Fairness in Forecasting and Learning Linear Dynamical Systems

Quan Zhou[1], Jakub Marecek[2], and Robert N. Shorten[1,3]

[1]University College Dublin, Ireland.
[2]Czech Technical Univeristy in Prague, the Czech Republic.
[3] Imperial College London, UK.

June 15, 2020

Abstract

As machine learning becomes more pervasive, the urgency of assuring its fairness increases. Consider training data that capture the behaviour of multiple subgroups of some underlying population over time. When the amounts of training data for the subgroups are not controlled carefully, under-representation bias may arise. We introduce two natural concepts of subgroup fairness and instantaneous fairness to address such under-representation bias in forecasting problems. In particular, we consider the learning of a linear dynamical system from multiple trajectories of varying lengths, and the associated forecasting problems. We provide globally convergent methods for the subgroup-fair and instant-fair estimation using hierarchies of convexifications of non-commutative polynomial optimisation problems. We demonstrate both the beneficial impact of fairness considerations on the statistical performance and the encouraging effects of exploiting sparsity on the estimators' run-time in our computational experiments.

1 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 [1]. This problem is also known as proper learning of linear dynamical systems (LDS) in System Identification [2]. As a rather general approach to time-series analysis, it has applications ranging from learning population-growth models in actuarial science and mathematical biology [3] to functional analysis in neuroscience [4]. Indeed, one encounters either partially observable processes [5] or questions of causality [6] that can be tied to proper learning of LDS [7] in almost any application domain.

A discrete-time model of a linear dynamical system $\mathcal{L} = (G, F, V, W)$ [1] 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 [2].

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 a set of subgroups $\mathcal{S}$.
- 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 for notations, the superscripts denote the trajectories and subgroups while subscripts indicates the periods.

In this setting, under-representation bias [8, 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. [9]). Notice that such anti-discrimination measures are increasingly stipulated by the 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 at certain triple (t, i, s), there is no observation and $Y_t^{(i,s)}$, $\text{loss}^{(i,s)}(f_t)$ are infeasible.

We propose two novel objective to address the under-representation bias:

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 in the objective:

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

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

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

1.1 Contributions

Overall, our contributions are the following:

- We introduce two new notions of fairness into forecasting.
- We cast proper and improper learning of a linear dynamical system with fairness considerations as a non-commutative polynomial optimisation problem (NCPOP).
- We prove convergence of an algorithm based on the convergent hierarchy of semi-definite programming (SDP) relaxations.
- We study the numerical methods for solving the resulting NCPOP and extracting its optimiser.

This presents an algorithmic approach to addressing the under-representation bias studied by Blum et al. [8] and presents a step forward within the fairness in forecasting studied recently by [9, 10, 11], as outlined in the excellent survey of [12]. It follows much work on fairness in classification, e.g., [13, 14, 15, 16, 12, 17]. It is complemented by several recent studies involving dynamics and fairness [18, 19, 20], albeit not involving learning of dynamics. It relies crucially on tools developed in non-commutative polynomial optimisation [21, 22, 23, 24, 25] and non-commutative algebra [26, 27, 28, 29], which have not seen much use in Statistics and Machine Learning, yet.

2 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. The insurer will hence pay more to a female insured over the course of her lifetime, compared to a male insured, on averag [30]. 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 [30]. Consequently, male insureds might leave the annuity-benefit scheme (known as the 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.

To be more specific, 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 in the end of each year for at most 10 years on the condition that it is still alive. Let s_t denotes how many insureds left in the scheme in the end of the t^{th} year. Suppose there are s_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.

$$\text{premium} := \frac{\sum_{t=1}^{10} s_t \times (1+i)^{-t}}{s_0} \tag{6}$$

The most important quality s_t is derived from estimating insureds' life expectancy. Suppose the insureds can be divided into female subgroup and male subgroup and each subgroup only have 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 in the end of the t^{th} year respectively. Both trajectories can be regarded as linear dynamic systems. We have

$$\begin{aligned} Y_t^{(\,\cdot\,,f)} &= G^{(f)} Y_{t-1}^{(\,\cdot\,,f)} + \omega_t^{(f)}, \ 2 \leq t \leq 10, && (7) \\ Y_t^{(\,\cdot\,,m)} &= G^{(m)} Y_{t-1}^{(\,\cdot\,,m)} + \omega_t^{(m)}, \ 2 \leq t \leq 10, && (8) \\ s_t &= Y_t^{(\,\cdot\,,f)} + Y_t^{(\,\cdot\,,m)}, \ 1 \leq t \leq 10, && (9) \end{aligned}$$

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:

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

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:

