Fair Policy Learning

Tennison Liu

Supervisor: Prof. Mihaela van der Schaar

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Master of Philosophy in Machine Learning and Machine Intelligence

St Edmund's College

August 2021

I would like to dedicate this thesis to my loving family, friends, and partner, without whom all of this would not be possible.

Declaration

I, Tennison Liu of St Edmund's College, being a candidate for the MPhil in Machine Learning and Machine Intelligence, hereby declare that this report and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose

All software used in this thesis was written from scratch in Python. Code for Fair-COCCO can be found at https://github.com/tennisonliu/fair-cocco and code for Fair-PoLe can be found at https://github.com/tennisonliu/fair-pole.

This dissertation contains fewer than 14208 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Tennison Liu
August 2021

Acknowledgements

First and foremost, I would like to acknowledge my supervisor, Mihaela van der Schaar, for her invaluable guidance on the technical aspects of my project as well as the patience and personal mentoring to help me develop as a researcher.

I would also like to express my gratitude to members of the van der Schaar Lab, especially Alex Chan, Boris van Breugel, and Alexis Bellot in invaluable discussions, thoughtful encouragements, and overall great banter.

Finally, I would like to thank my undergraduate thesis supervisors, Omid Kavehei and Nhan Duy Truong, for setting me on this path I find myself today.

Abstract

Ensuring machine learning algorithms deployed in the real world do not result in unexpected unfairness or social implications is becoming increasingly important. However, there exists a clear gap in literature for a measure of fairness that can detect discrimination against multiple sensitive attributes while also handling continuous or discrete outcomes. In this thesis, we propose a fairness measure, *Fair-COCCO*, based on the conditional cross-covariance operator on reproducing kernel Hilbert Spaces. This novel method generalise to the majority of existing fairness notions and naturally extends to settings with continuous outcomes and multi-dimensional sensitive attributes. Additionally, we demonstrate how the proposed measure can be readily implemented in stochastic gradient optimisation for fair policy learning in supervised learning settings. Empirical evaluations of Fair-COCCO on synthetic and real-world experiments reveal favourable comparisons to state-of-the-art techniques in balancing predictive power and fairness. We also see much potential in applying machine learning to analyse fairness in observed behaviour, especially in complex and high-dimensional real-world environments. To that end, we propose the first known definition of fairness for sequences of decisions and showcase how Fair-COCCO can be applied to quantify fairness in these problems. Building off these definitions, we turn to learning fair policies in real-world conditions, where learning is constrained to be performed offline. We propose *Fair-PoLe*, a novel inverse reinforcement learning that operates completely offline and is computationally efficient and functionally expressive when compared to existing methods. We illustrate the potential for Fair-PoLe to learn policies that balance imitation of expert policies with fair outcomes on the challenging problem of sepsis treatment.

Table of contents

Nomenclature

Roman Symbols

γ	Discount Factor
$\mathcal{A}$	Action Space
$\mathcal{D}$	Dataset
$\mathcal{H}$	Hilbert Space
$\mathcal{S}$	State Space
$\mathcal{X}$	Feature Space
$\mathcal{Y}$	Target Space
$\mathcal{Z}$	Sensitive Attributes Space

Greek Symbols

ε	Regularisation Constant
π	Policy
σ	Standard Deviation
τ	Trajectory

Other Symbols

$\Delta(\cdot)$	Space of Probability Distributions over Set
$\mathbb{E}(\cdot)$	Expectation Operator

$\phi(\cdot,\cdot)$	Feature Map
$k(\cdot,\cdot)$	Kernel Function
$Q(\cdot,\cdot)$	Action-value Function
$R(\cdot,\cdot)$	Reward Function
$T(\cdot,\cdot,\cdot)$	Transition Function
$V(\cdot)$	Value Function

Acronyms / Abbreviations

BC	Behavioural Cloning
BIRL	Bayesian Inverse Reinforcement Learning
CAL	Calibration
CCO	Cross-Covariance Operator
DEO	Difference in Equal Opportunity
DI	Disparate Impact
DP	Demographic Parity
EO	Equalised Odds
Fair-PoLe	Fair Policy Learning
FTU	Fairness Through Unawareness
HS	Hilbert-Schmidt
i.i.d.	Identical and Independently Distributed
IL	Imitation Learning
IRL	Inverse Reinforcement Learning
KDE	Kernel Density Esimation
MaxEnt	Maximum Entropy
MCC	Maximal Correlation Coefficient

MCMC	Markov Chain Monte Carlo
MDP	Markov Decision Process
MI	Mutual Information
NOCCO	Normalised Cross-Covariance OPerator
RL	Reinforcement Learning
TF	Test Fairness

Chapter 1

Introduction

As machine learning (ML) technologies become increasingly prevalent in everyday life, researchers and institutions have the responsibility of ensuring predictions made by the algorithms are not unfair towards personal attributes such as gender, ethnicity and disabilities. This issue was brought into the limelight by a ProPublica investigation into the COMPAS software, a predictive algorithm employed by U.S. courts to assess risks of recidivism (Angwin et al., 2016). Follow-up analysis of the algorithm revealed that African-Americans are almost twice as likely as White-Americans to be labelled a higher risk but not actually re-offend. Recently, this call for societal accountability and social understanding of ML has reached mainstream media. Governments around the world, such as the EU (Hacker, 2018), US (Podesta and Others, 2014) and NGOs (e.g. the Association of Internet Researchers Markham and Others (2012)) are increasingly concerned with the responsible use of machine learning.

In response to these challenges, the machine learning community has developed many fairness notions and methods, which has sparked a field of research known as *algorithmic fairness*. Within this growing body of literature, many works have pioneered new definitions and metrics of fairness; illustrated how human or algorithmic behaviour can be audited; and introduced data collection and modelling methodologies into creating policies in a fairness-aware setting. Despite the increased discussion around fairness, what constitutes fairness or how to quantify it are still open questions. The lack of universal standards means that most commercial and deployed ML products are yet to include means to ascertain social fairness.

While much discussion has been rightfully paid to the fairness of ML driven predictions, if we invert the perspective, there also exists massive potential for ML techniques to be used to analyse fairness in human decision making. Consider, as an example, the healthcare setting, where there exists significant variation in clinical practice. In such complex and noisy environments, how do we answer the question: *what amount of variation is unwarranted or unfair, and what amount if justified?* This thesis seeks to address this question, specifically focusing on developing methods to achieve two key goals:

1. Audit and quantify fairness of decisions made by humans,
2. Fairness-aware learning of policies for predictive tasks.

Achieving the first goal involves uncovering the motivations behind observed behaviour, a challenging ML task when environments are complex and temporal dynamics are important. For the second goal, we introduce a unifying framework of *fair policy learning*, which seeks to learn fair policies in one-shot and sequential decision making settings.

Most studies to date in algorithmic fairness have focused on the comparatively simpler *one-shot decision making* setting, or tasks where only a single decision is required. An example of this is approving loans and ensuring the approval outcomes are not unfair with respect to personal attributes of the applicant. These settings do not consider feedback or delayed impact of the decisions made. A plethora of methods to quantify fairness in these tasks have been proposed, but the generally consider discrete outcomes and personal attributes although many sensitive attributes (e.g. age) are continuous values. Additionally, fair predictions in the presence of multiple sensitive attributes has received less attention amidst existing work. It is, however, necessary for algorithms deployed for real-world use to be capable of analysing fairness when multiple sensitive attributes are present and when the outcomes are continuous. Our first contribution directly addresses these issues by introducing a novel kernel-based fairness measure. The proposed measure, which we term *Fair-COCCO* employs a strong characterisation of fairness and naturally handles multiple attributes and continuous outcomes. Additionally, if we formulate fair policy learning in one-shot settings as supervised learning problems, Fair-COCCO can be employed as a differentiable regulariser to perform fairness-aware learning.

Compared to the one-shot setting, auditing fairness in *sequential decision making* is a challenge yet to be addressed in the literature. The problem is more challenging as analysing fairness in sequential settings necessitates incorporating temporal dynamics and long-term outcomes. Inspecting fairness also means we must go beyond observed behaviour to scrutinise the underlying goals and motivations that induced the behaviour. To that end, we start by introducing definitions of fairness on sequences of decisions. We are also interested in learning fair policies, particularly about learning in complicated and high-stakes environments, such as clinical settings, where we do not have perfect knowledge of the environment and we cannot interact with it (e.g. testing policies on patients). We formulate this as a reinforcement learning problem, where the goal is to learn policies solely from observational datasets of expert demonstrations. We introduce *Fair-PoLe*, an *inverse reinforcement learning* algorithm, which, when incorporated with Fair-COCCO regularisation, learns to mimic expert behaviour while improving fairness in a purely offline fashion.

1.1 Contributions

The work completed in this thesis indicates a number of novel contributions, including:

- Introducing **Fair-COCCO, a kernel-based fairness measure**, with strong characterisation of fairness that naturally extend to multiple attributes and types of outcomes, addressing a clear gap in algorithmic fairness.
- Proposing a new **framework of analysing fairness in sequential decision making** scenarios using Markov Decision Processes and introducing notions of fairness in these settings.
- Presenting **Fair-PoLe, a novel inverse reinforcement learning algorithm** that is computationally efficient and can learn fair policies in complex environments solely from logged expert demonstrations.
- Demonstrating the **superiority of Fair-COCCO and potential for Fair-PoLe** to improve fairness outcomes in real world problems.

1.2 Thesis Outline

The rest of this thesis is structured as follows. In Chapter 2, we cover preliminaries, reviewing different notions of fairness and describing general concepts around sequential decision making and related works in inverse reinforcement learning.

In Chapter 3, we introduce Fair-COCCO, a kernel measure of fairness. We start by reviewing key shortcomings in current methods, including the inability to scale with multiple sensitive attributes and accounting for discrete and continuous outcomes. After briefly reviewing the theory of covariance operators on reproducing kernel Hilbert Spaces, we highlight the key features that make our novel measure superior. Then, we propose methods to learn fair policies in one-step decision making, which are formulated as supervised learning and leverage Fair-COCCO as a differentiable regulariser.

Moving on, in Chapter 4, we discuss the unifying framework of fair policy learning. Based off of existing notions of group fairness, we introduce new definitions of fairness in sequential decision making scenario, using Markov Decision Processes to model temporal dynamics. We then introduce Fair-PoLe, an offline inverse reinforcement learning algorithm that delivers policies balancing fairness and performance.

