Fairness in Forecasting of Observations of Linear Dynamical Systems

Quan Zhou   Q.ZHOU22@IMPERIAL.AC.UK
Dyson School of Design Engineering,
Imperial College London,
London, SW7 9EG, United Kingdom
School of Electrical and Electronic Engineering,
University College Dublin,
Dublin, D04 V1W8, Ireland

Jakub Mareček   JAKUB.MARECEK@FEL.CVUT.CZ
Department of Computer Science,
Czech Technical University in Prague,
Prague, 121 35, the Czech Republic

Robert Shorten   R.SHORTEN@IMPERIAL.AC.UK
Dyson School of Design Engineering,
Imperial College London,
London, SW7 9EG, United Kingdom
School of Electrical and Electronic Engineering,
University College Dublin,
Dublin, D04 V1W8, Ireland

Abstract

In machine learning, training data often capture the behaviour of multiple subgroups of some underlying human population. This behaviour can often be modelled as observations of an unknown dynamical system with an unobserved state. When the training data for the subgroups are not controlled carefully, however, under-representation bias arises. To counter under-representation bias, we introduce two natural notions of fairness in time-series forecasting problems: subgroup fairness and instantaneous fairness. These notion extend predictive parity to the learning of dynamical systems. We also show globally convergent methods for the fairness-constrained learning problems using hierarchies of convexifications of non-commutative polynomial optimisation problems. We also show that by exploiting sparsity in the convexifications, we can reduce the run time of our methods considerably. Our empirical results on a biased data set motivated by insurance applications and the well-known COMPAS data set demonstrate the efficacy of our methods.

1. Introduction

Forecasts affect almost all aspects of our daily life, as a basis for access-control mechanisms. As the quality of the forecasts impacts our lives, for better or worse, it is becoming more and more apparent that many of the tools that produce the forecasts seem to be, or indeed are, unfair, in a sense we formalise below. If the tools used to produce the forecasts are unfair, society suffers.

One such example of a forecasting tool that shapes the very pillars of our society is the *FICO Score* (Fair Isaac Corporation, 2021). FICO is a measure of an individual's creditworthiness (or conversely, credit-default risk), computed by the Fair Isaac Corporation. It has been suggested (Nickerson et al., 2016; Delis & Papadopoulos, 2019; Hegarty, 2019; Vogel, Bellet, & Clémencon, 2021) that the FICO Score may be unfair to certain minorities, although this has been disputed (Avery, Brevoort, & Canner, 2012).

As another example, consider college admissions, where results of standardised tests were often presented as forecasts of potential academic success. In the context of COVID-19, forecasting algorithms utilising previous grades and input from teachers replaced standardised tests in determining the satisfaction of college-admission requirements in many jurisdictions.

Other forecasting tools are perhaps less well known, but perhaps even more alarming, as they have the potential to shape the very core of society. One striking example is Northpointe's *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS). COMPAS is a criminal-risk assessment tool that is widely used in pretrial, parole, and sentencing decisions at courts in New York, Wisconsin, California, and Florida. COMPAS forecasts the likelihood that an individual will re-offend within two years. It has been suggested (Angwin, Larson, Mattu, & Kirchner, 2016; Dressel & Farid, 2021) that COMPAS under-predicts recidivism for Caucasian defendants, and over-predicts recidivism for African-American defendants.[1]

The applications, where fairness seems most important, often capture the behaviour of multiple subgroups of some underlying human population in the training data. Let us consider a model, where there are a number of individuals within a population $\mathcal{P}$. 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 model, $\mathcal{L}^{(s)}$. 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 model $\mathcal{L}^{(s)}$. Throughout, the superscripts distinguish the trajectories and subgroups, while subscripts indicate the periods.

In this setting, under-representation bias (Blum & Stangl, 2019, cf. Section 2.2) arises, where the trajectories of observations from one ("disadvantaged") subgroup are under-represented in the training data. This is particularly important if the forecasting is constrained to be subgroup-blind, i.e., we wish to learn a single subgroup-blind model $\mathcal{L}$. This is the case when the use of protected attributes distinguishing each subgroup can be regarded as discriminatory, such as in the case of gender and race (Gajane & Pechenizkiy, 2018; Kleinberg, Ludwig, Mullainathan, & Sunstein, 2018). Notice that such anti-discrimination measures are increasingly stipulated legally, e.g., within insurance pricing, where the sex of the applicant cannot be used, despite being known. More broadly, under-representation bias harms both the accuracy of the forecast and fairness in the sense of varying accuracy across the subgroups.

To address under-representation bias in the training of a forecasting model, it is natural to seek a notion of fairness that captures the overall behaviour across all subgroups, while taking into account the varying amounts of training data for the individual subgroups. To

---

1. We note that this has been disputed (Kleinberg, Mullainathan, & Raghavan, 2017; Dressel & Farid, 2021), and that it has been suggested that such forecasts (Neil, Sampson, & Nagin, 2021) are difficult to make, in general, due to the cohort differences in group-based arrest trajectories.

formalise this, suppose that we learn one model $\mathcal{L}$ from the multiple trajectories and define a loss function that measures the loss of accuracy for a certain observation $Y_t^{(i,s)}$ 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

$$\overset{(i,s)}{\text{loss}}(f_t) := ||Y_t^{(i,s)} - f_t||. \tag{1}$$

Let $\mathcal{T}^+ = \cup_{i \in \mathcal{I}^{(s)}, s \in \mathcal{S}} \mathcal{T}^{(i,s)}$. Following the definitions of $\mathcal{T}^{(i,s)}$, there is not an observation for $t \notin \mathcal{T}^+$, in which case, loss in Equation (1) does not exist. To evaluate the performance of the forecasts, we only consider f_t made in periods $t \in \mathcal{T}^+$. Note that, since each trajectory is of varying length, it is possible that for a certain triple (t, i, s), there is no observation $Y_t^{(i,s)}$.

Following much recent work on fairness in classification, e.g., (Zliobaite, 2015; Hardt, Price, & Srebro, 2016; Kilbertus, Carulla, Parascandolo, Hardt, Janzing, & Schölkopf, 2017; Kusner, Loftus, Russell, & Silva, 2017; Chouldechova & Roth, 2020; Aghaei, Azizi, & Vayanos, 2019), we propose two objectives to address the under-representation bias, which extend group fairness (Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 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:

$$\min_{f_t, t \in \mathcal{T}^+} \max_{s \in \mathcal{S}} \left\{ \frac{1}{|\mathcal{I}^{(s)}|} \sum_{i \in \mathcal{I}^{(s)}} \frac{1}{|\mathcal{T}^{(i,s)}|} \sum_{t \in \mathcal{T}^{(i,s)}} \overset{(i,s)}{\text{loss}}(f_t) \right\} \tag{2}$$

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

$$\min_{f_t, t \in \mathcal{T}^+} \left\{ \max_{t \in \mathcal{T}^{(i,s)}, i \in \mathcal{I}^{(s)}, s \in \mathcal{S}} \left\{ \overset{(i,s)}{\text{loss}}(f_t) \right\} \right\} \tag{3}$$

Specific Contributions: This paper builds on our preliminary work that was presented in AAAI 2021 (Zhou, Mareček, & Shorten, 2021). The present manuscript extends this work by:

- We have extended our fairness notions presented at (Zhou et al., 2021) to more generally applicable post-processing methods.

- We have added a comparison of our fairness notions with all post-processing methods in the AI Fairness 360 library, which are based on previous fairness notions of "calibrated equalised odds", "equalised odds" and "demographic parity", respectively.

With respect to the state of the art, this paper defines two new notations of fairness. We then cast the 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, which can be of independent interest. A comprehensive comparison is given to illustrate the efficacy of our approach.

2. Definitions and Related Work

Recent years have seen an unprecedented explosion in attention of notions of fairness in the field of artificial intelligence and machine learning (Ntoutsi, Fafalios, Gadiraju, Iosifidis, Nejdl, Vidal, Ruggieri, Turini, Papadopoulos, Krasanakis, et al., 2020; Chouldechova & Roth, 2020). In a typical machine learning process, the training set usually contains individuals' protected attributes (e.g., race, gender), remaining attributes, and a target variable Y. Other than not using the protected attributes ("fairness under unawareness"), several candidate definitions of fairness have been proposed and we would start from statistical notions of fairness, which could be roughly categorised into three types (Barocas, Hardt, & Narayanan, 2019): (i) independence; (ii) separation; and (iii) sufficiency; a more fine-grained discussion of fairness is presented in the sequel. The independence notion typically asks the output to be independent of protected attributes. A simple example is the "demographic parity" (Calder, Malthouse, & Schaedel, 2009), which requires each segment of a protected class (e.g., defined by gender) to receive the positive outcome at equal rates. The notion of separation, e.g., "equal odds" or "equal opportunity" in (Hardt et al., 2016), requires the predictor's output to be unrelated to protected attributes, but conditional on the target variables Y. Finally, the notion of sufficiency, derived from calibration, asked the target variables be independent from protected attributes conditional on the predictor output. For instance, we would expect the portion of defendants who were predicted to re-offend by the COMPAS system and actually re-offend to be equalised across subgroups. Calibration would require that for any given COMPAS score, the recidivism rates are similar.