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

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

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, we have

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

$$\min_{f_t, 1\leq t\leq 10} \max \left\{ \overset{(\,\cdot\,,f)}{\text{loss}}(f_1), \ldots, \overset{(\,\cdot\,,f)}{\text{loss}}(f_{10}), \overset{(\,\cdot\,,m)}{\text{loss}}(f_1), \ldots, \overset{(\,\cdot\,,m)}{\text{loss}}(f_{10}) \right\} \tag{14}$$

Personalised pricing Another application arises in personalised pricing (PP). The extent of personalised pricing is growing, as the amounts of data available to pricing strategies increase. Suitable data include user locations, IP address, web visits, past purchases and additional information volunteered by customers [31]. There are concerns that the practice may hurt overall trust, as it did in the well-known case [31] of a consumer, who found out that Amazon was selling products to regular consumers at higher prices, and that deleting the cookies on the computer could cause the inflated prices to drop.Furthermore, the practice can also violate anti-discrimination law [31] in many jurisdictions. For instance in the United States, the Federal Trade Commission enforces the Equal Credit Opportunity Act (ECOA), which bars offering prices for credit from utilising certain protected consumer characteristics such as race, colour, religion, national origin, sex, marital status, age, or the receipt of public assistance. This risk would force many entities offering financial products to set the same price for the subgroups, regardless of the significant differences in their willingness to pay.

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 $S = \{\text{female, male}\}$ 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 the 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:

$$
\begin{aligned}
\phi_t^f &= G^{(f)}\phi_{t-1}^{(f)} + \omega_t^{(f)} \quad, t \in \mathcal{T}^+, && (15)\\
Y_t^{(i,f)} &= F^{(f)\prime}\phi_t^{(f)} + \nu_t^{(i,f)} \quad, t \in \mathcal{T}^{(i,f)}, i = \{1, \dots, \mathcal{I}^{(f)}\}, && (16)\\
\phi_t^m &= G^{(m)}\phi_{t-1}^{(m)} + \omega_t^{(m)} \quad, t \in \mathcal{T}^+, && (17)\\
Y_t^{(i,m)} &= F^{(m)\prime}\phi_t^{(m)} + \nu_t^{(i,m)} \quad, t \in \mathcal{T}^{(i,m)}, i = \{1, \dots, \mathcal{I}^{(m)}\}. && (18)
\end{aligned}
$$

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:

$$
\begin{aligned}
m_t &= Gm_{t-1} + \omega_t \ , \ t \in \mathcal{T}^+ && (19)\\
f_t &= F'm_t + \nu_t \quad , \ t \in \mathcal{T}^+ && (20)
\end{aligned}
$$

For $\text{loss}^{(i,f)}(f_t) := ||Y_t^{(i,f)} - f_t||, t \in \mathcal{T}^{(i,f)}, i = \{1, \dots, \mathcal{I}^{(f)}\}$ and $\text{loss}^{(i,m)}(f_t) := ||Y_t^{(i,m)} - f_t||, t \in \mathcal{T}^{(i,m)}, i = \{1, \dots, \mathcal{I}^{(m)}\}$, we can consider

$$
\min_{f_t, t \in \mathcal{T}^+} \max \left\{ \frac{1}{|\mathcal{I}^{(f)}|} \sum_{i=1}^{\mathcal{I}^{(f)}} \frac{1}{|\mathcal{T}^{(i,f)}|} \sum_{t \in \mathcal{T}^{(i,f)}} \overset{(i,f)}{\text{loss}}(f_t), \frac{1}{|\mathcal{I}^{(m)}|} \sum_{i=1}^{\mathcal{I}^{(m)}} \frac{1}{|\mathcal{T}^{(i,m)}|} \sum_{t \in \mathcal{T}^{(i,m)}} \overset{(i,m)}{\text{loss}}(f_t) \right\} \tag{21}
$$

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

We also refer to [32] for further work on protecting customers' interests in personalised pricing via fairness considerations.

3 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.

$$
\begin{aligned}
m_t &= Gm_{t-1} + \omega_t && (23)\\
f_t &= F'm_t + \nu_t. && (24)
\end{aligned}
$$

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. (See the Supplementary Material.) Overall, the subgroup-fair and instant-fair formulations read:

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

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

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

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

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 [21]. Let $Q = \{q_i\}$ be the set of polynomials determining the constraints. The positivity domain $\mathbf{S}_Q$ of Q are tuples $X = (X_1, \dots, X_n)$ 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 [21], we assume:

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

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. Further, with suitably modified assumptions, the result holds also for the instant-fair considerations* (26)*.*

The proof is in the Supplementary Material. In summary, Theorem 2 makes it possible to recover the quadruple (G, F, V, W) of the subgroup-blind $\mathcal{L}$ using the technologies of NCPOP with guarantees of global convergence [21].