In Chapter 5 we provide empirical evidence to demonstrate that Fair-COCCO can effectively quantify fairness and that Fair-COCCO regularised learning compares favourably to state-of-the-art techniques on a variety of benchmarks. Additionally, we demonstrate Fair-PoLe as a method to analyse fairness in observed sequential decisions and derive fair policies on a challenging medical problem of sepsis treatment. We conclude in Chapter 6, provide closing thoughts, highlight limitations of our proposed study and indicate directions for future improvements.

Chapter 2

Fairness and Decision Making

We start this chapter by reviewing key notions of fairness and how they can be stated in terms of (conditional) dependencies. Subsequently, we set up general concepts around sequential decision making and reinforcement learning: the task of learning a policy based on reward signals provided to an agent. We then review existing methods in inverse reinforcement learning, one of the main focuses of our work. In these settings, rewards are not explicitly provided but rather learned from demonstrations provided by an expert. While traditional methods are primarily concerned with fully online settings, we review the offline version of inverse reinforcement learning, which is necessary when learning in the minimal possible settings.

2.1 Notions of Fairness

Merriam-Webster Dictionary defines fairness as:

> *"Fair or impartial treatment, especially lack of favouritism towards one side or another."*

Despite the straightforward definition, there is no universal measure of fairness, and the correct notion depends on ethical, legal and technical contexts. Indeed, Corbett-Davies et al. (2017) duly highlighted that many fairness notions are difficult or impossible to satisfy at the same time. Most approaches to quantify unfairness involve, to some extent, the concept of *sensitive*, *personal* or *protected attributes*, terms which are employed interchangeably. A *sub-group* is a group of individuals who share the same sensitive attributes. Sensitive attributes are defined on socio-cultural variables of an individual that might result in discrimination and unfair treatment. Some common examples include gender, ethnicity and age. While what is considered a sensitive attribute can differ drastically across different settings, this is not an aspect we consider in this thesis, instead resorting to those explicitly identified by legal frameworks [1][2].

Studies to date have adopted one of two broad families of fairness definitions: *group fairness* and *individual fairness*. Group fairness is typically defined on the outcomes received by protected sub-groups and algorithms attempting to enforce group fairness do so by equalising outcomes for all protected sub-groups. This is by far the most popular definition of fairness and includes a very wide body of literature (see Barocas et al. (2017) for comprehensive review). While group fairness guarantees are averaged over all individuals in a protected sub-group, individual fairness (Zemel et al. (2013), Dwork et al. (2012)), in contrast, ignores inter-group restrictions and requires similar individuals to be treated similarly. However, it is unclear how similarity metrics should be defined over individuals and outcomes. Additionally, analysis of individual fairness often make untenable functional assumptions (Fleisher, 2021) In this study, we focus on *group fairness* notions and introduce methods to quantify fairness that generalise to any specific notion of group fairness.

Next, we briefly overview four popular definitions of group fairness and how each corresponds to a different aspect of fairness. We make an effort to make clear the connection between fairness notions and measures of (conditional) dependence (see Table 2.1), an insight important in the formulation of Fair-COCCO in Chapter 3. Let random variables $X \in \mathcal{X} \subset \mathbb{R}^{d_X}$ and $Y \in \mathcal{Y} \subset \mathbb{R}^{d_Y}$ denote respectively the d_X-dimensional features and d_X-dimensional target. Let $Z \in \mathcal{Z} \subset \mathbb{R}^{d_Z}$ denote d_Z-dimensional sensitive attributes that we want to protect (e.g. gender or race). We do not make specific assumptions around whether Z is included in the feature vector. Consider the model $f : \mathcal{X} \to \mathcal{Y}$ with prediction $\hat{Y} = f(X)$.

Fairness through unawareness (FTU) (e.g. Grgic-Hlaca et al. (2016)) prohibits the algorithm from using sensitive attributes explicitly in making predictions. While straightforward to implement, this method ignores the indirect discriminatory effect of proxy covariates that are correlated with Z (e.g. “redlining” (Avery et al., 2009)). *Demographic parity*

[1] https://www.equalityhumanrights.com/en/equality-act/protected-characteristics
[2] https://www.gov.uk/discrimination-your-rights

Table 2.1 Popular definitions of fairness in terms of (conditional) independence requirements

Definition	Requirement
FTU	$Z \perp\!\!\!\perp \hat{Y} \mid X \backslash Z$
DP	$Z \perp\!\!\!\perp \hat{Y}$
EO	$Z \perp\!\!\!\perp \hat{Y} \mid Y$
CAL	$Z \perp\!\!\!\perp Y \mid \hat{Y}$

(DP) (Barocas and Selbst, 2016; Zafar et al., 2017) does not allow for indirect discrimination, by requiring statistical independence between predictions and attributes $\hat{Y} \perp\!\!\!\perp Z$, i.e. $p(\hat{Y}|Z=z) = p(\hat{Y}) \;\; \forall \;\; z \in \mathcal{Z}$. Evidently, this strict notion sacrifices predictive utility by ignoring all correlations between Y and Z, thereby precluding the optimal predictor. Dwork et al. (2012), most notably, argued that this approach permits laziness, which can hurt fairness in the long run.

To address some of these concerns, Hardt et al. (2016) introduced *equalised odds* (EO), requiring that predictions $\hat{Y}$ and attributes Z are independent given the true outcome Y, i.e. $\hat{Y} \perp\!\!\!\perp Z|Y$ or $p(\hat{Y}|Y=y, Z=z) = p(\hat{Y}|Y=y) \;\; \forall \;\; z \in \mathcal{Z}, \; y \in \mathcal{Y}$. This approach recognises that sensitive attributes have predictive value, but only allows Z to influence $\hat{Y}$ to the extent allowed for by the true outcome Y. Additional notions of fairness include *calibration* (CAL) (Kleinberg et al., 2016), which ensures that predictions are calibrated between subgroups ($Y \perp\!\!\!\perp Z|\hat{Y}$). For a comprehensive review of fairness notions, we refer the interested reader to §3 in Caton and Haas (2020).

2.1.1 Group Fairness Metrics

In performing retrospective audits of fairness in logged decisions, several metrics have been defined to quantify the level of unfairness. These definitions are generally aligned with the respective fairness notion that is being investigated or enforced. Problematically, the majority of these metrics are defined for when outcomes and sensitive attributes are binary or discrete, while a clear gap exists in literature for metrics that can handle continuous predictions and attributes.

Difference in equal opportunity (DEO) (Hardt et al., 2016) is defined for EO and highlights the different predictions made based on different group memberships:

$$DEO = |P(\hat{Y}|Z=1, Y=1) - P(\hat{Y}|Z=0, Y=1)|$$

which is 0 iff the false negative rates are the same for both groups, i.e. the probability of being mistreated is independent of Z.

Disparate impact (DI) (Zemel et al., 2013) is a parity metric based on DP and looks at the probability of being classified with the positive label:

$$\frac{P(\hat{Y}|Z=1)}{P(\hat{Y}|Z=0)}$$

DI considers the difference in impact experienced by sub-groups and originated from legal doctrine. The US Equal Employment Opportunity Commission Recommendation advocates that DI should not be lower than 0.8 (when $Z=0$ denotes the privileged group), a rule known as the '80%-Rule'.

Test Fairness (TF) (Chouldechova, 2017) is defined on CAL and attempts to evaluate if the probability of $Y=1$ is the same across sub-groups if the same predictions are made:

$$P(Y=1|\hat{Y}=\hat{y}, Z=1) = P(Y=1|\hat{Y}=\hat{y}, Z=0)$$

This concept cannot be computed directly as it is conditioned on a particular value of $\hat{Y}=\hat{y}$. Thus, CAL is usually visualised through the use of reliability diagrams.

2.2 One-Shot and Sequential Decision Making

In this section, we formally describe sequential decision making problems, but defer a full review of the relevant theory to Sutton and Barto (2018). Generally, sequential decision making describes any step-by-step decision making, where actions can have delayed impact and dynamics of the environment affect the outcome of the problem. In ML literature, such types of problems are usually set up through the *reinforcement learning* (RL) framework using *Markov Decision Processes* (MDPs).

Generally, RL refers to problems where agents taking actions in an environment in order to maximise the notion of cumulative rewards. We consider the standard MDP M, described by the tuple $M = \langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$, where:

- $s \in \mathcal{S}$ are states,
- $a \in \mathcal{A}$ are actions,
- $T \in \Delta(\mathcal{S})^{\mathcal{S}\times\mathcal{A}}$ describes the transition dynamics, where $T(s'|s,a)$ is the probability of transitioning to state s' after taking action a while currently in s,
- $R \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ are the rewards, where $R(s,a)$ are the rewards for being in state s and taking action a,
- $\gamma \in [0,1)$ is the discount factor.

Let $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ represent a stochastic policy, where $\pi(a|s)$ indicates the probability of taking action a given state s. In the standard RL problem, the aim is to find the optimal policy π^* that will maximise the expected discounted sum of rewards:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right] \tag{2.1}$$

Here, the expectation is with respect to trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots)$. More specifically, the expectation depends on: the starting state distribution $s_0 \sim \mu$, where $\mu : \mathcal{S} \to [0,1]$, policy $a_t \sim \pi(\cdot|s_t)$ and dynamics $s_{t+1} \sim T(\cdot|s_t, a_t)$. For completeness, we further define several auxiliary concepts, which simplify mathematical manipulation of policies. Let the value of a state s when following policy π be given as:

$$V_{\pi}(s) = \mathbb{E}_{T,\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s\right] \tag{2.2}$$

which can be interpreted as the expected discounted sum of rewards from starting in state s and following policy π. A closely related idea is the action-value function:

$$Q_\pi(s,a) = \mathbb{E}_{T,\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t,a_t)|s_0 = s, a_0 = a\right] \tag{2.3}$$

$$= R(s,a) + \gamma \mathbb{E}_{T,\pi}\left[Q_\pi(s',a')\right] \tag{2.4}$$

Also known as the Q-function, it defines the expected discounted sum of rewards starting in state s, taking action a and following policy π. The expression can be re-written to make explicit the connection to the reward R and the discounted, expected action-value of the next state-action pair (s',a'). The Q-function is important in that, under the Bellman Optimality Theorem, it fully describes the optimal policy. Namely, $\forall s \in \mathcal{S}$, $\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q_{\pi^*}(s,a)$

This is one of the fundamental pillars of RL and many successful algorithms have been developed around the Q-function. This includes the original Q-Learning (Watkins and Dayan, 1992), a dynamic programming approach to evaluating the quality of particular actions in states. Recently, these methods have been powered-up by leveraging the expressive function approximations of deep learning approaches to approximate the Q-function (Henderson et al., 2018; Lillicrap et al., 2015; Mnih et al., 2013). Regardless, these algorithms still leverage the recursive Bellman Optimality equations and *temporal difference error*, seeking to minimise the difference between estimated rewards and actual received rewards. However, these methods require extensive interaction with the environment to evaluate the learned Q-function in order to reach satisfactory performance (Li, 2017). Additionally, RL techniques are highly reliant on reward mechanisms that elicit the desirable learning behaviours in the agent (Dewey, 2014).