The notions of independence, separation, and sufficiency are all related to subgroups of the population and provide an average guarantee for individuals in the protected group (Awasthi, Cortes, Mansour, & Mohri, 2020). In contrast, the notion of individual fairness asks for constraints that bind on specific pairs of individuals, rather than on a quantity that is averaged over groups (Chouldechova & Roth, 2020). In other words, it requires "similar individuals should be treated similarly" (Petersen, Mukherjee, Sun, & Yurochkin, 2021; Dwork, Hardt, Pitassi, Reingold, & Zemel, 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, & Roth, 2019). Apart from notions based on correlations of statistical measures and fairness, the notion of counterfactual fairness, pioneered by (Kusner et al., 2017), operates at the individual level such that causal methods are used to examine whether a decision is the same as in situations whether an individual's protected attributes are altered or not. Its generalised variant path-specific fairness in (Nilforoshan, Gaebler, Shroff, & Goel, 2022; Chiappa, 2019) specifies the effects of protected attributes along certain path in a causal directed acyclic graph. The notion of "procedural fairness", or in other words, the fairness of the decision-making process, especially in processes that resolve disputes and allocate resources, deeply rooted in legal science, has recently been tied to causal methods and fair feature selection in (Belitz, Jiang, & Bosch, 2021; Grgić-Hlača, Zafar, Gummadi, & Weller, 2018). As the many notions of fairness arise, it is necessary to build up a comprehensive framework of multiple fairness criteria, especially when there is not a widely-recognised trivial fairness notion should be used (Awasthi et al., 2020) or when certain notions are incompatible with one another (Chakraborty, Segal-Halevi, & Suk-

sompong, 2022). One could consider dynamically learning fair policies using feedback (Wen, Bastani, & Topcu, 2021; D'Amour, Srinivasan, Atwood, Baljekar, Sculley, & Halpern, 2020).

Somewhat removed from the mainstream literature, there are some excellent works in (Foulds, Islam, Keya, & Pan, 2020; Yang, Cisse, & Koyejo, 2020), who give the definition of independent subgroups, intersectional subgroups, and gerrymandering subgroups for the situations of overlapping subgroups. (Tran, Dinh, & Fioretto, 2021; Pinzón, Palamidessi, Piantanida, & Valencia, 2021; Chang & Shokri, 2021; Cummings, Gupta, Kimpara, & Morgenstern, 2019) discuss the trade-off between differential privacy (Dwork, 2006) which is an important direction for further work. An empirical study of real-world problems in (Rodolfa, Lamba, & Ghani, 2021) challenged the existence or magnitude of the trade-off accuracy between fairness. While maintaining accuracy, there exist tension between fairness and interpretability (feature deduction) (Agarwal, 2021). For the cases of unknown or ambiguous protected attributes, (Amini, Soleimany, Schwarting, Bhatia, & Rus, 2019) introduce the latent variables which may over-represent some subgroups as the proxies of unknown protected attributes. (Wang, Guo, Narasimhan, Cotter, Gupta, & Jordan, 2020) use the technologies of distributional robust optimisation to minimise the worse-case expected loss of the predictor with an upper bound on the distance between the distribution of ambiguous and actual protected attributes. (Zhang & Long, 2021; Jeong, Wang, & Calmon, 2022) discuss the impact of missing data on fairness in uniformly sampled time series, which our model of non-uniformly sampled trajectories largely avoids. Some literature in (Caton, Malisetty, & Haas, 2022; Fernando, Cèsar, David, & José, 2021) also discusses imputation strategies for fairness with respect to missing data.

Our approach to addressing the under-representation bias is rooted within the imbalanced-learning literature e.g., (He & Ma, 2013; Rolf, Worledge, Recht, & Jordan, 2021) and presents a step forward within the fairness in forecasting studied recently by (Gajane & Pechenizkiy, 2018; Chouldechova, 2017; Locatello, Abbati, Rainforth, Bauer, Schölkopf, & Bachem, 2019; Jeong et al., 2022), as outlined in the excellent survey of (Chouldechova & Roth, 2020; Barocas et al., 2019). On a more technical level, our work on fairness in learning linear dynamical systems is complemented by several recent studies involving dynamics and fairness (Mouzannar, Ohannessian, & Srebro, 2019; Paaßen, Bunge, Hainke, Sindelar, & Vogelsang, 2019; Jung, Kannan, Lee, Pai, Roth, & Vohra, 2020), and several even more recent studies on the learning of non-linear dynamics (Salmela, Tsipinakis, Foi, Billet, Dudley, & Genty, 2021; Lu, Jin, Pang, Zhang, & Karniadakis, 2021). We rely crucially on tools developed in non-commutative polynomial optimisation (Pironio, Navascués, & Acín, 2010; Wang, Magron, & Lasserre, 2021b, 2021a) and non-commutative algebra (Gelfand & Neumark, 1943; Segal, 1947; McCullough, 2001; Helton, 2002), which have not seen much use in Statistics and Machine Learning, yet.

3. Our Models

As the simplest example of the use of subgroup fairness and instantaneous fairness, cf. Equations (2) and (3), consider their applications in linear regression. For simplicity, let us assume that the cardinality of each subgroup is the same and the lengths of all trajectories

are equal. Then:

$$\min_{z,A,f_t,t\in\mathcal{T}^+} z \qquad\qquad\qquad \text{Subgroup Fairness in LR}$$

$$\text{s.t.} \quad z \geq \sum_{i\in\mathcal{I}^{(s)},t\in\mathcal{T}^+} |Y_t^{(i,s)} - f_t|, \ \forall s \in \mathcal{S}$$

$$f_t = AX_t,$$

$$\min_{z,A,f_t,t\in\mathcal{T}^+} z \qquad\qquad\qquad \text{Instantaneous Fairness in LR}$$

$$\text{s.t.} \quad z \geq |Y_t^{(i,s)} - f_t|, \ \forall i \in \mathcal{I}^{(s)}, s \in \mathcal{S}, t \in \mathcal{T}^+$$

$$f_t = AX_t,$$

where A concatenates the regression coefficients. X_t concatenates explanatory variables. f_t is the dependent variable and $Y_t^{(i,s)}$ is the actual observation in a compatible fashion. The auxiliary scalar variable z is used to reformulate "max" in the objective in Equations (2) and (3).

Next, let us consider more elaborate models, which assume that there exists a linear dynamical system (LDS) corresponding to each subgroup $s \in \mathcal{S}$. A discrete-time model of a linear dynamical system $\mathcal{L} = (G, F, V, W)$ (West & Harrison, 1997) suggests that the random variable $Y_t \in \mathbb{R}^m$ capturing the observed component (i.e., output, observations or measurements) evolves over time $t \geq 1$ according to:

$$\begin{aligned} \phi_t &= G\phi_{t-1} + w_t, \\ Y_t &= F'\phi_t + v_t, \end{aligned} \tag{4}$$

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.

The objectives in Equations (2) and (3), subject to the state-evolution and observation equations, in Equation (4), 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 multipliers $\lambda_1, \lambda_2 > 0$. The operator-valued decision variables $\mathcal{O}$ include operators F, G, vectors m_t, ω_t, and scalars f_t, ν_t, 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 in Equations (2) and (3). Since the process noise and observation noise are assumed to be samples of mean-zero normally-distributed random variables, we add the sum of squares of ω_t (resp. ν_t) to the objective with the positive multiplier λ_1 (resp. λ_2), seeking a solution with ω_t (resp. ν_t) close to zero. Overall, the subgroup-fair and instant-fair formulations

read:

$$\min_{\mathcal{O}} \quad z + \lambda_1 \sum_{t \geq 1} \omega_t^2 + \lambda_2 \sum_{t \geq 1} \nu_t^2 \qquad \text{Subgroup-Fair}$$

$$\text{s.t.} \quad z \geq \frac{1}{|\mathcal{I}^{(s)}|} \sum_{i \in \mathcal{I}^{(s)}} \frac{1}{|\mathcal{T}^{(i,s)}|} \sum_{t \in \mathcal{T}^{(i,s)}} \overset{(i,s)}{\text{loss}}(f_t), s \in \mathcal{S},$$

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

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

(5)

$$\min_{\mathcal{O}} \quad z + \lambda_1 \sum_{t \geq 1} \omega_t^2 + \lambda_2 \sum_{t \geq 1} \nu_t^2 \qquad \text{Instant-Fair}$$

$$\text{s.t.} \quad z \geq \overset{(i,s)}{\text{loss}}(f_t) \qquad , t \in \mathcal{T}^{(i,s)}, i \in \mathcal{I}^{(s)}, s \in \mathcal{S},$$

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

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

(6)

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

$$\min_{\mathcal{O}} \quad \sum_{s \in \mathcal{S}} \sum_{i \in \mathcal{I}^{(s)}} \sum_{t \in \mathcal{T}^{(i,s)}} \overset{(i,s)}{\text{loss}}(f_t) + \lambda_1 \sum_{t \geq 1} \omega_t^2 + \lambda_2 \sum_{t \geq 1} \nu_t^2 \quad \text{Unfair}$$

$$\text{s.t.} \quad m_t = G m_{t-1} + \omega_t, t \in \mathcal{T}^+,$$

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

(7)

As we explain in the Appendix, the operator-valued optimisation problems (i.e., "Unfair", "Instant-Fair", and "Subgroup-Fair") can be convexified to any given accuracy, and thence solved efficiently, under a technical assumption related to the stability of the LDS, which entails that the estimates of states and observations remain bounded, and thus all operator-valued decision variables remain bounded.