4 Numerical illustrations

4.1 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 of Section 2, with varying degrees of the bias. Suppose there are two subgroups, 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 training set in the following ways:

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 $\mathcal{I}^{(d)}$ so that $\beta^{(a)} \geq \beta^{(d)}$. If $\mathcal{I}^{(a)}$ is fixed at 1, 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 small sample size, it is necessary to make sure there is at least one observation in each subgroup at each period.

For example, consider that 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. Set the time window to be 20 across 3 trajectories in the advantaged subgroup and 2 in disadvantaged

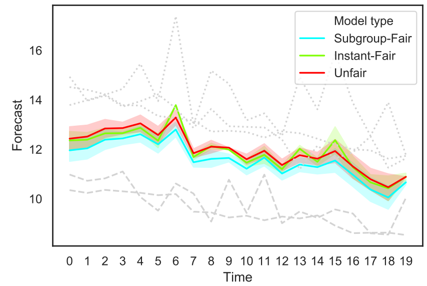


Figure 1: Forecast obtained using (25)-(27): the solid lines in primary colours with error bands display the mean and standard deviation of the forecasts over 30 experiments. For reference, dotted lines and dashed lines in grey denote the trajectories of observations of advantaged and disadvantaged subgroups, respectively, before discarding any observations.

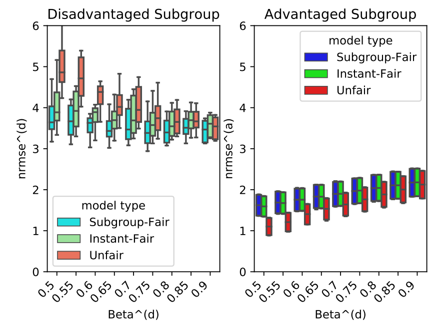


Figure 2: Accuracy as a function of the degree of under-representation bias: The boxplot of $\mathrm{nrmse}^{(s)}, s \in \mathcal{S}$ against $\beta^{(d)}$, where $\beta^{(d)} = [0.5, 0.55, \ldots, 0.9]$, with boxes for the quartiles of $\mathrm{nrmse}^{(s)}$ obtained from 5 experiments, using the observations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

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 30 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)-(26) are applied in each experiment with λ of 5 and 1, respectively, as chosen by iterating over integers 1 to 10. The mean of forecast f and its standard deviation are displayed as the solid curves with error bands.

4.2 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 $\beta^{(d)}$ within $0.5 \leq \beta^{(d)} \leq 1$. To measure the effect of the degree on accuracy, we introduce the normalised root mean squared error (nrmse) fitness value for each subgroup:

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

for $s \in \mathcal{S}$ and $\text{mean}^{(s)} := \frac{1}{|\mathcal{I}^{(s)}|}\sum_{i\in\mathcal{I}^{(s)}}\frac{1}{|\mathcal{T}^{(i,s)}|}\sum_{t\in\mathcal{T}^{(i,s)}}Y_t^{(i,s)}$. Higher $\text{nrmse}^{(s)}$ indicates lower accuracy for the subgroup, 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 advantaged subgroup is discarded. Then, the biased training data generalisation process in Section 4.1 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. At 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 $\text{nrmse}^{(s)}$ for each subgroup shown as boxes in Figure 2.

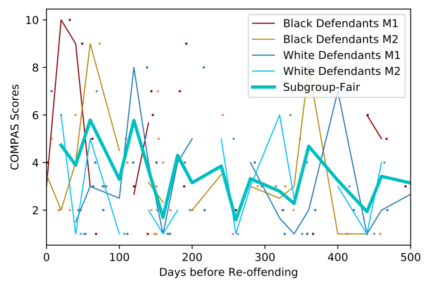
One could expect that nrmse fitness values of advantaged subgroup in Figure 2 to be generally lower than those of the disadvantaged subgroup ($\text{nrmse}^{(d)} \geq \text{nrmse}^{(a)}$), leaving a gap. Those gaps narrow down as $\beta^{(d)}$ increases, simply because more observations of disadvantaged 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 $\text{nrmse}^{(a)}$ decreases as $\beta^{(d)}$ gets close to 0.5, "Subgroup-Fair" model still can keep the $\text{nrmse}^{(d)}$ at almost the same level, indicating a rise in overall accuracy. This is in contrast with results [13, 33] from classification.

4.3 Run-time

Notice that minimising a multivariate polynomial in matricial variables (25)-(27) over a set defined by a finite intersection of polynomial inequalities in the same variables is non-trivial, but there exists the globally convergent Navascués-Pironio-Acín (NPA) hierarchy [34] of semidefinite programming (SDP) relaxations as explained in the Supplementary Material, and its sparsity-exploiting variant (TSSOS) as pioneered by Wang et al. [24, 25], which can be applied to such non-commutative polynomial optimisation problems. The SDP of a given order in the respective hierarchy can be constructed using `ncpol2sdpa`[1] of Wittek [22] or the `tssos`[2] of Wang et al. [24, 25] and solved by `sdpa` of Yamashita et al. [35]. Our implementation is available in the Supplementary Material for review purposes and will be open-sourced upon acceptance.