Importantly for our goal of analysing fairness in sequential decision making, if the rewards at each step are explicit, the problem becomes significantly simpler. However, this requirement rarely satisfied in the real world. Consider the healthcare setting—rewards for individual clinician action are rarely available, meaning that we often rely on spare indicators that are only available once the sequence of interactions is complete, such as indicators on recovery after treatment. Additionally, core RL algorithms, are in their nature, *online* algorithms that require repeated interactions with the environment. While this requirement is satisfied in game-playing or simulated environments, it is infeasible in real-world settings where online interaction will have meaningful consequences. For example, it is unjustifiable to deploy policies on patients mid-training to learn policies.

Thus, traditional RL techniques are not the main focus of this thesis, but the theory reviewed provide relevant context when we focus on the inverse problem, that is, recovering the underlying rewards from observed expert demonstrations.

2.3 Inverse Reinforcement Learning

Consider auditing real-world sequential decision making on the basis of demonstrated behaviour—i.e. from *logged trajectories* of observations and actions taken by an expert. To perform any meaningful analysis on the fairness of the actions, simply examining the actions are not enough. Instead, we need knowledge of the expert's preferences and motivations. Decision theory literature suggest maximisation of *expected utility* as a model of decision making in uncertainty. Thus, our task becomes that of recovering the utility function from observed human behaviour. In the ML community, this field of research is known as *inverse reinforcement learning* (IRL). Formally:

Definition 2.3.1 (Inverse Reinforcement Learning (Piot et al., 2016)) *IRL is a paradigm relying on Markov Decision Processes (MDPs), where the goal is to find a reward function from the expert demonstrations that could explain the expert behaviour.*

Imitation Learning (IL) is a closely related concept. IL is generally concerned with mimicking and generalising observed expert behaviour (see Hussein et al. (2017) for review). While IL is largely concerned with learning policies to match the expert, IRL directly models the underlying reward function that induces the observed behaviour. Once the underlying rewards are recovered, traditional RL techniques can be applied to learn a policy optimal with respect to learned rewards.

2.3.1 Behavioural Cloning

Before discussing more advanced methods, we start by introducing *behavioural cloning* (BC) (Bain and Sammut, 1995), the simplest of IL methods. In BC, the agent receives demonstrations including both the states encountered and actions taken by the expert. By assuming that each state-action pair is *i.i.d.*, the algorithm reduces the problem to that of supervised learning, learning a regressor or classifier to match the expert policy (Ross and Bagnell, 2010). BC methods are powerful as they do not require knowledge of the environment or the ability to interact with it. To that end, it has successfully been applied to a variety of IL tasks, including training a quadrotor to fly (Giusti et al., 2015), end-to-end learning for self-driving cars (Bojarski et al., 2016), and teaching robots complex movements (Niekum et al., 2015).

Of course, the *i.i.d.* assumption does not hold in sequential decision making as the future state distribution depends on predictions made in the current step. Ross and Bagnell (2010) proved that this leads to compounding errors and errors that grow quadratically with the time horizon (alternatively, the further away from the training state distribution). The issues of ignoring the effects on the underlying state distribution were elucidated in the success of ALVINN (Pomerleau, 1989), a self-driving car, where the agent does not know how to recover when the car occasionally strays away from the centre of the road.

2.3.2 Learning the Reward Function

State-of-the-art IL techniques address the generalisation issue of BC by accounting for the temporal dynamics through MDPs. Inverse reinforcement learning (IRL) was first introduced by Ng et al. (2000) and can be formally defined as recovering the unknown reward function $R(\cdot,\cdot)$ given an MDP/R (an MDP without a reward function) and a dataset of N logged trajectories $\mathcal{D} = \{\tau_i\}_{i=1}^{N}$. Once the underlying reward function is learned, an imitator policy can be learned through any variety of RL techniques.

Standard IRL solutions work by iterating through a few steps: 1) proposing reward function, 2) solving the MDP with respect to the candidate reward function to generate a candidate policy and 3) optimising the reward function against a loss defined on the divergence between candidate policy and expert policy. These steps are repeated until a satisfactory reward function is produced. We call the second step the *forward RL* step to distinguish it from the inverse procedure. This standard template introduces a set of challenges, the most important of which is that the inverse problem is fundamentally ill-posed. Notably, that for any given set of demonstrations $\mathcal{D}$, there are (infinitely) many reward functions for which the demonstrations are optimal (Ng et al., 2000). Consider as a simple example if the reward function is constant for all inputs.

Several broad approaches have been introduced to address this. *Max-margin* (Abbeel and Ng, 2004) seeks to learn a reward function for which the margin between the expected value of the demonstrated behaviour and that of any candidate policy is maximised. This class of methods assume the reward function is linear in the feature map constructed from state features $\phi(s,a)$, i.e. $R(s,a) = w \cdot \phi(s,a)$. The learning goal is to find reward weights w that induces a policy that minimises the margin between*feature expectations* of the expert and candidate policies (Ng et al. (2000), Abbeel and Ng (2004)). Nonetheless, as a heuristic, max-margin introduces a bias into the learned reward function.

Additionally, max-margin assumes the expert is always acting optimally, which is rarely satisfied in real-world situations involving human operators. In contrast, probabilistic methods have been developed, including around *maximum entropy* (MaxEnt) principle (Ziebart et al., 2008) and *Bayesian* IRL (BIRL), which allows for *near-optimal* expert behaviour.

MaxEnt IRL is formulated around globally normalised distribution over trajectories, which resolves ambiguity by selecting reward functions that best matches optimal behaviour while maintaining maximum entropy in the distribution over trajectories (Ziebart et al., 2008). This probability of a trajectory can be described as:

$$P(\tau|w) = \frac{1}{Z(w)} \exp\left(w^T \sum_{s_j \in \tau} f_{s_j}\right) \tag{2.5}$$

where f_{s_j} is the feature count of state s_j. Thus, trajectories with higher rewards are exponentially more probable and preferred.

Both max-margin and MaxEnt methods employ heuristics to distinguish between plausible reward hypothesis but at the cost of potentially dismissing the true reward function. *Bayesian* IRL (BIRL) (Ramachandran and Amir, 2007) takes a probabilistic view of the reward, where inference is performed on the posterior distribution of rewards having seen the demonstrations. The likelihood of an action given a state is defined using the Boltzmann distribution with inverse temperature β and respective state-action value $Q_R(s,a)$ given the rewards:

$$P(a|s,R,\theta) = \frac{1}{Z(\theta)} \exp\left(\beta Q_R(s,a;\theta)\right) \tag{2.6}$$

where θ are the parameters of the reward function. The posterior distribution over rewards is intractable and Ramachandran and Amir (2007) used a Markov Chain Monte

Carlo (MCMC) approach to sample from the posterior. As Brown and Niekum (2019) noted, BIRL suffers from linear reward formulation and the expensive sampling procedure limits the feasibility to small, solvable environments.

Another challenge common to all these methods is the repeated inner-loop calls to forward RL procedures. The forward RL procedures solves the MDP with the candidate reward function to generate learned policy, which is a pre-requisite to evaluating the candidate reward function against observed expert behaviour. While not an issue in small and known MDPs, this inner-loop procedure quickly becomes infeasible for real-world tasks that routinely involve continuous or large-dimensional environments. Consider as an example DQN learning introduced in Mnih et al. (2013), which easily takes several hours to complete, a procedure which will have to be repeated multiple times within the training loop of IRL methods.

2.3.3 Offline Inverse Reinforcement Learning

In most real-world problems, we are constrained to operate in the minimal possible setting, where we do not have perfect knowledge of the MDP and cannot perform further experimentation. Consider again the medical setting, where we do not have perfect knowledge of the environment and it is unethical and impractical to test policies during training. This effectively means that we must must resort to *offline* IRL, where we only have access to trajectories generated by an expert policy in the form of logged dataset $\mathcal{D}$. The IRL methods reviewed in the previous section are, by their nature, online algorithms that require knowledge of dynamics or interactions with the environment (e.g. Choi and Kim (2011a), Abbeel and Ng (2004)). Formally, we define offline IRL:

Definition 2.3.2 (Offline IRL) *Offline IRL has the same goals IRL but performed in a MDP/RT environment, where there is no knowledge of either the underlying reward R or the transitions T. The lacking knowledge of dynamics is strong in the sense that it cannot be approximated through simulation.*

Porting pre-existing approaches to strictly offline settings is highly non-trivial. Max-margin approaches are perhaps the easiest to adapt, as the linear parameterisation allows the use of off-policy evaluation to estimate feature expectations. This is the approach taken in

Klein et al. (2012), which utilises the feature expectations to parameterise the score function of multi-class classifier. Similarly, Klein et al. (2011) propose using LSTD-Q to estimate feature expectations, Bica et al. (2020) employ counter-factual reasoning and Lee et al. (2019) propose a deep successor feature network based on Q-learning. However, these approaches similarly assume reward functions that have a non-expressive linear parameterisation in state feature maps.