4. Numerical Illustrations

Under-representation bias considers the situation where some subgroups would be given unfair treatments, either due to the varying numbers or lengths of trajectories across subgroups. In response to under-representation bias, we have introduced two natural fairness notions for forecasting. We use a Linear Dynamic System to predict the next observation, as in Equation (4). Then, we have given two formulations associated with each notion. We tested our formulations on the famous Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset. Our implementation is available on-line at `https://github.com/Quan-Zhou/Fairness-in-Learning-of-LDS`.

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 to generate data with a given degree of bias, which is based on (Blum & Stangl, 2019, cf. Section 2.2). Suppose that there is one advantaged subgroup (a) and one disadvantaged subgroup (d), i.e., $S = \{a, d\}$, 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. 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 both subgroups are 5 and 7.

2. Observations $Y_t^{(i,s)}$ are sampled from corresponding LDS $\mathcal{L}^{(s)}$. Thus each $Y_t^{(i,s)} \sim \mathcal{L}^{(s)}$.

3. Let $\beta^{(d)}$ denote the probability that an observation from subgroup d stays in the training data, and $0 \leq \beta^{(d)} \leq 1$. It can be seen as the ratio of the number of observations in disadvantaged subgroup to that of advantaged subgroup. The degree of under-representation bias can be controlled by simply adjusting $\beta^{(d)}$. Smaller values of $\beta^{(d)}$ correspond to higher level of bias in the training set.

The last step makes the number of observations of the disadvantaged subgroup less than that of the advantaged subgroup when $0 \leq \beta^{(d)} < 1$. Hence, the advantaged subgroup becomes over-represented. Note that for a small sample size, it is necessary to make sure that there is at least one observation in each subgroup at each period.

4.2 Effects of Under-Representation Bias on Forecast

To illustrate the impact of our models on data with varying degrees of under-representation bias, suppose there is an advantaged subgroup (a) and a disadvantaged subgroup (d), i.e., $\mathcal{S} = \{a, d\}$. Figure 1 illustrates 10 experiments with general forecasting procedures. 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 curves and dashed curves, respectively. However, in each experiment, a subset of observations with the same cardinality is randomly selected and discarded and thus a new biased training set is generated, albeit based on the same "ground set" of observations. The three models in Equations (5)-(7) are applied in each experiment with λ_1 of 1, 3, and 5, respectively, as chosen by iterating over integers 1 to 10, while λ_2 remains 0.01, The mean of forecast f_t across 10 experiments and its standard deviation are shown as solid curves with error bands. The red curve gives an overview of how a prediction without considering fairness would cause an unevenly distributed prediction loss for each subgroup. This is simply because the advantaged subgroup is of larger cardinality, and the overall loss would decrease more steeply if the predicted trajectory gets closer to the advantaged subgroup.

4.3 Fairness as a Function of Bias

Figure 2 suggests how the degree of bias affects accuracy in each subgroup with and without considering fairness. With the number of trajectories in both subgroups set to two, i.e., $|I_a| = |I_d| = 2$, 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
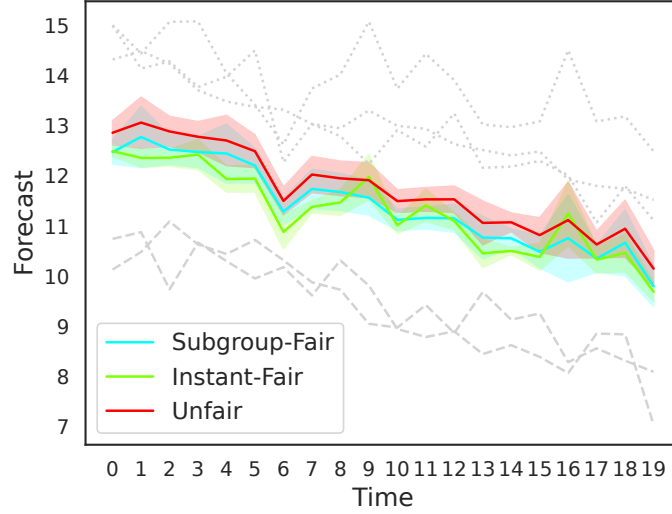
Figure 1: Forecast obtained using (5-7): the solid lines in primary colours with error bands display the mean and standard deviation of the forecasts over 10 experiments. For reference, dotted lines and dashed lines in grey denote the trajectories of observations of advantaged and disadvantaged subgroups, respectively, before discarding any observations.

error (nrmse) fitness value for each subgroup $s \in \mathcal{S}$:

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

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 this subgroup.

The training data are generated in the same way as the set of observations used in Figure 1, but with two trajectories in each subgroup ($|I_a| = |I_d| = 2$). Then, the biased training data generalisation process (described in Section 4.1) is applied in each experiment with the value of $\beta^{(d)}$ selecting from 0.5 to 0.9 at the step of 0.1. For each value of $\beta^{(d)}$, three models in Equations (5)-(7) are conducted for 10 experiments with a new biased training set in each experiment. Therefore, the quartiles of $\text{nrmse}^{(s)}$ across 10 experiments 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 ($\text{nrmse}^{(d)} \geq \text{nrmse}^{(a)}$), 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
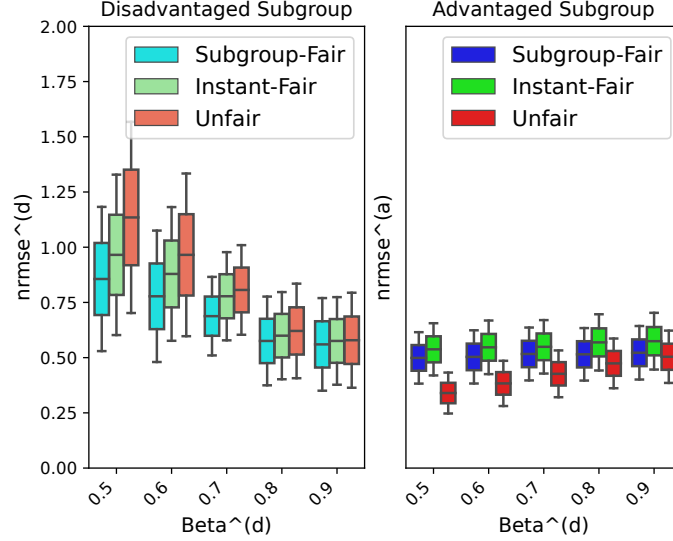
Figure 2: Accuracy as a function of the degree of under-representation bias: the boxplot of nrmse$^{(s)}$, $s \in \mathcal{S}$ against $\beta^{(d)}$, where $\beta^{(d)} = [0.5, 0.55, \ldots, 0.9]$. Each box for the quartiles of nrmse$^{(s)}$ is obtained from 10 experiments, with observations generated in the same way as the ones in Figure 1.

higher fairness between two subgroups. More surprisingly, when nrmse$^{(a)}$ decreases as $\beta^{(d)}$ gets close to 0.5, "Subgroup-Fair" model still can keep nrmse$^{(d)}$ at almost the same level, indicating a rise in overall accuracy. This is in contrast to the results of (Zliobaite, 2015; Dutta, Wei, Yueksel, Chen, Liu, & Varshney, 2019) in classification, but in line with recent work (Maity, Mukherjee, Yurochkin, & Sun, 2021).

4.4 Evaluation of Run Time of the Method

Minimising multivariate operator-valued polynomial optimisation problems (5-7) is a known non-trivial problem. We exploit sparsity-exploiting variants (TSSOS) of the globally convergent Navascués-Pironio-Acín (NPA) hierarchy used in the proof of Theorem 2, to develop fast computational methods. See (Klep, Magron, & Povh, 2022; Wang et al., 2021b, 2021a; Wang & Magron, 2021). The SDP of a given order in the respective hierarchy can be constructed using `ncpol2sdpa` 1.12.2[2] of (Wittek, 2015) or the tools of (Wang & Magron, 2021) [3] and then solved by `mosek` 9.2 of (MOSEK, ApS, 2020).

In Figure 3, we illustrate the run-time and size of the relaxations as a function of the length of the time window. Models "Subgroup-Fair" in Equation (5) and "Instant-Fair" in Equation (6) are implemented three times for each length of the time window, with the same data set used in Figure 2. The type of models, i.e., "Subgroup-Fair" (solid curves) and "Instant-Fair" (dashed curves), is distinguished by line styles. The deep-pink and cornflower-

---

2. `https://github.com/peterwittek/ncpol2sdpa`
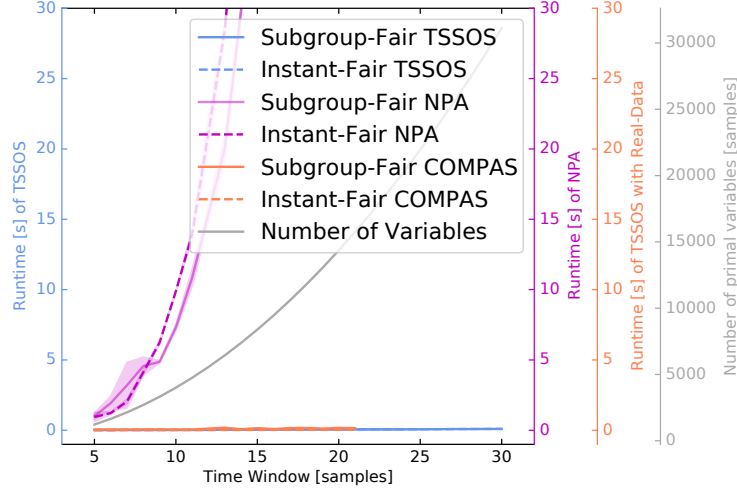
3. `https://github.com/wangjie212/TSSOS`

Figure 3: 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 TSSOS using the COMPAS dataset in Figure 4, is also displayed as coral-coloured curves. For run-time, the mean and mean ± one standard deviations across three experimental runs are presented by curves with shaded error bands.

blue curves show the run-time of the first-order SDP relaxation of NPA and the second-order SDP relaxation of TSSOS hierarchy, respectively, implemented with five CPUs and 64GB of memory per CPU. The mean and mean ± 1 standard deviation of run-time across 3 experimental runs are presented by curves with shaded error bands. The grey curve displays the number of variables in the first-order SDP relaxation of our models in Equations (5) against the length of time window. Further, models "Subgroup-Fair" in Equation (5) and "Instant-Fair" in Equation (6) are implemented once via TSSOS for each length of the time window, using COMPAS dataset, as the experiment in Figure 4, with the run-time displayed by a coral solid curve and a coral dashed curve, respectively. 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.