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).

[1] `https://github.com/peterwittek/ncpol2sdpa`
[2] `https://github.com/wangjie212/TSSOS`



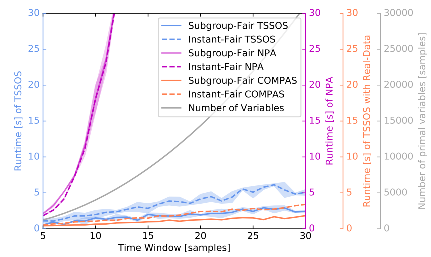


Figure 3: COMPAS recidivism scores of black and white defendants against the actual days before their re-offending. The sample of defendants' scores are divided into 4 sub-samples based on race and type of re-offending, distinguished by colours. Dots and curves with the same colour denote the scores 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.

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 surges as can be expected, given that the number of SDP variables is equivalent to that of relaxation variables or the entries in the moment matrix ($M_k(y)$, as defined in the Supplementary Material, cf. Eq. 32).

4.4 Experiments with COMPAS recidivism scores

Finally, we wish to suggest the broader applicability of the two notions of subgroup fairness and instantaneous fairness. We use the well-known benchmark dataset [36] 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. Broadly speaking, the defendants' risk scores (the higher the worse) are negatively correlated with the amount of time before defendants' recidivism. However, the correlation is different between black and white defendants' COMPAS scores.

We consider all defendants (N = 21) within the age range of 25-45, male, with two or less prior crime counts, labelled as belonging to either African-American or Caucasian ethnicity. The defendants are partitioned into two subgroups, by ethnicity. In each subgroup, defendants are divided by the type of their re-offending (M1 and M2). The COMPAS scores of 4 sub-samples are shown in Figure 3 by dots, where warm and cold tones denote African-American and Caucasian subgroups respectively. The trajectory shown by same colour is obtained from dots in corresponding sub-sample. The Subgroup-Fair outcome is presented in cyan. In Figure 4, the coral-coloured curve for the COMPAS dataset suggests that the run-time remains modest. While the COMPAS dataset calls for classification, rather than forecasting, our notion also seems to be applicable.

5 Conclusions

Overall, the two natural notions of fairness (subgroup fairness and instantaneous fairness), which we have introduced, contribute towards the fairness in forecasting and proper learning of linear dynamical systems. We have presented globally convergent methods for the estimation considering the two notions of fairness using hierarchies of convexifications of non-commutative polynomial optimisation problems, whose run-time is independent of the hidden state.

Acknowledgements

Quan's and Bob's work has been supported by the Science Foundation Ireland under Grant 16/IA/4610. Jakub acknowledges funding from RCI (reg. no. CZ.02.1.01/0.0/ 0.0/16_019/0000765) supported by the European Union.

References

[1] M. West and J. Harrison, *Bayesian Forecasting and Dynamic Models (2nd ed.)*. Berlin, Heidelberg: Springer-Verlag, 1997.

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

[3] P. H. Leslie, "On the use of matrices in certain population mathematics," *Biometrika*, vol. 33, no. 3, pp. 183–212, 11 1945.

[4] M. Besserve, B. Schölkopf, N. K. Logothetis, and S. Panzeri, "Causal relationships between frequency bands of extracellular signals in visual cortex revealed by an information theoretic analysis," *Journal of computational neuroscience*, vol. 29, no. 3, pp. 547–566, 2010.

[5] K. J. Åström, "Optimal control of markov processes with incomplete state information," *Journal of Mathematical Analysis and Applications*, vol. 10, no. 1, pp. 174 – 205, 1965.

[6] J. Pearl, *Causality*. Cambridge university press, 2009.

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

[8] A. Blum and K. Stangl, "Recovering from biased data: Can fairness constraints improve accuracy?" *arXiv preprint arXiv:1912.01094*, 2019.

[9] P. Gajane and M. Pechenizkiy, "On formalizing fairness in prediction with machine learning," *arXiv preprint arXiv:1710.03184*, 2017.

[10] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[11] F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem, "On the fairness of disentangled representations," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 14 611–14 624.

[12] A. Chouldechova and A. Roth, "A snapshot of the frontiers of fairness in machine learning," *Commun. ACM*, vol. 63, no. 5, p. 82–89, Apr. 2020.

[13] I. Zliobaite, "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*, 2015.

[14] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

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

[16] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, 2017, pp. 4066–4076.

[17] S. Aghaei, M. J. Azizi, and P. Vayanos, "Learning optimal and fair decision trees for non-discriminative decision-making," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 1418–1426.