Relatively speaking, transferring MaxEnt IRL for offline use is even more challenging. The formulation of MaxEnt is based on a probability distribution over trajectories, where the partition function (i.e. $Z(w)$ in Equation 2.5) is defined over all possible trajectories and is not straightforward to compute. This makes problems involving high-dimensional environments or unknown dynamics intractable due to exponentially increasing number of trajectories to consider. To address this, several model-free approximations for MaxEnt IRL have been proposed, including Finn et al. (2016) and Ho and Ermon (2016) but both require the ability to interact with a simulator to sample trajectories.

While there have been several extensions to the original BIRL formulation (e.g. Choi and Kim (2011b) considers maximum-a-posteriori inference), very little work has been done in the area compared to other approaches due to the difficulty in scaling the method to higher-dimensional state-spaces and convergence difficulties inherent to MCMC (Gamerman and Lopes, 2006). Thus, few works have extended BIRL methods, let alone to offline settings. More recently, Chan and van der Schaar (2021) proposed a variational inference approach to posterior reward learning, which has improved scalability and is applicable to the offline setting.

Chapter 3

Fair-COCCO: Kernel Measure of Fairness

The ability to quantify fairness variation across decision processes is the first step to learning fair policies. In this Chapter, we introduce Fair-COCCO, a kernel measure to evaluate fairness. Fair-COCCO employs a strong characterisation of fairness and naturally handles sensitive attributes of multiple dimensions as well as outcomes that are continuous or discrete, overcoming limitations of existing work. We further demonstrate how the differentiability of Fair-COCCO enables it to be employed as a regulariser for fair policy learning and conclude the chapter by illustrating this on one-shot decision making settings.

3.1 Makings of a Good Fairness Measure

Most studies to date on algorithmic fainess quantification have focused on classification settings, where the outcome and sensitive attribute are discrete. In these discrete settings, fairness criteria, such as DP and EO can be computed directly by comparing rates of outcomes between sub-groups. However, many sensitive attributes, e.g. age or ethnic proportions, are *continuous values*. Existing methods discretise continuous values into categorical bins, which lacks intuitive and applied appeal as it introduces thresholding effects into the modelling process and discards element order information.

Additionally, fairness in the presence of *multiple sensitive attributes* has received less attention amidst existing work. U.S. federal law protects groups from discrimination based on nine protected classes [1], highlighting the need to detect discrimination against individuals with multiple sensitive attributes. Existing methods approach this by introducing additional fairness conditions on each attributes, which can present computational challenges or additional hyperparameters for regularisation approaches during learning.

Lastly, and perhaps most importantly, we wish for a *strong characterisation of fairness*. §2.1 showcased how we can quantify fairness by evaluating dependence between random variables. We use the word 'strong' in the sense defined by Daudin (1980) in terms of how well a measure characterises dependence. A strong characterisation allows us to fully capture fairness information, instead of resorting to relaxations at the risk of forfeiting true fairness quantification.

Fair-COCCO applies kernel measures for evaluating the dependence of outcome and protected attribute, such that the fairness requirements hold (Table 2.1). Fair-COCCO, as an algorithmic fairness measure is based on the Conditional Cross-Covariance Operator on reproducing kernel Hilbert Spaces (RKHS). It is a strong measure and, to the best of our knowledge, it is the only measure that can naturally extend to sensitive attributes of multiple dimensions as well s continuous or discrete outcomes. Additionally, it can serve as a statistical test, leading to stronger guarantees and more transparency in practice.

3.1.1 Existing Approaches in Algorithmic Fairness

Binary Classification Much of the literature focus on the setting with single, binary label and attribute (Donini et al., 2018; Goel et al., 2018; Jiang et al., 2020; Kamishima et al., 2012). While these methods can be applied to continuous variables through discretisation, this introduces unwanted threshold effects. By contrast, limited attention has been paid to fairness in settings with continuous outcomes or settings with multiple protected attributes. Fair-COCCO naturally addresses these settings by applying kernel transformations of the distributions.

Conditional Dependence with Continuous Variables Table 2.1 explicitly connect fairness notions with (conditional) dependence measures. At the core of many algorithmic

[1] https://www.eeoc.gov/discrimination-type.

fairness techniques is how these measures are estimated and constrained. In simplistic cases where variables are categorical, a conditional independence measure can be obtained by testing for unconditional independence $X \perp\!\!\!\perp Y|Z = z$ for every possible realisation $z \in Z$. However, quantifying conditional dependence between continuous variables is much more difficult than the discrete case (Bergsma, 2004). For intuition, consider the tuple $\{(X_i, Y_i, Z_i)\}_{i=1}^N$, where X is continuous: there is, with probability one, at most one observed (X_i, Y_i) for any value of Z. Bergsma (2004) proved that at least two pairs would be needed in order to have any "information" on conditional dependence for corresponding values of D.

Relaxations The difficulty in measuring dependence in continuous variables prompted many *weak* dependence measures of fairness (Daudin, 1980), which make simplifying assumptions. Calders et al. (2013), Johnson et al. (2016b) and Bechavod and Ligett (2017) reduce the problem to measuring the distance between moments of distributions. Donini et al. (2018) generalises this to distance between first moments of functions of $\hat{Y}, X$, and Z. Zafar et al. (2017) measure second-order dependencies through conditional covariance between $\hat{Y}$ and Z, corresponding to linear correlations only and Kamishima et al. (2012) introduced a first moment relaxation of mutual information (MI). While these methods are computationally tractable, the relaxations reduce the measures to weak measures of fairness, whereas our proposed kernel measure can strongly capture statistical dependence.

Strong Measures Another approach is to adopt *strong* measures of dependence (Daudin, 1980), including Mutual Information (MI), maximal correlation coefficients (MCC) and kernel methods. Zhang et al. (2012) and Cho et al. (2020) focus on MI but rely on kernel density estimation (KDE) for explicit estimation of the densities. Lowy et al. (2021) and Mary et al. (2019) similarly develop MCC as a measure, but similarly rely on KDE. However, density estimation is not straight forward for continuous variables, does not scale to higher dimensions and can introduce bias through choice of kernels. Steinberg et al. (2020a) and Steinberg et al. (2020b) adopts a MI-theoretic measure using density ratio estimation, which requires the training of a separate probabilistic classifier in each training loop. In contrast, our method is computationally efficient and naturally scales to high-dimensional sensitive attributes. Perhaps most similar to our work is Pérez-Suay et al. (2017), who seek zero correlation in RKHS to capture statistical independence but it cannot extend to conditional dependence notions, limiting its use to DP fairness.

Multiple attributes Few existing methods support multiple attributes, even though this is common in practice. Some works, including Cho et al. (2020); Donini et al. (2018); Zafar et al. (2017) can be scaled to multiple attributes by calculating fairness against individual attributes and aggregating, which quickly becomes cumbersome. Additionally, when performing fairness-aware learning, this approach results in more tune-able penalty parameters. By comparison, the fairness-fitness trade-off of Fair-COCCO can be tuned through a single hyperparameter.

3.2 Kernel Measure of Dependence

Most fairness metrics rely on estimating and enforcing (conditional) independence between the algorithm predictions and the sensitive attributes to protect. As this dependence can be arbitrarily complex, determining dependence is challenging. In this section, we briefly overview a kernel measure of dependence, which Fair-COCCO is based on. While we frame this discussion around conditional dependence, it is straightforward to generalise to the unconditional case.

Setup To avoid confusion, we emphasise the distinction between the notation employed in this section and elsewhere. Let X,Y,Z be random variables with domains $\mathcal{X}$,$\mathcal{Y}$, $\mathcal{Z}$, respectively. Let $\mathcal{H}_{\mathcal{X}}$ be the RKHS on $\mathcal{X}$, with positive definite kernel $k_{\mathcal{X}}$. $k_{\mathcal{Y}}, \mathcal{H}_{\mathcal{Y}}$ and $k_{\mathcal{Z}}, \mathcal{H}_{\mathcal{Z}}$ are defined similarly. The spaces of squared integrable functions of X are denoted by L^2_X i.e. $L^2_X = \{g(X)|\mathbb{E}[g^2] < \infty\}$. We assume that all involved RKHSs are separable and square integrable. Additionally, to ensure RKHSs are subsets of the L^2 space, we assume kernels are characteristic and integrality i.e. $\mathbb{E}[k(X,X)] < \infty$, (which is satisfied by Gaussian, Laplacian kernels). Formally, the problem of interest is quantifying the conditional dependence between X and Y given Z on finite samples.

Fortunately, a strong characterisation of dependence can be efficiently measured without explicitly estimating the densities or making relaxations by using kernel-based methods. We propose using a kernel measure, based on the conditional cross-covariance operator in Reproducing Kernel Hilbert Space (RKHS) (Fukumizu et al., 2007). A RKHS is a Hilbert Space in infinite dimensions, where each point in the space is a linear continuous function. Distributions of variables can be embedded in the RKHS, allowing for comparison of higher-order moments and inference on dependence properties between distributions. There is a large body of literature on kernel-based methods developed to capture dependence of variables (see Bach and Jordan (2002); Gretton et al. (2005) for more insights).

The cross-covariance operator (CCO) is a unique, bounded operator $\Sigma_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ defined by the relation:

$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}_{XY}[f(X)g(Y)] - \mathbb{E}_X[f(X)]\mathbb{E}_Y[g(Y)] \tag{3.1}$$

for all $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$. Intuitively, the Σ_{YX} operator extends the covariance matrix defined on Euclidean spaces and represents higher-order covariance of X and Y through functional transformations$f(X)$ and $g(Y)$ with non-linear kernels. A closely related operator is the normalised cross-covariance operator (NOCCO) (Baker, 1973):

Definition 3.2.1 (Normalised Cross-Covariance Operator (NOCCO)) *V_{YX} is defined by:*

$$V_{YX} = \Sigma_{YY}^{-\frac{1}{2}} \Sigma_{YX} \Sigma_{XX}^{-\frac{1}{2}} \tag{3.2}$$

which can be used to measure the dependence of X and Y. Σ_{YY}, Σ_{YY} are defined similarly to (3.1).