5. Numerical Results of 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, cf. Appendix C. The COMPAS dataset, analysed by ProPublica, comprises of defendants' gender, race, age, charge degree, COMPAS recidivism scores, two-year recidivism label, as well as information on prior incidents. The COMPAS recidivism scores, ranging from 1 to 10, are positively

related to the estimated likelihood of recidivism, given by the COMPAS system. The two-year recidivism label denotes whether a person actually got rearrested 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.

5.1 An Alternative Approach to COMPAS Dataset

From the COMPAS dataset cf. Appendix C, we choose 119 defendants with recidivism label being 1, who are either African-American or Caucasian, male, within the age range of 25-45, and with prior crime counts less than two, 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 4, dots suggest COMPAS recidivism scores of the four 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 4, the dots are far 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, Caucasian 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 re-offends 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.

5.2 A Comparison Against the State of the Art on COMPAS Dataset

Generally speaking, fairness objectives or constraints might not be easily applied to models that are already in use in applications. For such systems, revision of the model output with some post-processing tools would be a widely applicable and practical solution. Since we have shown the existence of unfairness in COMPAS recidivism scores, we now illustrate this approach to improve upon the COMPAS scores by using post-processing methods that embed our fairness notions. We then compare our methods using the AI Fairness 360 toolkit
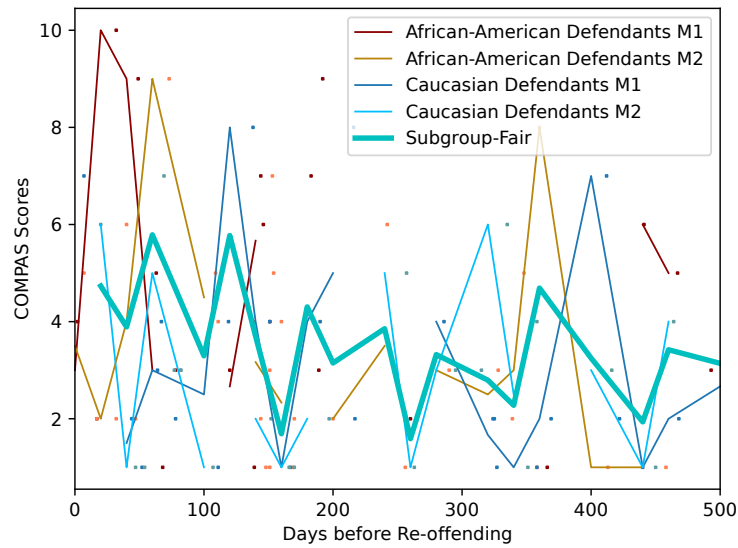
Figure 4: COMPAS recidivism scores of African-American and Caucasian defendants against the actual days before their re-offending. The sample of defendants' scores is divided into four sub-samples based on race and the type of re-offending, distinguished by colours. Dots and curves with the same colour 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.

`AIF360`[4] of (Bellamy, Dey, Hind, Hoffman, Houde, Kannan, Lohia, Martino, Mehta, Mojsilovic, Nagar, Ramamurthy, Richards, Saha, Sattigeri, Singh, Varshney, & Zhang, 2018).

The training set and test sets: The sample set contains 1005 defendants, whose race is either African-American or Caucasian, selected from the first 1200 rows of the COMPAS dataset cf. Appendix C. For a single trial, we randomly pick 80% of samples as the training set then test the output on the rest 20% of the samples. Each trial uses a new batch of the training set and the test set generated from the same sample set of 1005 defendants.

Notice that the sample set is biased as there are only 403 Caucasian defendants. Since existing data may generally contain biases, stemming for example from poor information acquisition process (Bertail, Clémençon, Guyonvarch, & Noiry, 2021), due to historical and social injustices (Ferrer, van Nuenen, Such, Coté, & Criado, 2021), we seek other methods to validate our approaches. To this end we randomly remove some observations of African-American defendants from the original test set, such that the number of defendants in both subgroups are the same. The resulted subset is called the re-weighted test set.

Performance indices: We use three baseline fairness metrics (i.e., independence, separation, and sufficiency), as well as prediction inaccuracy, to measure the performance of post-processing models. Essentially, the sample set includes two race subgroups $\mathcal{S} = \{$African-American defendants (AA), Caucasian defendants (C)$\}$. The recidivism label and the prediction of the recidivism label outputted from a model, are denoted by binary variables Y and f respectively. Let $P(f = f | Y = Y, s = s)$ be the probability of a defendant from subgroup s with recidivism label Y being predicted to recidivism label f. We set $Y = 1$ and $f = 1$ to be a defendant re-offending and being predicted to re-offend, thus they are negative events. Further, we define the indices of three baseline fairness metrics (IND, SP, SF), inaccuracy (INA) and their re-weighted versions (i.e., INDrw, SPrw, SFrw, INArw) in Equation (8):

$$
\begin{aligned}
\text{IND(rw)} &:= |P(f = 1 \mid s = \text{AA}) - P(f = 1 \mid s = \text{C})|, \\
\text{SP(rw)} &:= |P(f = 0 \mid Y = 1, s = \text{AA}) - P(f = 0 \mid Y = 1, s = \text{C})| \\
&\quad + |P(f = 1 \mid Y = 0, s = \text{AA}) - P(f = 1 \mid Y = 0, s = \text{C})|, \\
\text{SF(rw)} &:= |P(Y = 1 \mid f = 1, s = \text{AA}) - P(Y = 1 \mid f = 1, s = \text{C})| \\
&\quad + |P(f = 0 \mid Y = 0, s = \text{AA}) - P(f = 0 \mid Y = 0, s = \text{C})|, \\
\text{INA(rw)} &:= P(Y \neq f \mid s = \text{AA}) + P(Y \neq f \mid s = \text{C}),
\end{aligned}
\tag{8}
$$

where, for example, IND(rw) implies both indices IND and INDrw. The difference between IND and INDrw is that IND measures the performance of a model on the original test set while INDrw on the re-weighted test set. The same applies for SP(rw), SF(rw), INA(rw). To interpret the definitions in Equation (8): IND(rw) are race-wise absolute difference of negative rate; SP(rw) combine race-wise absolute difference of false false positive and false negative rates; SF(rw) captures the race-wise absolute difference of positive predictive value and negative predictive value; INA(rw) measure inaccuracy of test set. In our setting, smaller values of IND(rw), SP(rw), SF(rw), INA(rw) indicate better performance in terms of independence, separation, sufficiency and accuracy, respectively.

---

4. `https://github.com/Trusted-AI/AIF360`

Classification thresholds: Since the outputs of all post-processing tools implemented in this paper and COMPAS system are scores from varying intervals and, to transfer the scores to binary labels, we would need a threshold such that $f = 1$ when the score, denoted by g is higher than this threshold, and $f = 0$ otherwise. For ease of comparison, we define uni-race thresholds which differ across different models but all of them are defined as the xth percentile of all scores outputted by the corresponding model, where $x \in [0, 100]$ is fixed. Notice that there is a gap between the percentage of recidivism in African-American defendants (46%) and Caucasian defendants (59%) in terms of the sample set, and we call those percentages base rates, as in (Pleiss, Raghavan, Wu, Kleinberg, & Weinberger, 2017). In fairness to African-American defendants, we introduce race-wise thresholds using base rates: for each model, the percentage of defendants in a subgroup whose scores are higher than the subgroup's threshold needs to be the same as the subgroup's base rate.