[18] H. Mouzannar, M. I. Ohannessian, and N. Srebro, "From fair decision making to social equality," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 359–368.

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

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

[21] S. Pironio, M. Navascués, and A. Acín, "Convergent relaxations of polynomial optimization problems with noncommuting variables," *SIAM Journal on Optimization*, vol. 20, no. 5, pp. 2157–2180, 2010.

[22] P. Wittek, "Algorithm 950: Ncpol2sdpa—sparse semidefinite programming relaxations for polynomial optimization problems of noncommuting variables," *ACM Transactions on Mathematical Software (TOMS)*, vol. 41, no. 3, pp. 1–12, 2015.

[23] I. Klep, J. Povh, and J. Volcic, "Minimizer extraction in polynomial optimization is robust," *SIAM Journal on Optimization*, vol. 28, no. 4, pp. 3177–3207, 2018.

[24] J. Wang, V. Magron, and J.-B. Lasserre, "Tssos: A moment-sos hierarchy that exploits term sparsity," *arXiv preprint arXiv:1912.08899*, 2019.

[25] ——, "Chordal-tssos: a moment-sos hierarchy that exploits term sparsity with chordal extension," *arXiv preprint arXiv:2003.03210*, 2020.

[26] I. Gelfand and M. Neumark, "On the imbedding of normed rings into the ring of operators in Hilbert space," *Rec. Math. [Mat. Sbornik] N.S.*, vol. 12, no. 2, pp. 197–217, 1943.

[27] I. E. Segal, "Irreducible representations of operator algebras," *Bulletin of the American Mathematical Society*, vol. 53, no. 2, pp. 73–88, 1947.

[28] S. McCullough, "Factorization of operator-valued polynomials in several non-commuting variables," *Linear Algebra and its Applications*, vol. 326, no. 1-3, pp. 193–203, 2001.

[29] J. W. Helton, ""Positive" noncommutative polynomials are sums of squares," *Annals of Mathematics*, vol. 156, no. 2, pp. 675–694, 2002.

[30] Y. Thiery and C. Van Schoubroeck, "Fairness and equality in insurance classification," *The Geneva Papers on Risk and Insurance-Issues and Practice*, vol. 31, no. 2, pp. 190–211, 2006.

[31] OECD, "Personalised pricing in the digital era," in *the joint meeting between the Competition Committee and the Committee on Consumer Policy*, 2018.

[32] R. Dong, E. Miehling, and C. Langbort, "Protecting consumers against personalized pricing: A stopping time approach," *arXiv preprint arXiv:2002.05346*, 2020.

[33] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. R. Varshney, "An information-theoretic perspective on the relationship between fairness and accuracy," *The 37th International Conference on Machine Learning (ICML 2020)*, 2020, arXiv preprint arXiv:1910.07870.

[34] S. Pironio, M. Navascués, and A. Acin, "Convergent relaxations of polynomial optimization problems with noncommuting variables," *SIAM Journal on Optimization*, vol. 20, no. 5, pp. 2157–2180, 2010.

[35] M. Yamashita, K. Fujisawa, and M. Kojima, "Implementation and evaluation of sdpa 6.0 (semidefinite programming algorithm 6.0)," *Optimization Methods and Software*, vol. 18, no. 4, pp. 491–505, 2003.

[36] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica, May*, vol. 23, p. 2016, 2016.

[37] S. Burgdorf, I. Klep, and J. Povh, *Optimization of polynomials in non-commuting variables.* Springer, 2016.

[38] A. K. Tangirala, *Principles of system identification: theory and practice.* Crc Press, 2014.

[39] K.-J. Åström and B. Torsten, "Numerical identification of linear dynamic systems from normal operating records," *IFAC Proceedings Volumes*, vol. 2, no. 2, pp. 96–111, 1965, 2nd IFAC Symposium on the Theory of Self-Adaptive Control Systems, Teddington, UK, September 14-17, 1965.

[40] O. Anava, E. Hazan, S. Mannor, and O. Shamir, "Online learning for time series prediction," in *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, 2013.

[41] C. Liu, S. C. H. Hoi, P. Zhao, and J. Sun, "Online arima algorithms for time series prediction," ser. AAAI'16, 2016.

[42] M. Kozdoba, J. Marecek, T. Tchrakian, and S. Mannor, "On-line learning of linear dynamical systems: Exponential forgetting in kalman filters," in *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019, arXiv preprint arXiv:1809.05870.

[43] A. Tsiamis and G. Pappas, "Online learning of the kalman filter with logarithmic regret," *arXiv preprint arXiv:2002.05141*, 2020.

[44] T. Katayama, *Subspace methods for system identification.* Springer Science & Business Media, 2006.