We note the inverses in Equation 3.2 may not exist as bounded operators (although the theoretical definition is justified). In practise, to ensure invertibility, regularisation is introduced. This operator can be used to evaluate the unconditional dependence between Y and X. Note that V_{YX} contains the same dependence information as CCO but is preferential to work with as it disentangles the influence of the marginals (analogous to the relationship between covariance and correlation). For many definitions of fairness (e.g. EO), we need a measure of the *conditional* dependence of two variables given a third variable. Based on NOCCO, the normalized conditional cross-covariance operator of (X,Y) given Z is given in the following definition (Fukumizu et al., 2007):

Definition 3.2.2 (Conditional Normalized Cross-Covariance Operator) *$V_{YX|Z}$ is defined by:*

$$V_{YX|Z} = V_{YX} - V_{YZ}V_{ZX}, \tag{3.3}$$

which can be used to measure the conditional dependence of X and Y given Z. V_{YZ}, V_{ZX} are defined similarly to (3.2).

To extend the intuition established previously, this operator can be interpreted as the partial correlation between $\{f(X)\, \forall f \in \mathcal{H}_{\mathcal{X}}\}$ and $\{g(Y)\, \forall g \in \mathcal{H}_{\mathcal{Y}}\}$ given $\{h(Z)\, \forall h \in \mathcal{H}_{\mathcal{Z}}\}$.

We round up this discussion by characterising the relation between the $V_{YX|Z}$ operator and conditional independence (Fukumizu et al., 2007).

Lemma 3.2.1 *Denote $\ddot{X} \triangleq (X,Z), k_{\ddot{\mathcal{X}}} \triangleq k_{\mathcal{X}} k_{\mathcal{Z}}$ and $\mathcal{H}_{\ddot{\mathcal{X}}}$ the RKHS corresponding to $k_{\ddot{\mathcal{X}}}$. Assume $\mathcal{H}_{\mathcal{X}} \subset L^2_X, \mathcal{H}_{\mathcal{Y}} \subset L^2_Y, \mathcal{H}_{\mathcal{Z}} \subset L^2_Z$. Further assume that $k_{\ddot{\mathcal{X}}} k_{\mathcal{Y}}$ is a characteristic kernel on $(\mathcal{X} \times \mathcal{Y}) \times \mathcal{Z}$, and that $\mathcal{H}_{\mathcal{Z}} + \mathbb{R}$ (direct sum of two RKHSs) is dense in $L^2(P_Z)$. Then:*

$$V_{Y\ddot{X}|Z} = O \iff X \perp\!\!\!\perp Y|Z \tag{3.4}$$

Alternatively, Lemma 3.2.1 can be inspected using Daudin's characterisation of conditional independence (Daudin, 1980), which explicitly enforces *uncorrelatedness of functions in L^2 space*. $V_{Y\ddot{X}|Z}$ efficiently exploits the spaces corresponding to certain characteristic kernels, which are usually smaller, making it more practically appealing (Zhang et al., 2012). Note that $\ddot{X}$ is simply the extended variable set—for ease of notation, we write $V_{XY|Z}$ in place of $V_{\ddot{X}Y|Z}$ from this point on-wards.

3.3 Fair-COCCO for Fairness Evaluation

We define the Fair-COCCO score as the Hilbert-Schmidt norm of the normalised conditional cross-covariance operator:

$$I = ||V_{YX|Z}||^2_{HS} \tag{3.5}$$

By Lemma 3.2.1, $||V_{YX|Z}||^2_{HS} = 0 \iff X \perp\!\!\!\perp Y|Z$. This is also applicable to the unconditional case—if we consider, as a special case, $\mathcal{Z} = \emptyset$ and that the same assumptions hold, $||V_{YX}||^2_{HS} = 0 \iff X \perp\!\!\!\perp Y$. Thus, Fair-COCCO scores are *non-negative* and scores closer to zero indicate higher levels of conditional independence. Consequently, this score can be used to quantify and enforce fairness notions around conditional independence, by replacing Y, X and Z with $\hat{Y}, Z, Y$, in different manners, such that the fairness condition holds. Notably, this score can also naturally accommodate continuous/discrete and multiple attributes. To make the connection with fairness notions explicit (Table 2.1), the Fair-COCCO score for the different notions can be expressed as:

$$I_{EO} = ||V_{Z\hat{Y}|Y}||^2_{HS}, \ I_{CAL} = ||V_{ZY|\hat{Y}}||^2_{HS}, \ I_{DP} = ||V_{Z\hat{Y}}||^2_{HS}. \tag{3.6}$$

3.3.1 Closed Form Expression of Fair-COCCO

The previous sections introduced covariance operators on RKHSs and made explicit the connection to fairness. In this section, we flesh out the closed-form expression of the empirical estimator of Fair-COCCO.

Fair-COCCO is based on the Hilbert-Schmidt (HS) norm of the covariance operators. An operator $A : \mathcal{H}_1 \to \mathcal{H}_2$ is called HS if, for complete orthonormal systems $\{\phi_i\}$ of $\mathcal{H}_1$ and $\{\psi_j\}$ of $\mathcal{H}_2$, the sum $\sum_{i,j} \langle \psi_j, A\phi_i \rangle_{HS}^2$ is finite (Reed and Simon, 1980). Thus, for a HS operator A, the HS norm, $||A||_{HS}$ is defined as $||A||_{HS}^2 = \sum_{i,j} \langle \psi_j, A\phi_i \rangle_{HS}^2$. Provided that $V_{\ddot{Y}\ddot{X}|Z}$ and V_{YX} are HS operators, Fair-COCCO scores can be expressed as:

$$||V_{\ddot{Y}\ddot{X}|Z}||_{HS}^2 \text{ (conditional fairness measure)} \tag{3.7}$$

$$||V_{YX}||_{HS}^2 \text{ (unconditional fairness measure)} \tag{3.8}$$

We denote using $||\hat{V}_{YX}^{(N)}||_{HS}^2$ and $||\hat{V}_{\ddot{Y}\ddot{X}|Z}^{(N)}||_{HS}^2$ the empirical estimators of the scores. More details on the derivatinos can be found at Gretton et al. (2005); Fukumizu et al. (2007). Let G_X be the centered Gram matrices, such that:

$$G_{X,ij} = \left\langle k_{\mathcal{X}}(\cdot, X_i) - \hat{m}_X^{(N)}, k_{\mathcal{X}}(\cdot, X_j) - \hat{m}_X^{(N)} \right\rangle_{\mathcal{H}_{\mathcal{X}}} \tag{3.9}$$

Additionally, $\hat{m}_X^{(N)} = 1/N \sum_{i=1}^N k_{\mathcal{X}}(\cdot, X_i)$ is the empirical mean. G_Y, G_Z are defined similarly. Based on this, R_X can be defined as follows:

$$R_X = G_X(G_X + \varepsilon N I_N)^{-1} \tag{3.10}$$

where ε is a regularisation constant, used in the same way as Bach and Jordan (2002), I_N is an identity matrix and R_Y, R_Z are defined similarly. The empirical estimator of $||\hat{V}_{\ddot{Y}\ddot{X}|Z}||_{HS}^2$ can then be computed:

$$||\hat{V}_{\ddot{Y}\ddot{X}|Z}||_{HS}^2 = \mathrm{Tr}[R_{\ddot{Y}} R_{\ddot{X}} - 2 R_{\ddot{Y}} R_{\ddot{X}} R_Z + R_{\ddot{Y}} R_Z R_{\ddot{X}} R_Z] \tag{3.11}$$

The unconditional fairness score can similarly be estimated empirically as follows (note that unconditional dependence does not entail using extended variables):

$$||\hat{V}_{YX}||^2_{HS} = \text{Tr}[R_Y R_X] \tag{3.12}$$

3.3.2 Connections to Other Strong Measures

Many works in algorithmic fairness have proposed different fairness measures. We are primarily concerned with other strong measures of fairness, which includes Mutual Information (MI) and Maximal Correlation Coefficient (MCC). In this section, we hope to make clear the theoretical relationships between Fair-COCCO and other measures and illustrate the factors that make Fair-COCCO superior. We include a brief overview in this section and formally derive the connections in Appendix A.

Fair-COCCO, MCC and (conditional) MI are the three main 'strong' characterisations of (conditional) dependence ('strong' in the sense defined by Daudin (1980)). Representative works in algorithmic fairness based on MCC include Baharlouei et al. (2019); Mary et al. (2019), which both rely on Witsenhausen's characterisation (Witsenhausen, 1975) of MCC. In the MI domain, representative works include Steinberg et al. (2020a,b).

Fair-COCCO and MCC, while different approaches in principle, are equivalent ideal regularisers. Additionally, Fair-COCCO is an upper bound on MI. The key advantages of Fair-COCCO lie in having the superior empirical estimator, which is straightforward to compute and does not rely on further approximations, thus retaining all theoretical guarantees around detecting dependence. On the other hand, the methods developed by Baharlouei et al. (2019); Mary et al. (2019) involve multiple levels of relaxations. Even in the relaxed formulation, it is not straightforward to estimate for continuous variables, is infeasible for higher numbers of dimensions due to the requirement of explicit density estimation, and relies on discretising continuous intervals for efficient marginalisation. The same challenges exist for MI-approaches (Steinberg et al., 2020a), due to the requirement of probability density and intractability of marginalisation over continuous variables. This is reflected in the superior performance and better accuracy-fairness tradeoff characteristics of Fair-COCCO on a benchmark of experiments.

3.3.3 On Computation of Fair-COCCO

While general kernel dependence measures depend not only on variable distributions, but also the choice of kernel, Fukumizu et al. (2007) proved that in the limit of infinite data and assumptions on richness of the RKHS (dense in the sense of continuous functions on $\mathcal{X}$ with the supremum norm, see Hofmann et al. (2008)), the estimates converges to a kernel-independence value. For our experiments, we use a Gaussian RBF kernel, $k(X_i, X_j) = \exp\left(-\frac{||X_i - X_j||^2}{2\sigma^2}\right) \forall\, i, j \in N$, where σ is the tuneable bandwidth parameter. We employ the median heuristic introduced by Schölkopf et al. (2002), i.e. $\sigma = median\{X_i - X_j, \forall\, i \neq j \in N\}$ to select the bandwidth.

A limitation of the kernel-based measure is the computational complexity. We address this in two ways, 1) by employing a low-rank Cholesky decomposition of the Gram matrix (of rank r), resulting in $\mathcal{O}(r^2 N)$ complexity and 2) by estimating the score on subsets of each minibatch. We empirically demonstrate that these lead to strong results on real-world experiments.