Post-processing methods: Associated with our fairness notions, we propose two post-processing methods. Both methods use simple race-wise linear regression models

$$g^{(i,s)} = A^{(s)} X^{(i,s)} + e^{(s)}, i \in \mathcal{I}^{(s)}, s \in \mathcal{S}, \tag{9}$$

where the subscript t is removed such that we only consider prediction in one period. In other words, we cast the problem of prediction into classification. $A^{(s)}$ concatenates the regression coefficients. $X^{(i,s)}$ concatenates explanatory variables, including COMPAS recidivism score, prior incidents (i.e., the sum of "prior counts", "juv_ fel_ count" and "juv_ misd_ count"), age category (i.e., 1 if age is less than 25 and 0 otherwise), and recidivism label. $e^{(s)}$ corresponds to a noise to the linear relationship. $g^{(i,s)}$ is the post-processed recidivism score of the defendant i in subgroup s, and $Y^{(i,s)}$ is the actual recidivism label (i.e., the ground truth). Let $\text{loss}^{(i,s)}(g) := \|Y^{(i,s)} - g^{(i,s)}\|$, our post-processing methods are Equation (10) subject to Equation (9), with $\lambda_3 = 0.05$. Further, the score $g^{(i,s)}$ would be mapped to the binary prediction of recidivism label $f^{(i,s)}$ using a threshold.

$$
\begin{aligned}
\textbf{Subgroup-Fair} \quad &\min_{g,A,e} \left\{ \max_{s \in \mathcal{S}} \left\{ \frac{1}{|\mathcal{I}^{(s)}|} \sum_{i \in \mathcal{I}^{(s)}} \text{loss}^{(i,s)}(g) \right\} + \lambda_3 \sum_{s \in \mathcal{S}} \left(e^{(s)} \right)^2 \right\} \\
\textbf{Instant-Fair} \quad &\min_{g,A,e} \left\{ \max_{i \in \mathcal{I}^{(s)}, s \in \mathcal{S}} \left\{ \text{loss}^{(i,s)}(g) \right\} + \lambda_3 \sum_{s \in \mathcal{S}} \left(e^{(s)} \right)^2 \right\}
\end{aligned} \tag{10}
$$

In Figure 5, we test the performance of all post-processing methods implemented in AI Fairness 360 toolkit:

- "AIF360": calibrated equalised odds post-processing with cost constraint being a combination of both false negative rate and false positive rate, as suggested by the authors of the AI Fairness 360 toolkit (Bellamy et al., 2018),

- "CaliEqOdds(fnr)": calibrated equalised odds post-processing with cost constraint being the false negative rate,

- "CaliEqOdds(fpr)": calibrated equalised odds post-processing with cost constraint being the false positive rate,

- "EqOdds": equalised odds post-processing,

- "RejectOption": reject option classification.

Note that "AIF360", "CaliEqOdds(fnr)", "CaliEqOdds(fpr)" are based on the fairness notion of "calibrated equalised odds" in (Pleiss et al., 2017). "EqOdds" is derived from the fairness notion of "equalised odds" in (Hardt et al., 2016). "RejectOption" comes from (Kamiran, Karim, & Zhang, 2012), which is rooted in the fairness notion of "demographic parity" (Calder et al., 2009). Those five methods are implemented in five trials for each of three uni-race thresholds $x = [47, 53, 60]$ ($5 \times 5 \times 3$ runs). The left subplot displays mean values of eight indices across all five trials and three thresholds, with each angular axis representing one index, and each colour denoting one post-processing method. The right subplot represents the values of all experimental runs as dots in a circular sector, with each sector representing one index. Each sector is labelled with the index immediately counter-clockwise to it. For instance, the sector between the labels "IND" and "INDrw" displays the "IND" values of all experimental runs. In both subplots, the value represented by a dot is displayed by its distance from the original point, with shorter distances indicating better performance.

Figure 5 depicts a summary of the state of the art on COMPAS dataset, and how we select the appropriate method to benchmark our own algorithms. Referring to this figure, since in both subplots, most of yellow ("AIF360") dots are relatively closer to the origin than other dots, it seems fair to consider "AIF360" as the state of the art post-processing method, at least within those implemented in AI Fairness 360 toolkit, and to compare our methods against it in the following.

In Figure 6, COMPAS scores ("COMPAS" red), the state of the art ("AIF360", yellow), and the outputs of our methods "Subgroup-Fair" (blue) and "Instant-Fair" (green) are evaluated across 50 trials using base rates as race-wise thresholds (50×4 runs). The left subplot illustrates the average performance of four methods, where dots represent the mean values of original indices, and bars are those of re-weighted indices. The right subplot displays fairness performance of all experimental runs in a triangular area. For a single run, the original fairness metrics (i.e., IND, SP, SF) are denoted by one dot and re-weighted ones (i.e., INDrw, SPrw, SFrw) are shown as one square. A marker, that is, a dot or a square, represents the value of IND(rw), SP(rw), SF(rw), by its positions along the left, right, and bottom axes in ternary coordinates. The colour of this marker denotes the method used in this run.

Figure 6 illustrates the performance of our methods compared with the state of the art and COMPAS scores when using base rates as race-wise thresholds. As we can see on the left, there is not much difference between the values of original indices and their re-weighted versions, except that "Instant-Fair" generally performs worse in re-weighted version than in original version. It implies that "Instant-Fair" might not be appropriate to use in this case because its performance varies with the test set being re-weighted or not. The performance of "Subgroup Fair" is similar to that of "COMPAS", but with a slight improvement in IND(rw) and SP(rw). "AIF360" seems to sacrifice a lot of accuracy for lower average fairness indices, while its fairness performance shows a lot of variability, as shown in the right subplot. On the contrary, the concentration of blue and green markers (i.e., dots and squares) indicates less variability of our methods.

In Figure 7, COMPAS scores ("COMPAS" red), the state of the art ("AIF360", yellow), and the outputs of our methods "Subgroup-Fair" (blue) and "Instant-Fair" (green) are evaluated across 50 trials, with 10 different uni-race thresholds $x = [20, 27, \ldots, 80]$ ($50 \times 10 \times 4$ runs). Each subplot represents the mean (curves) and mean $\pm$ one standard deviations
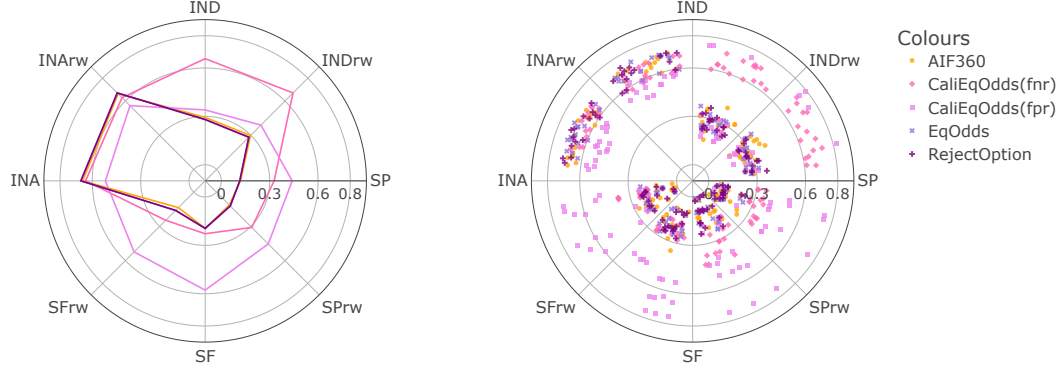
Figure 5: The state of the art in post-processing for improving fairness. Five post-processing methods implemented in AI Fairness 360 toolkit, i.e., "AIF360" (yellow), "CaliEqOdds(fnr)" (pink), "CaliEqOdds(fpr)" (violet), "EqOdds" (light purple), and "RejectOption" (purple), are run in five trials per each of three uni-race thresholds $x = [47, 53, 60]$ ($5 \times 5 \times 3$ runs). Left: The mean values of eight indices on eight angular axes. Right: Each sector is labelled with the index immediately counter-clockwise to it. Dots in one sector denote the values of the corresponding index of all experimental runs.

(shaded error bands) of the corresponding index across 50 trials, against 10 different uni-race thresholds $x = [20, 27, \ldots, 80]$ with four methods distinguished by the same palette as in Figure 6.