[45] P. Van Overschee and B. De Moor, *Subspace identification for linear systems. Theory, implementation, applications. Incl. 1 disk*, 01 1996, vol. xiv.

[46] A. Tsiamis, N. Matni, and G. J. Pappas, "Sample complexity of Kalman filtering for unknown systems," *arXiv preprint arXiv:1912.12309*, 2019.

[47] A. Tsiamis and G. J. Pappas, "Finite sample analysis of stochastic system identification," *arXiv preprint arXiv:1903.09122*, 2019.

[48] E. Hazan, K. Singh, and C. Zhang, "Learning linear dynamical systems via spectral filtering," in *Advances in Neural Information Processing Systems*, 2017, pp. 6702–6712.

[49] E. Hazan, H. Lee, K. Singh, C. Zhang, and Y. Zhang, "Spectral filtering for general linear dynamical systems," in *Advances in Neural Information Processing Systems*, 2018, pp. 4634–4643.

[50] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, 2018.

[51] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," in *Conference On Learning Theory*, 2018, pp. 439–473.

[52] M. Simchowitz, R. Boczar, and B. Recht, "Learning linear dynamical systems with semi-parametric least squares," *arXiv preprint arXiv:1902.00768*, 2019.

[53] T. Sarkar and A. Rakhlin, "Near optimal finite time identification of arbitrary linear dynamical systems," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 5610–5618.

[54] V. V. Vazirani, *Approximation algorithms.* Springer Science & Business Media, 2013.

[55] S. Samadi, U. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. Vempala, "The price of fair pca: One extra dimension," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 976–10 987.

[56] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, 2018, pp. 77–91.

[57] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in neural information processing systems*, 2016, pp. 4349–4357.

[58] S. Sharifi-Malvajerdi, M. Kearns, and A. Roth, "Average individual fairness: Algorithms, generalization and experiments," in *Advances in Neural Information Processing Systems*, 2019, pp. 8240–8249.

[59] U. Tantipongpipat, S. Samadi, M. Singh, J. H. Morgenstern, and S. Vempala, "Multi-criteria dimensionality reduction with applications to fairness," in *Advances in Neural Information Processing Systems*, 2019, pp. 15 135–15 145.

[60] N. I. Akhiezer and M. Krein, *Some questions in the theory of moments.* American Mathematical Society, 1962, vol. 2.

[61] D. Henrion and J.-B. Lasserre, "Detecting global optimality and extracting solutions in gloptipoly," in *Positive polynomials in control.* Springer, 2005, pp. 293–310.

[62] J. Dixmier, *Les C*-algèbres et leurs représentations.* Paris, France: Gauthier-Villars, 1969, English translation: C*-algebras (North-Holland, 1982).

6 Background

In this paper, we would like to consider the case of multiple variants of the LDS and conduct proper learning of the LDS in a way of fairness using the technologies of non-commutative polynomial optimisation. In Section 6.1, we set our work in the context of system identification and control theory. In Section 6.2, we introduce the concept of fairness, which can be used to deal with multiple variants of the LDS. In Section 6.3, we provide a brief overview of non-commutative polynomial optimisation, pioneered by [21] and nicely surveyed by [37], which is our key technical tool.

6.1 Related Work in System Identification and Control

Research within System Identification variously appears in venues associated with Control Theory, Statistics, and Machine learning. We refer to [2] and [38] for excellent overviews of the long history of research in the field, going back at least to [39]. In this section, we focus on pointers to key more recent publications. In improper learning of LDS, a considerable progress has been made in the analysis of predictions for the expectation of the next measurement using auto-regressive (AR) processes. In [40], first guarantees were presented for auto-regressive moving-average (ARMA) processes. In [41], these results were extended to a subset of autoregressive integrated moving average (ARIMA) processes. [42] have shown that up to an arbitrarily small error given in advance, AR(s) will perform as well as *any* Kalman filter on any bounded sequence. This has been extended by [43] to Kalman filtering with logarithmic regret. Another stream of work within improper learning focuses on sub-space methods [44, 45] and spectral methods. [46, 47] presented the present-best guarantees for traditional sub-space methods. Within spectral methods, [48] and [49] have considered learning LDS with input, employing certain eigenvalue-decay estimates of Hankel matrices in the analyses of an auto-regressive process in a dimension increasing over time. We stress that none of these approaches to improper learning are "prediction-error": They do *not* estimate the system matrices.

In proper learning of LDS, many state-of-the-art approaches consider the least-squares method, despite complications encountered in unstable systems [50]. [51] have provided non-trivial guarantees for the ordinary least-squares (OLS) estimator in the case of stable G and there being no hidden component, i.e., F' being an identity and $Y_t = \phi_t$. Surprisingly, they have also shown that more unstable linear systems are easier to estimate than less unstable ones, in some sense. [52] extended the results to allow for a certain pre-filtering procedure. [53] extended the results to cover stable, marginally stable, and explosive regimes.