Statistical Testing

Equation (3.5) can be employed as a test-statistic for CI testing where the null hypothesis is conditional independence $H_0 : X \perp\!\!\!\perp Y|Z$ and the alternative hypothesis $H_1 : X \not\perp\!\!\!\perp Y|Z$. As the distribution of the measure under the null hypothesis is unknown, local permutation is employed to determine the rejection region. As the statistical test is not a novel contribution or main focus of this work, we defer to Gretton et al. (2007); Tillman et al. (2009); Zhang et al. (2012) for full review.

3.4 Fair Policies in One-Shot Decision Making

Having introduced Fair-COCCO as a means to evaluate fairness, we now turn our attention to how it can be incorporated into *fair policy learning*. Existing approaches can be categorised into three main types: prior to modelling (pre-processing), during modelling (in-processing) or after modelling (post-processing) (del Barrio et al., 2020). We focus on developing in-processing techniques, which achieve fairness by incorporating either constraints or

Table 3.1 Comparison of Fair-COCCO with fairness-aware learning algorithms. (1) can handle continuous targets, (2) can handle continuous attributes, (3) can handle multiple attributes, (4) can be developed as a test statistic. MI=mutual information, CC=conditional covariance, MCC=Maximal Correlation correlation, LL=linear loss, KM=kernel measure.

Method	Regularisation	Notion	Models	(1)	(2)	(3)	(4)
Zemel et al. (2013)	MI	All	Linear	✗	✗	✗	✗
Zafar et al. (2017)	CC	DP	Linear/Kernel	✗	✗	✗	✗
Donini et al. (2018)	LL	EO	Linear/Kernel	✗	✗	✗	✗
Cho et al. (2020)	DEO	All	Any	✗	✗	✗	✗
Mary et al. (2019)	MCC, MI	All	Any	✓	✓	✗	✗
Steinberg et al. (2020b)	MI	All	Any	✗	✓	✗	✗
Pérez-Suay et al. (2017)	KM	DP	Linear/Kernel	✓	✓	✓	✓
Fair-COCCO	KM	All	Any	✓	✓	✓	✓

regularisers. An interesting future direction could be employing Fair-COCCO for pre-processing to learn fair representations or post-processing, to perform post-hoc fairness driven corrections.

In this section, we consider one-shot decision making scenarios or those with a single interaction, where each data-point can be considered *i.i.d.*. This fits in the domain of conventional supervised learning tasks, which much of algorithmic fairness literature focuses on. Note that the Fair-COCCO score is differentiable, which means it can be employed as a regulariser in any gradient-based method. We compare our approach to related methods, which we have already reviewed in §3.1.1, in Table 3.1. Formally, given N training triplets $\{(X_i,Y_i,A_i)\}_{i=1}^N$, method $f(\cdot;\theta) : \mathcal{X} \to \mathcal{Y}$ and original training loss $\mathcal{L}(\cdot)$, the Fair-COCCO score is computed for each minibatch and can be added to the original training objective. For the EO notion of fairness, this gives:

$$\arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(\theta, f(X_i;\theta), Y_i) + \lambda I(f(X;\theta), A|Y) \tag{3.13}$$

where the first term is the original loss (possibly including other regularisation) and the second term term is the Fair-COCCO regularisation. Hyperparameter $\lambda > 0$ determines the fairness-performance trade-off: a higher λ guarantees higher fairness, but this typically leads to lower predictive performance.

Chapter 4

Fair-PoLe: Fair Policy Learning

In Chapter 3, we detailed Fair-COCCO as a measure to evaluate fairness. We also illustrated how Fair-COCCO can be applied for *fair policy learning* (Fair-PoLe) in the simpler one-shot decision making tasks. In this chapter, we move on to the more challenging sequential setting where temporal effects and long-term outcomes have to be taken into account. We model these dynamics through Markov Decision Processes (MDPs), introducing fairness definitions to policies in MDPs. We then propose a novel inverse reinforcement learning algorithm that can learn to imitate expert actions while maintaining fairness outcomes in sequential decision settings, modelled as MDPs.

4.1 Fairness in Sequential Decision Making

It is interesting to wonder whether we can simply apply fairness techniques developed in the previous chapters to sequential settings—indeed, if each state-action pair is considered *i.i.d.*, fairness can be evaluated on the outcomes of each individual action. If we continue to flesh out this idea, fair policies can be learned by naïvely enforcing fairness conditions on action predictions $\hat{a}_t$, sensitive attribute, z_t and/or true outcome y_t, depending on the fairness notion enforced. Upon further examination, we uncover two main problems with this setup: $1)$ the *i.i.d.* assumption does not hold in the sequential setting, resulting in the even larger problem that $2)$ enforcing fairness at the action level cannot guarantee fairness over the sequence of interaction.

The *i.i.d.* assumption does not hold in sequential settings as future state distributions depend on predictions made in the current step. Ross and Bagnell (2010) proved that this naïve *i.i.d.* assumption leads to compounding errors that grow quadratically with the time horizon. The assumption also ignores long-term planning in sequential settings, ignoring that agents will frequently take actions that generate negative utility in the current step to accomplish a long-term goal.

There are also several issues with the second assumption. Firstly, imposing fairness constraints without considering feedback effects of decisions actually hurts fairness. Several studies, including Creager et al. (2020), Liu et al. (2018) and D'Amour et al. (2020) have demonstrated that even ignoring one-step feedback, common fairness criteria will not longer promote fairness over time. A second challenge is that the true outcome of individual actions, which is required to evaluate certain fairness notions, is not always known. In many real-world settings, rewards are sparse and only available when the sequence terminates. Consider patients in an ICU, we have access to terminal statistics (e.g. whether the patient successfully recovered) but it is highly non-trivial to extract true outcomes of individual actions taken by clinicians.

Evidently, temporal effects need to be modelled to incorporate long-term outcomes into fairness evaluations. We incorporate dynamics by using MDPs and introduce novel ways to quantify group fairness in the MDP setting.

Setup Unless otherwise stated, we continue using the same notation established in §2.2. We denote by MDP/RT a MDP without access to the reward function $R(\cdot,\cdot)$ and transition dynamics $T(\cdot|\cdot,\cdot)$. Let $\mathcal{D} = \{\tau_i\}_{i=1}^N$ be N logged trajectories of expert demonstrations. Each trajectory $\tau_i = \{s_0, a_0, \ldots, s_{T_i}, a_{T_i}\}$ records the state-action pairs, where T_i is the max time horizon. Let each state $s \in \mathcal{S}$ consist of a feature vector $\tilde{s} \in \tilde{\mathcal{S}}$ and $z \in \mathcal{Z}$ that encode the sensitive attribute, i.e. $s = (\tilde{s}, z)$ where the state space can be described by $S = \tilde{S} \times Z$. We further denote by $J(\pi)$ the infinite horizon discounted total return of a policy π:

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right] \tag{4.1}$$

Here, the expectation is taken with respect to $\tau \sim \pi$, indicating that the distribution over trajectories depends on π: $s_0 \sim \mu, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim T(\cdot|s_t, a_t)$. We define another concept to aid explanation in this chapter:

$$J_z(\pi) = J(\pi) \mid s_0 = (z, \tilde{s}_0) \tag{4.2}$$

This is the value of a policy conditioned on starting in initial state s_0 and sub-group $z \in \mathcal{Z}$. We note that the condition of belonging to sub-group z is static as sensitive attributes generally do not change in a logged trajectory. Thus, $J_z(\pi)$ essentially denotes the expected value of a policy for sub-group z.

4.1.1 Fairness Notions for Sequential Settings

In this section, we introduce group fairness notions for the MDP setting. As opposed to defining fairness on an action level (i.e. whether a particular a_t was fair given the features (z, s_t)), we define fairness at a policy level, that is, we consider whether expected value of a policy is fair with respect to the sensitive attributes. Evaluating fairness at a policy level introduces transition dynamics and long-term outcomes into the evaluation, making it more suitable for sequential settings. We introduce definitions of fairness based on notions of DP and EO in MDPs.

Definition 4.1.1 (Demographic Parity in MDP) *We say a policy π satisfies demographic parity (DP) if:*

$$J(\pi) \perp Z \tag{4.3}$$

the expected value of the policy is independent of the sensitive attributes.

Definition 4.1.2 (Equalised Odds in MDP) *We say a policy π satisfies equalised odds (EO) if:*

$$J(\pi) \perp Z|Y \tag{4.4}$$

the expected value of a policy is independent of sensitive attributes given the true outcome.

We emphasise that Y in Definition 4.1.2 is not defined on the action level but rather the true outcome after a sequence terminates. Thus, by considering defining fairness on expected value and outcomes of a sequence of ecisions, we address the difficulties of enforcing fairness in sequential settings.

4.1.2 Fairness Evaluation in Sequential Settings

Based on Definitions 4.1.1 and 4.1.2, it is not immediately obvious how Fair-COCCO can be applied to evaluate fairness in MDP settings. The key challenge lies in estimating the discounted total return $J(\pi)$ of a policy π, which must be carried out using dataset $\mathcal{D}$ of logged expert trajectories.

We apply the *Inverse Reinforcement Learning* (IRL) framework to perform this analysis. Recallig our review from §2.3, IRL offers a principled way of modelling human behaviour by recovering the unknown reward function being maximised when given $\mathcal{D}$ (Abbeel and Ng, 2004; Ng et al., 2000; Ziebart et al., 2008). The proposed method is not limited in the type of IRL procedure that can be applied. Once the reward function is learned, the expected value of the policy $J_z(\pi)$ for a sub-group $z \in \mathcal{Z}$ can be estimated empirically:

$$\hat{J}_z(\pi) = \frac{1}{N} \sum_{\tau_i \in \mathcal{D}} \sum_{t=0}^{T_i} \left[\gamma^t R(s_t, a_t)\right] \mid s_0 = (z, \tilde{s}_0) \tag{4.5}$$

$$\tag{4.6}$$

As before, we make make the connection with group fairness notions explicit, Fair-COCCO scores for different notions in the sequential setting can be expressed as:

$$I_{EO} = ||V_{ZJ(\pi)|Y}||^2_{HS}, \;\; I_{CAL} = ||V_{ZY|J(\pi)}||^2_{HS}, \;\; I_{DP} = ||V_{ZJ(\pi)}||^2_{HS}. \tag{4.7}$$

4.2 Fair-PoLe: An Inverse Reinforcement Learning Algorithm

Having established our definitions of fairness in the sequential settings, we now turn to the problem of *fair policy learning* (Fair-PoLe) or learning a policy that maximises the expected utility while maintaining fairness outcomes.