Figure 7 depicts an investigation of the performance of "Subgroup-Fair", "Instant-Fair", "AIF360" and "COMPAS" for different uni-race thresholds. Notice that the error bands generally overlap each other in subplots of the first and third rows (which represent the indices IND(rw) and SF(rw)). One potential implication of this work is that there might not be significant differences amongst the four methods, in terms of IND(rw) and SF(rw), when using uni-race thresholds. If we look at the remaining indices, "Subgroup-Fair" (blue) surpasses "COMPAS" (red) in terms of the performance indexes SP(rw), and both achieve the best in INA(rw). Furthermore, "AIF360" has relatively low values of SP(rw), but at a large expense of INA(rw).
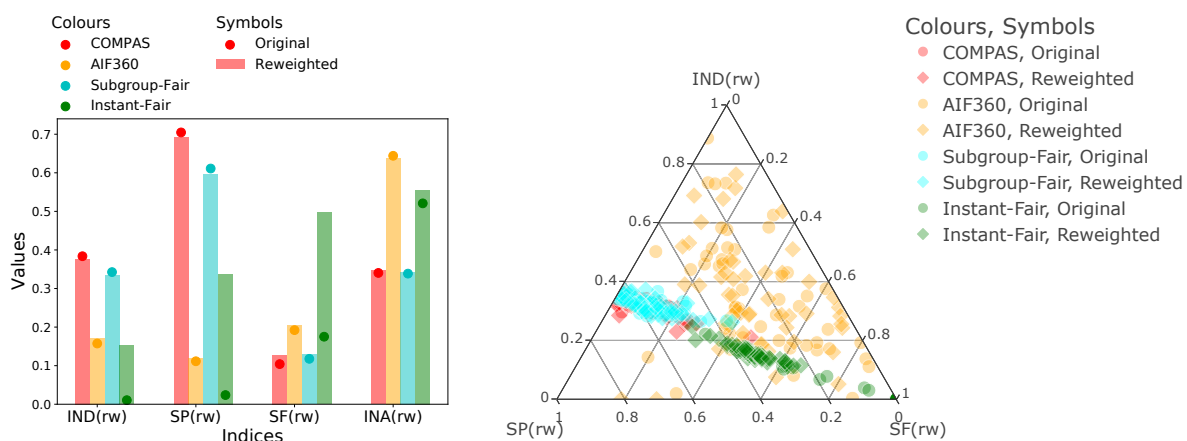
Figure 6: The effects of our methods, when applied as post-processing. Three post-processing methods "subgroup-fair" (blue), "instant-fair" (green), "AIF360" (yellow) are run in 50 trials, and their outputs are compared with the COMPAS scores (red) with race-wise thresholds being base rates. Left: The mean values of original (dots) and re-weighted (bars) indices associated with outputs from four models. Right: The original (dots) and re-weighted (squares) fairness metrics (i.e., INDrw, SPrw, SFrw) of all experimental runs.
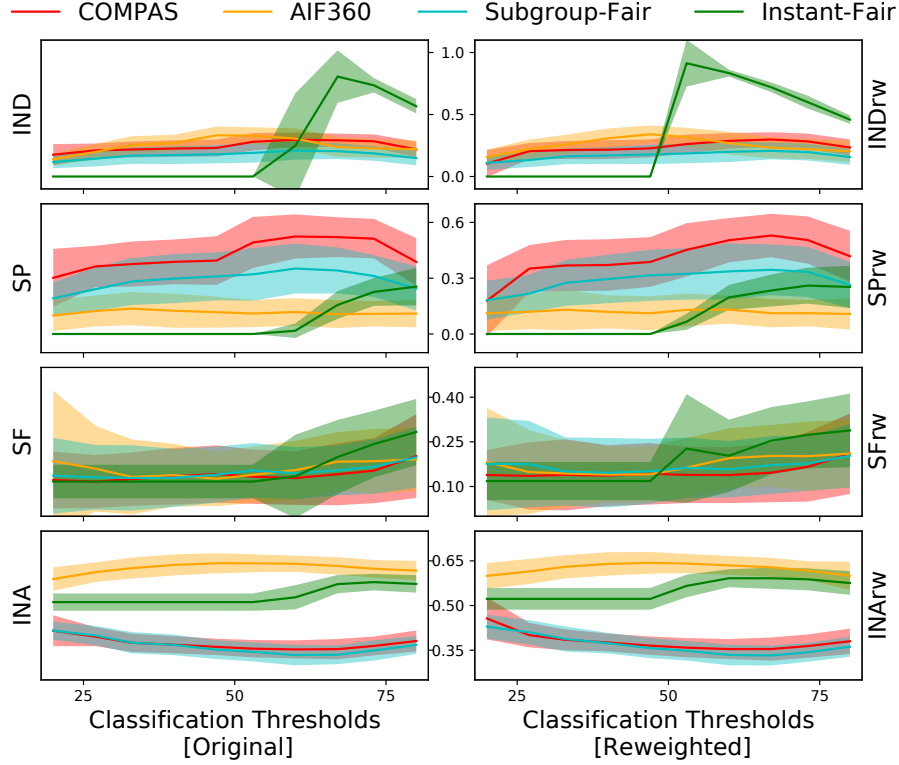
Figure 7: Three post-processing methods "subgroup-fair" (blue), "instant-fair" (green), "AIF360" (yellow) are run in 50 trials, and their outputs are compared with the COMPAS scores (red), with 10 uni-race thresholds $x = [20, 26.7, \ldots, 80]$. Each subplot displays the mean (curves) and mean $\pm$ one standard deviations (shaded error bands) of the corresponding index across the 50 trials, against 10 uni-race thresholds.

6. Conclusions

We have introduced the two natural notions of fairness in forecasting. When the corresponding optimisation problems are solved to global optimality, the solutions outperform the COMPAS system in terms of independence and separation indices IND(rw) and SP(rw), cf. Equation (8).

As a further technical contribution, we have presented globally convergent methods for solving the optimisation problems arising from the two notions of fairness using hierarchies of convexifications of non-commutative polynomial optimisation problems. Also, we have shown that the run-time of standard solvers for the convexifications is independent of the dimension of the hidden state. This provides a technical tool in machine learning and statistics that is of independent interest, and that can also be applied in other settings.

Acknowledgments

Appendix A. Motivation

A.1 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 behaviour 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 & Salm, 2020). The insurer will hence pay more to a female insured during her lifetime, compared to a male insured, on average (Thiery & 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 & 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 tth 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 Equation (A.1), thus summing up the present values of payment in each year and then divided by the number of insureds in the beginning.

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

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 tth year, respectively. Both trajectories can be regarded as linear dynamic systems. We have

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

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:

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

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:

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

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 Eq.(2)-Eq.(3)has the form:

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

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

A.2 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 $S = \{\text{female, male}\}$ 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 ith customer in subgroup s. These evolve as:

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

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:

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

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 Eq.(2)-Eq.(3) have the form:

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

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

We also refer to (Dong, Miehling, & Langbort, 2020) for further work on protecting customers' interests in personalised pricing via fairness considerations.

Appendix B. Background

In this paper, we 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 B, we firstly set our work in the context of system identification and control theory. Secondly, we introduce the concept of fairness, which can be used to deal with multiple variants of the LDS. In the end of this section, we provide a brief overview of non-commutative polynomial optimisation, pioneered by (Pironio et al., 2010) and nicely surveyed by (Burgdorf, Klep, & Povh, 2016), which is our key technical tool.

B.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 (Ljung, 1998) and (Tangirala, 2014) for excellent overviews of the long history of research in the field, going back at least to (Åström & Torsten, 1965). 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 (Anava, Hazan, Mannor, & Shamir, 2013), first guarantees were presented for auto-regressive moving-average (ARMA) processes. In (Liu, Hoi, Zhao, & Sun, 2016), these results were extended to a subset of autoregressive integrated moving average (liu2016online) processes. (Kozdoba, Marecek, Tchrakian, & Mannor, 2019) 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 (Tsiamis & Pappas, 2020) to Kalman filtering with logarithmic regret. Another stream of work within improper learning focuses on sub-space methods (Katayama, 2006; Van Overschee & De Moor, 2012) and spectral methods. (Tsiamis, Matni, & Pappas, 2020; Tsiamis & Pappas, 2019) presented the present-best guarantees for traditional sub-space methods. Within spectral methods, (Hazan, Singh, & Zhang, 2017) and (Hazan, Lee, Singh, Zhang, & Zhang, 2018) 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 (Faradonbeh, Tewari, & Michailidis, 2018). (Simchowitz, Mania, Tu, Jordan, & Recht, 2018) 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. (Simchowitz, Boczar, & Recht, 2019) extended the results to allow for a certain pre-filtering procedure. (Sarkar & Rakhlin, 2019) 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 (Vazirani, 2013), as it allows for ϵ error for any $\epsilon > 0$.

B.2 Learning from Imbalanced Data

Traditional machines learning algorithms can be biased towards majority class over-prevalence (Chawla, 2003), i.e., the under-representation bias (Blum & Stangl, 2019). Also, the cost of mis-classifying an abnormal event (minority class) as a normal event (majority class) is often relatively high (Chawla, Cieslak, Hall, & Joshi, 2008; Chawla, Bowyer, Hall, & Kegelmeyer, 2002). For example, in the case of fraud, diseases, those cases are rare but able to cause serious damages, so it is of great interest to research. The benchmark of learning from imbalanced data was pioneered by (Chawla et al., 2002). They proposed the Synthetic Minority Over-sampling Technique (SMOTE), such that a combination of over-sampling the minority class and under-sampling the majority class can efficiently improve the classifier performance.

The research in learning from imbalanced data has been extensively studied with a particular focus on classification and other predictive contexts as many real-world applications are already facing this problem (Torgo, Krawczyk, Branco, & Moniz, 2017; Fernández, Garcia, Herrera, & Chawla, 2018). SMOTE has been successfully extended to a variety of applications because of its simplicity and robustness (Cieslak & Chawla, 2008). Surprisingly, (Chawla et al., 2008) provides an algorithm that automatically discovers the amount of re-sampling. One the other hand, (Moniz, Branco, & Torgo, 2017) proposed the concept of temporal and relevance bias in extension of re-sampling strategies. For the clear journey of SMOTE, please refer to (Fernández et al., 2018).

Unlike the common solution of re-sampling, we address the under-representation bias from the view of optimisation, such that the "loss", or other statistical performance is equalised over majority and minority subgroups.

B.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 (Pironio et al., 2010). Here, we provide a brief summary of their results, and refer to (Burgdorf et al., 2016) for a book-length introduction. NCPOP is an operator-valued optimisation problem with a standard form in Equation (11):

$$
P: \quad
\begin{aligned}
p* = \min_{(\mathcal{H}, \mathbf{X}, \phi)} \quad & \langle \phi, p(\mathbf{X})\phi \rangle \\
\text{s.t.} \quad & q_i(\mathbf{X}) \succcurlyeq 0, i = 1, \ldots, m, \\
& \langle \phi, \phi \rangle = 1,
\end{aligned}
\tag{11}
$$

where $\mathbf{X} = (X_1, \ldots, X_n)$ is a n-tuple of bounded operators on a Hilbert space $\mathcal{H}$ in this section. 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. $p(\mathbf{X})$ and $q_i(\mathbf{X})$ are polynomials and $q_i(\mathbf{X}) \succcurlyeq 0$ denotes that the operator $q_i(\mathbf{X})$ is positive semi-definite.