Our work could be seen as a continuation of the least squares method to processes with hidden components, with guarantees of global convergence. In Computer Science, our work could be seen as an approximation scheme [54], as it allows for ϵ error for any $\epsilon > 0$.

6.2 Fairness

In machine learning, the training set might have biased representations of its subgroups, even when sampled with equal weight [55]. Algorithms focusing on maximising the overall accuracy might cause different distribution of errors in different subgroups.

In facial recognition, [56] find out that darker-skinned females are the most misclassied subgroup with error rates of up to 34.7% while the maximum error rate for lighter-skinned males is 0.8% as a result of the imbalanced gender and skin type distribution of the datasets of facial analysis benchmarks. Another threat facing us is gender bias shown in word embedding where the word *female* is tender to be associated to *receptionist* [57].

In concerns of the uneven distribution of error over subgroups, fairness was introduced to the field of machines learning. According to a clear summary in [12], the definition of fairness can be derived from a statistical notion and an individual notion. The statistical definition of fairness is to request a classifier's statistic, such as false positive or false negative rates be equalized across the subgroups so that the error caused by the algorithm be proportionately spread across subgroups [12]. The statistical definition has a natural connection with Principal Component Analysis (PCA). Introduced in [55], the Fair-PCA problem aims to minimize the maximum construction loss of different subgroups when looking for a lower dimensional representation. To solve the Fair-PCA problem, [58] design an oracle-efficient algorithm while [59] propose an algorithms based on extreme-point solutions of semi-definite programs. The individual definition is discussed less on account of its requirement of making significant assumptions even through it has strong individual level semantics that one's risk of being harmed by the error of the classifier are no higher than they are for anyone else [58].

We can introduce fairness to learning of LDS when dealing with multiple variants of the LDS. When estimating the next observation, one might be given several trajectories of observations from unknown variants of the LDS. In this case, fairness asks to find a suitable model that treats each LDS equally.

6.3 Non-Commutative Polynomial Optimisation

In learning of the LDS, the key technical tool of this paper is non-commutative polynomial optimisation (NCPOP), first introduced by [21]. Here, we provide a brief summary of their results, and refer to [37] for a book-length introduction. NCPOP is an operator-valued optimisation problem with a standard form in Problem 29:

$$P: \quad \begin{array}{rcl} p* = & \min\limits_{(H,X,\phi)} & \langle \phi, p(X)\phi \rangle \\ & \text{s.t.} & q_i(X) \succcurlyeq 0, i = 1, \ldots, m, \\ & & \langle \phi, \phi \rangle = 1, \end{array} \tag{29}$$

where $X = (X_1, \ldots, X_n)$ is a bounded operator on a Hilbert space $\mathcal{H}$. The normalised vector ϕ, i.e., $\|\phi\|^2 = 1$ is also defined on $\mathcal{H}$ with inner product $\langle \phi, \phi \rangle$ equals to 1. $p(X)$ and $q_i(X)$ are polynomials and $q_i(X) \succcurlyeq 0$ denotes that the operator $q_i(X)$ is positive semi-definite. Polynomials $p(X)$ and $q_i(X)$ of degrees $\deg(p)$ and $\deg(q_i)$, respectively, can be written as:

$$p(X) = \sum_{|\omega| \leq \deg(p)} p_\omega \omega, \quad q_i(X) = \sum_{|\mu| \leq \deg(q_i)} q_{i,\mu} \mu, \tag{30}$$

where $i = 1, \ldots, m$. Following [60], we can define the moments on field $\mathbb{R}$ or $\mathbb{C}$, with a feasible solution (H, X, ϕ) of problem (29):

$$y_\omega = \langle \phi, \omega(X)\phi \rangle, \tag{31}$$

for all $\omega \in \mathcal{W}_\infty$ and $y_1 = \langle \phi, \phi \rangle = 1$. Given a degree k, the moments whose degrees are less or equal to k form a sequence of $y = (y_\omega)_{|\omega| \leq 2k}$. With a finite set of moments y of degree k, we can define a corresponding k^{th} order moment matrix $M_k(y)$:

$$M_k(y)(\nu, \omega) = y_{\nu^\dagger \omega} = \langle \phi, \nu^\dagger(X)\omega(X)\phi \rangle, \tag{32}$$

for any $|\nu|, |\omega| \leq k$ and a localising matrix $M_{k-d_i}(q_i y)$:

$$\begin{aligned} M_{k-d_i}(q_i y)(\nu, \omega) &= \sum_{|\mu| \leq \deg(q_i)} q_{i,\mu} y_{\nu^\dagger \mu \omega} \\ &= \sum_{|\mu| \leq \deg(q_i)} q_{i,\mu} \langle \phi, \nu^\dagger(X)\mu(X)\omega(X)\phi \rangle, \end{aligned} \tag{33}$$

for any $|\nu|, |\omega| \leq k - d_i$, where $d_i = \lceil \deg(q_i)/2 \rceil$. The upper bounds of $|\nu|$ and $|\omega|$ are lower than the that of moment matrix because $y_{\nu^\dagger \mu \omega}$ is only defined on $\nu^\dagger \mu \omega \in \mathcal{W}_{2k}$ while $\mu \in \mathcal{W}_{\deg(q_i)}$.