The task of learning a policy to maximise the cumulative rewards in stochastic environments is usually formulated as a reinforcement learning (RL) problem. RL methods have been successfully applied to multiple sequential decision making problems, including in the healthcare domains (Weng et al. (2017), Parbhoo et al. (2017) and Daskalaki et al. (2016)). However, existing RL applications are grounded by the requirement of explicit rewards signals to guide the RL agents on the goals to be achieved and ability to interact with an enviornment, both requirements generally not satisfied in the real-world.

In such situations, it is necessary to consider an approach to RL where the policy can be learned from a set of presumably optimal expert trajectories. This means that we must consider solutions in the *inverse reinforcement learning* (IRL) framework. We present Fair-PoLe as an IRL algorithm with goals of imitating the actions of the experts while ensuring fair outcomes. We are especially interested in fair policy learning in the minimal possible setting, with imperfect knowledge of the environment dynamics and inability to interact with the environment to test policies. This essentially means we are restricted to learning solely on the basis of observational data in the form of logged trajectories, making this an *offline* IRL task.

§2.3.3 highlighted some of the limitations of existing approaches, including repeated inner-loop calls to MDP solves, linear reward environments and feasibility only in small, solve-able environments. We address the aforementioned challenges by setting our sights on three key criteria:

- Operate in the minimal possible setting, where environment dynamics are unknown and interactions online are infeasible. We propose learning rewards in a *completely offline* fashion,
- Avoid *inner-loop MDP solves* in each IRL iteration, which is computationally costly and untenable for high-dimensional environments,
- Overcome the assumptions on linear reward functions and derive more *functionally expressive* reward function approximators.

One of the key contributions of Fair-PoLe is that it can naturally address all three concerns, without additional requirements or assumptions. In subsequent sections, we develop different aspects of the algorithm before finally putting it together.

One of the fundamental components of Fair-PoLe is a well-known relationship between the action-value function and the reward function. We forego learning an explicit reward function but obtain an implied reward given as the difference between the Q-values of the current step and the expected, discounted Q-values in the subsequent step. This is a straight-forward re-arrangement of the theory presented in §2.2:

$$Q(s,a) = R(s,a) + \gamma \mathbb{E}_{a' \sim \pi, s' \sim T}[Q(s',a')] \tag{4.8}$$

$$R(s,a) = Q(s,a) - \gamma \mathbb{E}_{a' \sim \pi, s' \sim T}[Q(s',a')] \tag{4.9}$$

Moreover, we use $Q^*_R(s,a)$ to denote the optimal action-value function with the property that π^* is an optimal policy relative to R, if $\pi^*(s) \in \arg\max_{a \in A} Q^*(s,a)\ \forall\ s \in \mathcal{S}$. The connection between Equation 4.9 and $Q^*_R(s,a)$ can be written using the Bellman Optimality Theorem that $\forall s \in \mathcal{S}, a \in \mathcal{A}$:

$$R(s,a) = Q^*_R(s,a) - \gamma \mathbb{E}_{s' \sim T}\left[\max_{a' \in A} Q^*_R(s',a')\right] \tag{4.10}$$

This links the optimal action-value function to the reward function through a simple calculation. Piot et al. (2014); Reddy et al. (2019) take advantage of this formulation to impose sparsity regularisation on the implied reward. Additionally Jarrett et al. (2020) proved that behavioural cloning techniques can be seen as maximising this implicit reward through the energy-based model implied by the corresponding policy.

4.2.1 Maximum Entropy Model of Expert Behaviour

For the model of expert behaviour, we move within the Maximum Entropy (MaxEnt) framework (Ziebart et al. (2008), Levine (2018)). Unlike Max-margin, which by definition assumes that the observed expert behaviours are always optimal, the probabilistic formulation of MaxEnt allows for sub-optimal expert behaviour, which is common in the real-world involving human demonstrators. The optimal policy in this framework is parameterised by a Boltzmann distribution:

$$\pi(s,a) = \frac{\exp(Q(s,a))}{\sum_{a' \in A} \exp(Q(s,a'))} \tag{4.11}$$

We introduce a function approximator $Q_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ parameterised by θ to predict the Q-function. The θ parameters can be trained by directly maximising the likelihood of the input data. In the next section, we introduce a fairness regularisation term on the learned reward in Equation 4.10. As this expression involves the Bellman operator (maximisation term), it is not differentiable. We address this by substituting the Q-functions with *soft Q functions*, given by the soft Bellman equations:

$$Q(s,a) = R(s,a) + \gamma \mathbb{E}_{s' \sim T}\left[\log\left(\sum_{a' \in A} \exp(Q(s',a'))\right)\right] \tag{4.12}$$

We note that this expression essentially repllaces the maximisation term with a *softmax* function. Next, we briefly sketch the maximum likelihood procedure. Given the Markov assumption, we can formulate the probability of a trajectory τ as such:

$$p(\tau|\theta) = p(s_0^\tau) \prod_{t=0}^{T_\tau - 1} \left[\pi_\theta(s_t^\tau, a_t^\tau) T(s_{t+1}^\tau | s_t^\tau, a_t^\tau)\right] \tag{4.13}$$

Further assuming that trajectories are *i.i.d.*, the likelihood of demonstrations in $\mathcal{D}$ can be expressed as:

$$L_\theta(\mathcal{D}) = \log \prod_{\tau \in \mathcal{D}} p(\tau|\theta) = \sum_{\tau \in \mathcal{D}} \log p(\tau|\theta) \tag{4.14}$$

4.2.2 Fairness Regularisation

Training an imitation learning algorithm directly using Equation 4.14 corresponds to finding a model π_θ that maximises the log likelihood. However, the optimisation procedure will not yield a valid soft Q function which satisfies the Bellman conditions in Equation 4.12. Indeed, it is easy to see that this loss corresponds to the standard behavioural cloning (BC) setup. The loss does not incorporate any temporal dynamics and the training procedure encourages greedily maximising Q-values assigned to demonstrated actions, without considering delayed impacts. The naive assumptions about state distributions leads to compounding errors as the underlying distribution shifts (Ross and Bagnell, 2010).

We introduce fairness regularisation that adds constraints ensuring the learned Q-vaoues are valid with respect to the reward function. More importantly, the fairness constraints are placed on expected value of policy for sub-groups that we defined in §4.1.1, allowing fair policies to be learned while accounting for temporal effects. Under the soft Q framework, the value of following a policy can be empirically estimated by substituting:

$$\begin{aligned}\hat{J}_z(\pi) &= \frac{1}{N}\sum_{\tau_i \in \mathcal{D}}\sum_{t=0}^{T_i}\left[\gamma^t R(s_t^{\tau_i}, a_t^{\tau_i})\right] \mid s_0 = (z, \tilde{s}_0) \\ &= \frac{1}{N}\sum_{\tau_i \in \mathcal{D}}\sum_{t=0}^{T_i}\left[\gamma^t \left(Q(s_t^{\tau_i}, a_t^{\tau_i}) - \gamma \mathbb{E}_{s' \sim T}\left[\log\left(\sum_{a' \in A}\exp(Q(s', a'))\right)\right]\right)\right] \mid s_0 = (z, \tilde{s}_0)\end{aligned} \tag{4.15}$$

To achieve fairness-aware learning, we introduce the regularisation term $I(J(\pi), A|Y)$ (for EO, or alternatively $I(J(\pi), A)$ for DP). The regularisation term injects information about the state transition dynamics into the imitation learning objective, as $R(s,a)$ is a function of an expectation over next state $s' \sim T(\cdot|s,a)$. This encourages learning long-horizon behaviour instead of greedy maximisation in standard BC.

Regularisation to Resolve Degeneracy

IRL in unregularised MDPs have been shown to be an ill-posed problem. As an example, any constant reward function can perfectly rationalise observed expert behaviour. Additionally Ng et al. (2000) demonstrated that multiple reward hypothesis meet the criteria of being a solution. This obstacle has been approached from the perspective of regularisation theory, which have been adopted for ill-posed problems in general (see the theory presented by Tikhonov (1963)). Tikhonov (1963) saw regularisation as a constraint on the hypothesis space when approximating functions from sparse data. Indeed, if the hypothesis space is unconstrained, it is possible to find functions that exactly fit the data but lack generalisation ability (Vapnik, 1999).

Specifically in the context of IRL, BC can be seen as unconstrained search, which maximally fits the data given but can fail to generalise when encountering new data or when the distribution changes. MaxEnt IRL (Ziebart et al. (2008), Ziebart (2010)) and its variants (Finn et al. (2016), Ho and Ermon (2016)) use the Shannon Entropy as regulariser on the policy and can address the degeneracy issue by maximising expert's return along with entropy of the expert policy. Jeon et al. (2020) generalised MaxEnt IRL and demonstrated that IRL with a class of strongly convex policy regularisers $\Omega : \pi(\cdot|s) \to \mathbb{R}$ does not suffer from a degenerate solution (Geist et al., 2019). Formally:

Theorem 4.2.1 (Unique Solutions in Regularised IRL (Jeon et al., 2020)) *IRL with a class of strongly convex regularisers $\Omega : \pi(\cdot|s) \to \mathbb{R}$:*

$$IRL_\Omega(\pi_E) := \arg\max_R \left\{ J_\Omega(R, \pi_E) - \max_\pi J_\Omega(R, \pi) \right\} \tag{4.16}$$

were π_E is the expert policy, does not suffer from degenerate solutions since there is a unique optimal policy in the regularised MDP (Geist et al., 2019).

Viewed through the lens of regularisation, Fair-PoLe can be interpreted as a policy learning method with regularisation on implied rewards, requiring them to be fair and consistent across temporal dynamics. Thus, if a convex regulariser is adopted, IRL will no longer suffer from degenerate solutions. Unfortunately, our Fair-COCCO regulariser rely on non-linear mappings to the Hilbert Space, rendering it not convex. We see this as an interesting direction to explore in future works. Conceptually strong characterisations of fairness dependence generally result in non-convex regularisation, alternative measures that are weak but convex are covariance and conditional covariance in the original Euclidean spaces:

$$||\Sigma_{J(\pi),Z}||^2_{\mathcal{F}} \tag{4.17}$$

$$||\Sigma_{J(\pi),Z|Y}||^2_{\mathcal{F}} \tag{4.18}$$

where $||\cdot||_{\mathcal{F}}$ denotes the Frobeniuns norm. We note that these measures are equivalent to Fair-COCCO without the non-linear extensions to higher-order moments.