In contrast to traditional scalar-valued, vector-valued, or matrix-valued optimisation techniques, the dimension of operators $\mathbf{X}$ is unknown *a priori*. Let $[\mathbf{X}, \mathbf{X}^\dagger]$ denotes these $2n$ operators, with the †-algebra being conjugate transpose. Monomials ω, μ and ν in following text are products of powers of variables from $[\mathbf{X}, \mathbf{X}^\dagger]$. The degree of a monomial, denoted by $|\omega|$, refers to the sum of the exponents of all operators in the monomial ω. Let $\mathcal{W}_k$

denote the collection of all monomials whose degrees $|\omega| \leq k$. Polynomials $p(\mathbf{X})$ and $q_i(\mathbf{X})$ of degrees $\deg(p)$ and $\deg(q_i)$, respectively, can be written as:

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

where $i = 1, \ldots, m$. Following (Akhiezer & Krein, 1962), we can define the moments on field $\mathbb{R}$ or $\mathbb{C}$, with a feasible solution $(\mathcal{H}, \mathbf{X}, \phi)$ of problem in Equation (11):

$$y_\omega = \langle \phi, \omega \phi \rangle,$$

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 $M_k(y)$:

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

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

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

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 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 $(\mathcal{H}, \mathbf{X}, \phi)$ is feasible, one can utilise the Sums-of-Squares theorem of (Helton, 2002) and (McCullough, 2001) to derive semidefinite programming (SDP) relaxations. In particular, we can obtain a kth order SDP relaxation of the non-commutative polynomial optimisation problem in Equation (11) 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:

$$
R_k : \quad
\begin{aligned}
p^k = &\min_{y = (y_\omega)_{|\omega| \leq 2k}} \sum_{|\omega| \leq d} p_\omega y_\omega \\
\text{s.t.} \quad & M_k(y) \succcurlyeq 0, \\
& M_{k-d_i}(q_i y) \succcurlyeq 0, i = 1, \ldots, m, \\
& \langle \phi, \phi \rangle = 1.
\end{aligned}
\quad (12)
$$

Let us define the quadratic module, following (Pironio et al., 2010). Let $Q = \{q_i, i = 1, \ldots, m\}$ be the set of polynomials determining the constraints. The *positivity domain* $\mathbf{S}_Q$ of Q are n-tuples of bounded operators $\mathbf{X}$ on a Hilbert space $\mathcal{H}$ making all $q_i(\mathbf{X})$ positive semidefinite. The *quadratic module* $\mathbf{M}_Q$ 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 in the $2n$ operators in $[\mathbf{X}, \mathbf{X}^\dagger]$. As in (Pironio et al., 2010), if the Archimedean assumption is satisfied, $\lim_{k \to \infty} p^k = p^*$ for a finite k.

B.4 The Formal Statement

In order to utilise subgroup fairness or instantaneous fairness, one needs to be able to guarantee that global optima for the corresponding optimisation problems can be found. Such guarantees are non-trivial in the case of non-convex, non-commutative optimisation problems. Following (Zhou & Mareček, 2020), one can formalise the guarantees we provide:

Assumption 1 (Archimedean). *Quadratic module* $\mathbf{M}_Q$ *of Eq.(5) or Eq.(6) is Archimedean, i.e., there exists a real constant* C *such that* $C^2 - (X_1^\dagger X_1 + \cdots + X_{2n}^\dagger X_{2n}) \in \mathbf{M}_Q$, *for these* $2n$ *operators in* $[\mathbf{X}, \mathbf{X}^\dagger]$.

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 Eq.(5), 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 Eq.(6).*

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. As explained in the subsection above, (Pironio et al., 2010) 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 (Pironio et al., 2010), which requires Assumption 1.

Notice that we can use the so-called rank-loop condition of (Pironio et al., 2010) to detect global optimality. Once optimality is detected, it is possible to extract the global optimum $(\mathcal{H}^*, \mathbf{X}^*, \phi^*)$ from the optimal solution y of problem R_k, by Gram decomposition; cf. Theorem 2 in (Pironio et al., 2010). Simpler procedures for the extraction have been considered, cf. (Henrion & Lasserre, 2005), but remain less well understood.

More broadly, we would like to show 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 (Gelfand & Neumark, 1943; Segal, 1947), as explained in Section 2.2 of (Klep, Povh, & Volcic, 2018), which does not require the rank-loop condition to be satisfied, We refer to Section 2.2 of (Klep et al., 2018) and Section 2.6 of (Dixmier, 1969) for details. □

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 (Pironio et al., 2010). To do so, we need to have the non-commutative versions of those formulations firstly, thus introduce a a Hilbert space $\mathcal{H}$ with all operators and a normalised vector defined on $\mathcal{H}$.

Appendix C. The COMPAS dataset

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's

likelihood of re-offending (recidivism). It has been shown that the algorithm is biased in favour of white defendants, based on a 2 year follow-up study (i.e., who actually committed crimes or violent crimes after 2 years).

Downloaded from `https://github.com/propublica/compas-analysis`, this dataset is what (Angwin et al., 2016) used in analysing the racial bias in COMPAS recidivism scores. The COMPAS dataset comprises of defendants' gender, race, age, charge degree, COMPAS recidivism scores, two-year recidivism label, as well as information on prior incidents. The COMPAS recidivism scores, ranging from 1 to 10, are positively related to the estimated likelihood of recidivism, given by the COMPAS system. The two-year recidivism label denotes whether a person actually got rearrested 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.

References

Åström, K.-J., & Torsten, B. (1965). Numerical identification of linear dynamic systems from normal operating records. *IFAC Proceedings Volumes*, *2*(2), 96–111. 2nd IFAC Symposium on the Theory of Self-Adaptive Control Systems, Teddington, UK, September 14-17, 1965.

Abdou, D. S. (2019). Gender-based price discrimination: The cost of being a woman. *Proceedings of Business and Economic Studies*, *2*(5).

Agarwal, S. (2021). Trade-offs between fairness and interpretability in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*.

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

Akhiezer, N. I., & Krein, M. (1962). *Some questions in the theory of moments*, Vol. 2. American Mathematical Society, Providence, RI.

Amini, A., Soleimany, A. P., Schwarting, W., Bhatia, S. N., & Rus, D. (2019). Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 289–295.

Anava, O., Hazan, E., Mannor, S., & Shamir, O. (2013). Online learning for time series prediction. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica, May*, *23*, 2016.

Avery, R. B., Brevoort, K. P., & Canner, G. (2012). Does credit scoring produce a disparate impact?. *Real Estate Economics*, *40*, S65–S114.

Awasthi, P., Cortes, C., Mansour, Y., & Mohri, M. (2020). Beyond individual and group fairness. *CoRR*, *abs/2008.09490*.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairml-book.org, `http://www.fairmlbook.org`.

Belitz, C., Jiang, L., & Bosch, N. (2021). Automating procedurally fair feature selection in machine learning. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 379–389.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias..

Bertail, P., Clémençon, S., Guyonvarch, Y., & Noiry, N. (2021). Learning from biased data: A semi-parametric approach. In Meila, M., & Zhang, T. (Eds.), *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *Proceedings of Machine Learning Research*, pp. 803–812. PMLR.

Blum, A., & Stangl, K. (2019). Recovering from biased data: Can fairness constraints improve accuracy?. *arXiv preprint arXiv:1912.01094, abs/1912.01094*.

Burgdorf, S., Klep, I., & Povh, J. (2016). *Optimization of polynomials in non-commuting variables*. Springer.

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

Caton, S., Malisetty, S., & Haas, C. (2022). Impact of imputation strategies on fairness in machine learning. *Journal of Artificial Intelligence Research, 74*, 1011–1035.

Chakraborty, M., Segal-Halevi, E., & Suksompong, W. (2022). Weighted fairness notions for indivisible items revisited. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, pp. 4949–4956.

Chang, H., & Shokri, R. (2021). On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 292–303. IEEE.

Chawla, N. V. (2003). C4.5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In *Proceedings of the ICML*, Vol. 3, p. 66.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321–357.

Chawla, N. V., Cieslak, D. A., Hall, L. O., & Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery, 17*(2), 225–252.

Chiappa, S. (2019). Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 7801–7808.

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

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

Cieslak, D. A., & Chawla, N. V. (2008). Learning decision trees for unbalanced data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 241–256. Springer.

Cummings, R., Gupta, V., Kimpara, D., & Morgenstern, J. (2019). On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pp. 309–315.

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., & Halpern, Y. (2020). Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 525–534.

Delis, M. D., & Papadopoulos, P. (2019). Mortgage lending discrimination across the us: New methodology and new evidence. *Journal of Financial Services Research, 56*(3), 341–368.

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

Dong, R., Miehling, E., & Langbort, C. (2020). Protecting consumers against personalized pricing: A stopping time approach. *arXiv preprint arXiv:2002.05346, abs/2002.05346*.

Dressel, J., & Farid, H. (2021). The dangers of risk prediction in the criminal justice system..

Dutta, S., Wei, D., Yueksel, H., Chen, P.-Y., Liu, S., & Varshney, K. R. (2019). An information-theoretic perspective on the relationship between fairness and accuracy. *ArXiv, abs/1910.07870*.

