, there is also information concerning the recharge degree and the number of days until the person gets rearrested. The dataset also consists of information on 'Days before Re-offending', which is the date difference between the defendant's crime offend date and recharge offend date. It could be negatively correlated to the defendant's actual risk level while the COMPAS recidivism scores would be the estimated risk level.

We choose 119 defendants with recidivism label 1, who are either African-American or Caucasian, male, within the age range of 25-45, and with prior crime counts less than 2, with charge degree M and recharge degree M1 or M2. The defendants are partitioned into two subgroups by their ethnicity and then partitioned by the type of their recharge degree (M1 or M2). Hence, we obtain the 4 sub-samples.

In the days-to-reoffend-vs-score plot, such as Figure 3, dots suggest COMPAS recidivism scores of the 4 sub-samples against the days before rearrest. Each curve represents one model, either subgroup-dependent (plotted thin) or Subgroup-Fair (plotted thick). The thick cyan curve is the race-blind prediction from our Subgroup-Fair method, which equalises scores across the two subgroups. Ideally, one should like to see smooth, monotonically decreasing curves, overlapping across all subgroup-dependent models. For each sub-sample, the aggregate deviation from the Subgroup-Fair curve would be similar to the aggregate deviations of other sub-samples.

In Figure 3, the dots are far removed from the ideal monotonically decreasing curve. Furthermore, the subgroup-specific curves (plotted thin) are very different from each other ("subgroup-specific models are unfair"). Specifically, the red and yellow curves are above the sky blue and cornflower blue curves ("at the same risk level, white defendants get lower COMPAS scores"). Notice that the subgroup-dependent models are obtained as follows: we discretise time to 20-day periods. For each subgroup, we check if anyone reoffends within 20 days (the first period). If so, the (average) COMPAS score (for all cases within the 20 days) is recorded as the observation of the first period of the trajectory of the sub-sample. If not, there is no observation of this period. We repeat this for the subsequent periods and for the three other sub-samples.

In Figure 4, the coral-coloured curve (for the COMPAS dataset) suggests that the run-time remains modest, even as the length of the time window grows to 30.

Conclusions

Overall, the two natural notions of fairness (subgroup fairness and instantaneous fairness), which we have introduced, may help establish the study of fairness in forecasting and learning of linear dynamical systems. We have presented globally convergent methods for estimation considering the two notions of fairness using hierarchies of convexifications of non-commutative polynomial optimisation problems. The run-time of standard solvers for the convexifications is independent of the dimension of the hidden state.

An interesting direction for further research extends the two notions of fairness towards distributional robustness (Hashimoto et al. 2018) and uses the notions of fairness in constraints (Donini et al. 2018), as well as in the objective. Following (Agarwal, Dudik, and Wu 2019; Calmon et al. 2017) one could also try to pre-process the observations or augment the data to remove the under-representation bias, although this may not be a realistic option in many applications.

Acknowledgements

Quan's and Bob's work has been supported by the Science Foundation Ireland under Grant 16/IA/4610. Jakub acknowledges support of the OP RDE funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics".

References

Åström, K. J. 1965. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications* 10(1): 174 – 205. ISSN 0022-247X.

Abdou, D. S. 2019. Gender-Based Price Discrimination: The Cost of Being a Woman. *Proceedings of Business and Economic Studies* 2(5).

Agarwal, A.; Dudik, M.; and Wu, Z. S. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, 120–129. PMLR.

Aghaei, S.; Azizi, M. J.; and Vayanos, P. 2019. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1418–1426.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias. *ProPublica, May* 23: 2016.

Awasthi, P.; Cortes, C.; Mansour, Y.; and Mohri, M. 2020. Beyond Individual and Group Fairness. *arXiv preprint arXiv:2008.09490* .

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Blum, A.; and Stangl, K. 2019. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy? *arXiv preprint arXiv:1912.01094* .

Brabec, J.; Komárek, T.; Franc, V.; and Machlica, L. 2020. On Model Evaluation Under Non-constant Class Imbalance. In *Computational Science – ICCS 2020*, 74–87. Cham: Springer International Publishing.

Calder, B. J.; Malthouse, E. C.; and Schaedel, U. 2009. An experimental study of the relationship between online engagement and advertising effectiveness. *Journal of interactive marketing* 23(4): 321–331.

Calmon, F.; Wei, D.; Vinzamuri, B.; Natesan Ramamurthy, K.; and Varshney, K. R. 2017. Optimized Pre-Processing for Discrimination Prevention. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5(2): 153–163.

Chouldechova, A.; and Roth, A. 2020. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM* 63(5): 82–89.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J. S.; and Pontil, M. 2018. Empirical Risk Minimization Under Fairness Constraints. In *Advances in Neural Information Processing Systems*, volume 31, 2791–2801. Curran Associates, Inc.

Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4(1): eaao5580.

Dutta, S.; Wei, D.; Yueksel, H.; Chen, P.-Y.; Liu, S.; and Varshney, K. R. 2020. An Information-Theoretic Perspective on the Relationship Between Fairness and Accuracy. *The 37th International Conference on Machine Learning (ICML 2020)* ArXiv preprint arXiv:1910.07870.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Gajane, P.; and Pechenizkiy, M. 2018. On formalizing fairness in prediction with machine learning. In Friedler, S. A.; and Wilson, C., eds., *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*. PMLR.