If (H, X, ϕ) is feasible, one can utilize the Sums of Squares theorem of [29] and [28] to derive semidefinite programming (SDP) relaxations. In particular, we can obtain a k^{th} order SDP relaxation of the non-commutative polynomial optimization problem (29) by choosing a degree k that satisfies the condition of $2k \geq \max\{\deg(p), \max_i \deg(q_i)\}$. The SDP relaxation of order k, which we denote R_k, has the form:

$$\begin{array}{llll} & p^k = & \min\limits_{y = (y_\omega)_{|\omega| \leq 2k}} & \sum\limits_{|\omega| \leq d} p_\omega y_\omega \\ R_k : & & \text{s.t.} & M_k(X) \succcurlyeq 0, \\ & & & M_{k-d_i}(q_i X) \succcurlyeq 0, i = 1, \ldots, m, \\ & & & \langle \phi, \phi \rangle = 1, \end{array} \tag{34}$$

Let us define the quadratic module, following [21]. 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 [21], we assume:

Assumption 3 (Archimedean)**.** *Quadratic module M_Q of* (29) *is Archimedean, i.e., there exists a real constant C such that $C^2 - (X_1^\dagger X_1 + \cdots + X_{2n}^\dagger X_{2n}) \in M_Q$, where $X_{n+i}, i \in \{1, \ldots, n\}$ are defined to be $X_i^\dagger$.*

If the Archimedean assumption is satisfied, [21] have shown that $\lim_{k\to\infty} p^k = p^*$ for a finite k. We can use the so-called rank-loop condition of [21] to detect global optimality. Once detected, it is possible to extract the global optimum (H^*, X^*, ϕ^*) from the optimal solution y of problem R_k, by Gram decomposition; cf. Theorem 2 in [21]. Simpler procedures for the extraction have been considered, cf. [61], but remain less well understood.

More complicated procedures for the extraction are also possible. Notably, the Gelfand–Naimark–Segal (GNS) construction [26, 27] does not require the rank-loop condition to be satisfied, as is well explained in Section 2.2 of [23], cf. also Section 2.6 of [62].

7 Proof of Theorem 2

Putting the elements together, we an prove the Theorem 2:

Proof. First, we need to show the existence of a sequence of convex optimisation problems, whose objective function approaches the optimum of the non-commutative polynomial optimisation problem. [21] shows that, indeed, there are natural semidefinite programming problems, which satisfy this property. In particular, the existence and convergence of the sequence is shown by Theorem 1 of [21], which requires Assumption 1. Second, we need to show that the extraction of the minimiser from the SDP relaxation of order $k(\epsilon)$ in the series is possible. There, one utilises the Gelfand–Naimark–Segal (GNS) construction [26, 27], as explained in Section 2.2 of [23]. □

8 Details of Experiments with COMPAS recidivism scores

For the experiment with COMPAS recidivism scores, we use the "COMPAS" dataset from [36]. The dataset include 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 within two years (label 1) or not (label 0). If the two-year recidivism label is 1, there is also information of the time between a person received the COMPAS score and got rearrested, and the recharge degree.

We choose defendants that are with recidivism label 1, African-American or Caucasian, within the age range of 25-45, male and with prior crime counts less than 2, with charge degree M and recharge degree M1 or M2. The sample size is 119. We plot their COMPAS recidivism scores against the days before they got rearrested, which are shown by the dots in Figure 3. We distinguish the ethnicity of the defendants by warm-tone and cold-tone colours. Further, defendants with recharge degree M1 are displayed with darker colours than those with recharge degree M2. Note that each colour denotes one sub-sample.

Every 20 days are regarded as one period. Then, we try to extract trajectories from COMPAS scores of each sub-sample. For African-American defendants with recharge degree M1, we check if anyone re-offend within 20 days. If there is one, its COMPAS score is recorded as the observation of first period of the trajectory of "Black Defendants M1"; if there are more than one, their average are recorded as the observation. Then we check the following periods, up to 21 periods. Also, the same procedure can be applied to other three sub-samples. In the end, we get four trajectories of each sub-sample, which are shown by the curves in Figure 3. With the four trajectories, we can apply the Subgroup-Fair and the results is shown by the cyan curve.