Dwork, C. (2006). Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., & Wegener, I. (Eds.), *Automata, Languages and Programming*, pp. 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

Fair Isaac Corporation (2021). About.. https://www.ficoscore.com/about.

Faradonbeh, M. K. S., Tewari, A., & Michailidis, G. (2018). Finite time identification in unstable linear systems. *Automatica, 96*, 342–353.

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

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research, 61*, 863–905.

Fernando, M.-P., Cèsar, F., David, N., & José, H.-O. (2021). Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems, 36*(7), 3217–3258.

Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and discrimination in ai: a cross-disciplinary perspective. *IEEE Technology and Society Magazine*, *40*(2), 72–80.

Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. (2020). An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1918–1921. IEEE.

Gajane, P., & Pechenizkiy, M. (2018). On formalizing fairness in prediction with machine learning. In Friedler, S. A., & Wilson, C. (Eds.), *Conference on Fairness, Accountability and Transparency, FAT 2018*, Vol. 81 of *Proceedings of Machine Learning Research*, New York, NY, USA. PMLR.

Gelfand, I., & 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.

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

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

Hazan, E., Lee, H., Singh, K., Zhang, C., & Zhang, Y. (2018). Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, pp. 4634–4643.

Hazan, E., Singh, K., & Zhang, C. (2017). Learning linear dynamical systems via spectral filtering. In *Advances in Neural Information Processing Systems*, pp. 6702–6712.

He, H., & Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley, Germany.

Hegarty, P. (2019). Inequality brokered. *Proceedings of the National Academy of Sciences*, *116*(19), 9152–9154.

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

Henrion, D., & Lasserre, J.-B. (2005). Detecting global optimality and extracting solutions in gloptipoly. In *Positive polynomials in control*, pp. 293–310. Springer, Berlin-Heidelberg-New York.

Huang, S., & 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.

Jeong, H., Wang, H., & Calmon, F. P. (2022). Fairness without imputation: A decision tree approach for fair prediction with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, pp. 9558–9566.

Jung, C., Kannan, S., Lee, C., Pai, M., Roth, A., & Vohra, R. (2020). Fair prediction with endogenous behavior. In *Proceedings of the 21st ACM Conference on Economics and Computation*, EC '20, p. 677–678, New York, NY, USA. Association for Computing Machinery.

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929. IEEE.

Katayama, T. (2006). *Subspace methods for system identification.* Springer Science & Business Media, London, UK.

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

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis*, *10*, 113–174.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017).* Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Klep, I., Magron, V., & Povh, J. (2022). Sparse noncommutative polynomial optimization. *Mathematical Programming*, *193*(2), 789–829.

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

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

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

Liu, C., Hoi, S. C., Zhao, P., & Sun, J. (2016). Online arima algorithms for time series prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30(1).

Ljung, L. (1998). *System Identification: Theory for the User.* Pearson Education, New Jersey, USA.

Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., & Bachem, O. (2019). On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems 32*, pp. 14611–14624. Curran Associates, Inc., Vancouver, Canada.

Lu, L., Jin, P., Pang, G., Zhang, Z., & Karniadakis, G. E. (2021). Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, *3*(3), 218–229.

Maity, S., Mukherjee, D., Yurochkin, M., & Sun, Y. (2021). Does enforcing fairness mitigate biases caused by subpopulation shift?. *Advances in Neural Information Processing Systems*, *34*.

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

Moniz, N., Branco, P., & Torgo, L. (2017). Resampling strategies for imbalanced time series forecasting. *International Journal of Data Science and Analytics*, *3*(3), 161–181.

MOSEK, ApS (2020). *The MOSEK Optimizer API for Python 9.2.*

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

Neil, R., Sampson, R. J., & Nagin, D. S. (2021). Social change and cohort differences in group-based arrest trajectories over the last quarter-century. *Proceedings of the National Academy of Sciences*, *118*(31).

Nickerson, D., et al. (2016). Asset price volatility, credit rationing and rational lending discrimination. *International Journal of Economics and Finance*, *8*(10), 140–158.

Nilforoshan, H., Gaebler, J. D., Shroff, R., & Goel, S. (2022). Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, pp. 16848–16887. PMLR.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., et al. (2020). Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1356.

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., & 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)*.

Petersen, F., Mukherjee, D., Sun, Y., & Yurochkin, M. (2021). Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, *34*, 25944–25955.

Pinzón, C., Palamidessi, C., Piantanida, P., & Valencia, F. (2021). On the impossibility of non-trivial accuracy under fairness constraints. *arXiv preprint arXiv:2107.06944*, *abs/2107.06944*.

Pironio, S., Navascués, M., & Acín, A. (2010). Convergent relaxations of polynomial optimization problems with noncommuting variables. *SIAM Journal on Optimization*, *20*(5), 2157–2180.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in neural information processing systems*, *30*.

Rodolfa, K. T., Lamba, H., & Ghani, R. (2021). Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, *3*(10), 896–904.

Rolf, E., Worledge, T. T., Recht, B., & Jordan, M. (2021). Representation matters: Assessing the importance of subgroup allocations in training data. In Meila, M., & Zhang, T. (Eds.), *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *Proceedings of Machine Learning Research*, pp. 9040–9051. PMLR.

Salmela, L., Tsipinakis, N., Foi, A., Billet, C., Dudley, J. M., & Genty, G. (2021). Predicting ultrafast nonlinear dynamics in fibre optics with a recurrent neural network. *Nature Machine Intelligence*, *3*(4), 344–354.

Sarkar, T., & Rakhlin, A. (2019). Near optimal finite time identification of arbitrary linear dynamical systems. In Chaudhuri, K., & Salakhutdinov, R. (Eds.), *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 5610–5618, Long Beach, California, USA. PMLR.

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

Sharifi-Malvajerdi, S., Kearns, M., & Roth, A. (2019). Average individual fairness: Algorithms, generalization and experiments. In *Advances in Neural Information Processing Systems*, pp. 8240–8249.

Simchowitz, M., Boczar, R., & Recht, B. (2019). Learning linear dynamical systems with semi-parametric least squares. In Beygelzimer, A., & Hsu, D. (Eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, Vol. 99 of *Proceedings of Machine Learning Research*, pp. 2714–2802. PMLR.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., & Recht, B. (2018). Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pp. 439–473.

Tangirala, A. K. (2014). *Principles of system identification: theory and practice*. Crc Press, USA.

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

Torgo, L., Krawczyk, B., Branco, P., & Moniz, N. (2017). Learning with imbalanced domains: Preface. In *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pp. 1–6. PMLR.

Tran, C., Dinh, M., & Fioretto, F. (2021). Differentially private empirical risk minimization under the fairness lens. *Advances in Neural Information Processing Systems*, *34*.

Tsiamis, A., Matni, N., & Pappas, G. (2020). Sample complexity of kalman filtering for unknown systems. In Bayen, A. M., Jadbabaie, A., Pappas, G., Parrilo, P. A., Recht, B., Tomlin, C., & Zeilinger, M. (Eds.), *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, Vol. 120 of *Proceedings of Machine Learning Research*, pp. 435–444. PMLR.

Tsiamis, A., & Pappas, G. (2020). Online learning of the kalman filter with logarithmic regret. *arXiv preprint arXiv:2002.05141, abs/2002.05141*.

Tsiamis, A., & Pappas, G. J. (2019). Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 3648–3654.

Van Overschee, P., & De Moor, B. (2012). *Subspace identification for linear systems: Theory — Implementation — Applications*. Kluwer, Dordrecht, NL.

Vazirani, V. V. (2013). *Approximation algorithms*. Springer Science & Business Media, Berlin-Heidelberg, DE.

Vogel, R., Bellet, A., & Clémencon, S. (2021). Learning fair scoring functions: Bipartite ranking under roc-based fairness constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 784–792. PMLR.

Wang, J., & Magron, V. (2021). Exploiting term sparsity in noncommutative polynomial optimization. *Computational Optimization and Applications*, *80*(2), 483–521.

Wang, J., Magron, V., & Lasserre, J.-B. (2021a). Chordal-tssos: a moment-sos hierarchy that exploits term sparsity with chordal extension. *SIAM Journal on Optimization*, *31*(1), 114–141.

Wang, J., Magron, V., & Lasserre, J.-B. (2021b). Tssos: A moment-sos hierarchy that exploits term sparsity. *SIAM Journal on Optimization*, *31*(1), 30–58.

Wang, S., Guo, W., Narasimhan, H., Cotter, A., Gupta, M., & Jordan, M. (2020). Robust optimization for fairness with noisy protected groups. *Advances in Neural Information Processing Systems*, *33*, 5190–5203.

Wen, M., Bastani, O., & Topcu, U. (2021). Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*, pp. 1144–1152. PMLR.

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

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.

Yang, F., Cisse, M., & Koyejo, S. (2020). Fairness with overlapping groups; a probabilistic perspective. *Advances in neural information processing systems*, *33*, 4067–4078.

Zhang, Y., & Long, Q. (2021). Assessing fairness in the presence of missing data. *Advances in Neural Information Processing Systems*, *34*.

Zhou, Q., & Mareček, J. (2020). Proper learning of linear dynamical systems as a noncommutative polynomial optimisation problem. *arXiv preprint arXiv:2002.01444*, *abs/2002.01444*.

Zhou, Q., Mareček, J., & Shorten, R. N. (2021). Fairness in forecasting and learning linear dynamical systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 11134–11142.

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*.