Geiger, P.; Zhang, K.; Schoelkopf, B.; Gong, M.; and Janzing, D. 2015. Causal inference by identification of vector autoregressive processes with hidden components. In *International Conference on Machine Learning*, 1917–1925.

Gelfand, I.; and Neumark, M. 1943. On the imbedding of normed rings into the ring of operators in Hilbert space. *Rec. Math. [Mat. Sbornik] N.S.* 12(2): 197–217.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323.

Hashimoto, T.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 1929–1938. Stockholmsmässan, Stockholm Sweden: PMLR.

He, H.; and Ma, Y. 2013. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons.

Helton, J. W. 2002. "Positive" noncommutative polynomials are sums of squares. *Annals of Mathematics* 156(2): 675–694.

Huang, S.; and Salm, M. 2020. The effect of a ban on gender-based pricing on risk selection in the German health insurance market. *Health economics* 29(1): 3–17.

Jung, C.; Kannan, S.; Lee, C.; Pai, M. M.; Roth, A.; and Vohra, R. 2020. Fair prediction with endogenous behavior. *arXiv preprint arXiv:2002.07147* .

Kilbertus, N.; Carulla, M. R.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, 656–666.

Klep, I.; Magron, V.; and Povh, J. 2019. Sparse Noncommutative Polynomial Optimization. *arXiv preprint arXiv:1909.00569* .

Klep, I.; Povh, J.; and Volcic, J. 2018. Minimizer extraction in polynomial optimization is robust. *SIAM Journal on Optimization* 28(4): 3177–3207.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 4066–4076.

Ljung, L. 1998. *System Identification: Theory for the User*. Pearson Education.

Locatello, F.; Abbati, G.; Rainforth, T.; Bauer, S.; Schölkopf, B.; and Bachem, O. 2019. On the Fairness of Disentangled Representations. In *Advances in Neural Information Processing Systems 32*, 14611–14624. Curran Associates, Inc.

McCullough, S. 2001. Factorization of operator-valued polynomials in several non-commuting variables. *Linear Algebra and its Applications* 326(1-3): 193–203.

Mouzannar, H.; Ohannessian, M. I.; and Srebro, N. 2019. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 359–368.

OECD. 2018. Personalised Pricing in the Digital Era. In *the joint meeting between the Competition Committee and the Committee on Consumer Policy*.

Paaßen, B.; Bunge, A.; Hainke, C.; Sindelar, L.; and Vogelsang, M. 2019. Dynamic fairness-Breaking vicious cycles in automatic decision making. In *Proceedings of the 27th European Symposium on Artificial Neural Networks (ESANN 2019)*.

Pearl, J. 2009. *Causality*. Cambridge University Press.

Pironio, S.; Navascués, M.; and Acín, A. 2010. Convergent Relaxations of Polynomial Optimization Problems with Noncommuting Variables. *SIAM Journal on Optimization* 20(5): 2157–2180.

Segal, I. E. 1947. Irreducible representations of operator algebras. *Bulletin of the American Mathematical Society* 53(2): 73–88.

Sharifi-Malvajerdi, S.; Kearns, M.; and Roth, A. 2019. Average Individual Fairness: Algorithms, Generalization and Experiments. In *Advances in Neural Information Processing Systems*, 8240–8249.

Thiery, Y.; and Van Schoubroeck, C. 2006. Fairness and equality in insurance classification. *The Geneva Papers on Risk and Insurance-Issues and Practice* 31(2): 190–211.

Wang, J.; and Magron, V. 2020. Exploiting term sparsity in Noncommutative Polynomial Optimization. *arXiv preprint arXiv:2010.06956* .

Wang, J.; Magron, V.; and Lasserre, J.-B. 2019. TSSOS: A Moment-SOS hierarchy that exploits term sparsity. *arXiv preprint arXiv:1912.08899* .

Wang, J.; Magron, V.; and Lasserre, J.-B. 2020. Chordal-TSSOS: a moment-SOS hierarchy that exploits term sparsity with chordal extension. *arXiv preprint arXiv:2003.03210* .

West, M.; and Harrison, J. 1997. *Bayesian Forecasting and Dynamic Models (2nd ed.)*. Berlin, Heidelberg: Springer-Verlag. ISBN 0-387-94725-6.

Wittek, P. 2015. Algorithm 950: Ncpol2sdpa—sparse semidefinite programming relaxations for polynomial optimization problems of noncommuting variables. *ACM Transactions on Mathematical Software (TOMS)* 41(3): 1–12.

Yamashita, M.; Fujisawa, K.; and Kojima, M. 2003. Implementation and evaluation of SDPA 6.0 (semidefinite programming algorithm 6.0). *Optimization Methods and Software* 18(4): 491–505.

Zhou, Q.; and Marecek, J. 2020. Proper Learning of Linear Dynamical Systems as a Non-Commutative Polynomial Optimisation Problem. *arXiv preprint arXiv:2002.01444* .

Zhou, Q.; Marecek, J.; and Shorten, R. N. 2020. Fairness in forecasting and learning linear dynamical systems. *arXiv preprint arXiv:2006.07315* .

Zliobaite, I. 2015. On the relation between accuracy and fairness in binary classification. In *The 2nd workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML) at ICML'15*.