4.2.3 Putting It All Together

We wrap up our discussion on Fair-PoLe by putting the components together. One aspect that we have left out thus far is how the expectation term in Equation 4.15 is computed. The expectation is taken with respect to transition dynamics $T(\cdot|s,a)$, which is intractable to compute in continuous cases. We consider two approaches to address this.

Monte-Carlo (MC) Sampling We introduce another function approximator $T_\phi : S \times A \to S$ parameterised by ϕ to learn transition dynamics $T(s'|s,a)$. T_ϕ can be trained using gradient descent and Mean-Squared Error Loss problem directly from the data $(s,a,s') \sim \mathcal{D}$. The expectation can then be estimated using MC integration with M samples drawn from T_ϕ:

$$\mathbb{E}_{s'\sim T}\left[\log\left(\sum_{a'\in A}\exp(Q(s',a'))\right)\right] = \frac{1}{M}\sum_{i=1}^{M}\log\left(\sum_{a'\in A}\exp(Q_\theta(s^i,a'))\right), \quad s^i \sim T_\phi \tag{4.19}$$

Sampling from True Dynamics We also consider the special case of using observed transition tuples (s,a,s') in $\mathcal{D}$ to approximate the expectation:

$$\mathbb{E}_{s'\sim T}\left[\log\left(\sum_{a'\in A}\exp(Q(s',a'))\right)\right] = \log\left(\sum_{a'\in A}\exp(Q_\theta(s',a'))\right) \tag{4.20}$$

While the MC sampling approach is performed with respect to the learned transition dynamics T_ϕ, utilising the next observed state in $\mathcal{D}$ can be viewed as sampling from the true dynamics. However, there naturally exists the bias-variance trade-off between the two approaches, which we compare in subsequent experiments.

Next, we give how Fair-COCCO can be employed as a regulariser in Fair-PoLe. Formally, given N training triplets $\{(\tau_i, Y_i, A_i)\}_{i=1}^N$, consisting of trajectories, outcome of trajectory and sensitive attributes, the objective function (for the EO notion of fairness) is given as:

$$\arg\min_\theta\ J(\theta) := -\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\tau_i, A_i, Y_i) + \lambda I(J(\pi),\, A|\, Y) \tag{4.21}$$

Here the first term is the log likelihood as described in Equation 4.14 and the second term is the fairness regularisation. As the function is differentiable, it is amenable to minibatch gradient-based optimisation. The corresponding optimisation process for the MC sampling-variant of Fair-PoLe is presented in Algorithm 1.

Algorithm 1: Fair-PoLe

```
Result: $\theta$, $\phi$
Input: $\mathcal{D}, \gamma, \lambda$, learning rate $\eta$, batch size $B$, and $M$ MC samples;
Initialise $\theta, \phi$;
Train $T_\phi$ using MLE with $\mathcal{D}$;
while not converged do
    Sample $\mathcal{D}_{mini} := \{(\tau_i, A_i, Y_i)\}_{i=1}^{B} \sim \mathcal{D}$;
    Sample $\{s^i\}_{i=1}^{M} \sim T_\phi$ to compute $\frac{1}{M}\sum_{i=1}^{M} \log\left(\sum_{a' \in A} \exp(Q_\theta(s^i, a'))\right)$ ;
    Compute $\hat{J}(\pi)$ using (4.15) and $I_{EO}(\hat{J}(\pi), A|Y)$ using (4.7);
    Compute $J(\theta)$ using $\mathcal{D}_{mini}$, (4.21) and $I_{EO}(\hat{J}(\pi), A|Y)$;
    $\theta \leftarrow \theta + \nabla_\theta J(\theta)$;
end
```

Implicit Rewards and Interpretability Most conventional IRL techniques seek to learn a parameterised reward function $R_\psi(\cdot,\cdot)$, upon which forward RL techniques are applied to recover an optimal policy with respect to the reward function (Abbeel and Ng, 2004; Ng et al., 2000; Ramachandran and Amir, 2007). In our outlined approach, the rewards are learned implicitly, meaning the learned rewards are no longer coupled to a specific parameterised form, affording more flexibility when it comes to interpretation. Once we recover the rewards, any variety of techniques can be applied, thus making our method *model-agnostic* (e.g. statistical testing for increased transparency or decision trees for hierarchical modelling of decision trees are possible).

To conclude, Fair-PoLe is a policy learning method where the implied rewards are regularised at a policy level to ensure fair outcomes. We revisit the key desiderata outlined at the beginning of this section. Evidently, our proposed method works in a stricly offline sense, without requiring knowledge of dynamics or interactions with the environment. Additionally, by formulating rewards as implied in the learned Q-values, we remove the need for expensive inner-loop MDP solve. Lastly, we decouple the learned rewards from a parameterised form, opening up the possibility for a variety of function approximators to interpret and explain the rewards inducing the observed behaviour.

Chapter 5

Insights on Synthetic and Real Datasets

Having proposed Fair-COCO as a fairness measure and regulariser and Fair-PoLe as a fair imitation learning algorithm, we now turn our attention to how it works (and can be applied) in practice. There are a number of areas that require empirical demonstration and so we proceed as follows:

1. We first consider a synthetic validation of Fair-COCCO — demonstrating that Fair-COCCO can *successfully learn a range of fairness trade-offs*, and inspecting the empirical impact this has on predictive distributions.

2. We then take incorporate Fair-COCCO into Fair-PoLe for one-shot decisions. On a range of standard benchmarks we compare to existing methods on protecting single binary attributes and outcomes, resulting in competitive (and usually superior) predictive performance on these tasks while *consistently producing the best fairness score*.

3. Further, we apply Fair-COCCO to real data with multiple attributes and continuous outcomes. This is an area that to the best of our knowledge *no other method naturally extends to*, and one that Fair-COCCO now sets a strong benchmark for future work.

4. Finally we move into the setting of sequential decision making. Here we demonstrate on the problem of sepsis treatment how fair imitation learning is a significant application of Fair-PoLe, establishing the *first known study on fairness in sequential policies*.

In this section, we perform experiments within the EO framework, since it is usually considered the most challenging and it covers the middle-ground between the strict DP and lenient FTU definitions. However, we re-iterate that our method is *framework-agnostic* and attach further results performed under alternative definitions in Appendix B.

The Python implementation for Fair-COCCO is available at https://github.com/tennisonliu/fair-cocco and the implementation for Fair-PoLe is available at https://github.com/tennisonliu/fair-pole.

5.1 Experiments on Fair-COCCO

5.1.1 Inspecting Fairness on Synthetic Data

Experimental Setup We simulate a synthetic binary classification dataset with two non-sensitive and one binary sensitive attribute, encoded into group memberships. For data points belonging to group 1 and label 1, samples are drawn from an isotropic Gaussian with $\mu = (-1,-1), \sigma^2 = 1.0$. Likewise, samples are drawn from Gaussians parameterised by $\mu = (1,1), \sigma^2 = 1.0$, $\mu = (-0.5, 0.5), \sigma^2 = 0.5$ and $\mu = (0.5, 0.5), \sigma^2 = 0.8$ for group 2, label 1, group 1, label 0 and group 1, label 0, respectively. 1000 samples are drawn for each combination, except for group 2, label 1, where 500 samples are drawn. Training standard machine learning models on this dataset will be unfair w.r.t. to group 2, in that the classifier tends to negatively classify the examples in this group.

Evaluation Metrics We are mainly concerned with the performance and fairness trade-offs of our proposed methods. We use accuracy as the key metric for performance evaluation and metrics that were introduced in §2.1.1 to evaluate fairness.

We train a logistic regression model for this task. The unregularised model achieved an accuracy of 79.00% with DEO of 0.38. By contrast, a model trained with $\lambda = 500$ achieved 73.71% accuracy and DEO of 0.07. In Figure 5.1, we show the decision boundaries produced by models with varying levels of regularisation. In Figure 5.2, we plot the distributions of predictions with ground truth label 1 at different regularisation settings. Note that regularisation encourages the distributions to match, better satisfying EO desiderata.

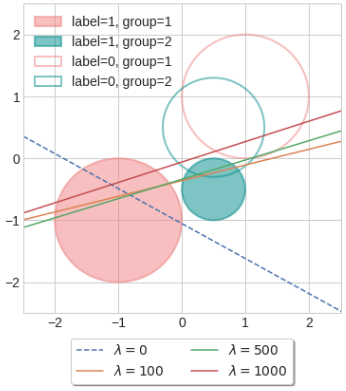
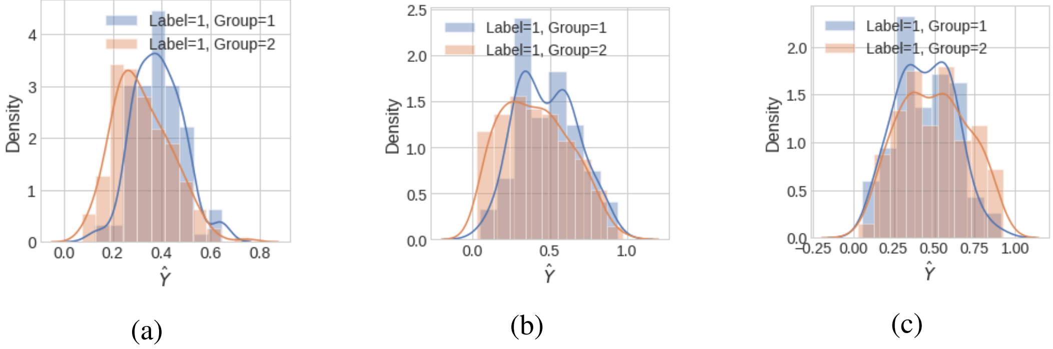


Fig. 5.1 Decision boundaries produced by logistic regression models with varied levels of regularisation ($\lambda = 0, \lambda = 100, \lambda = 500, \lambda = 1000$).



(a)

(b)

(c)

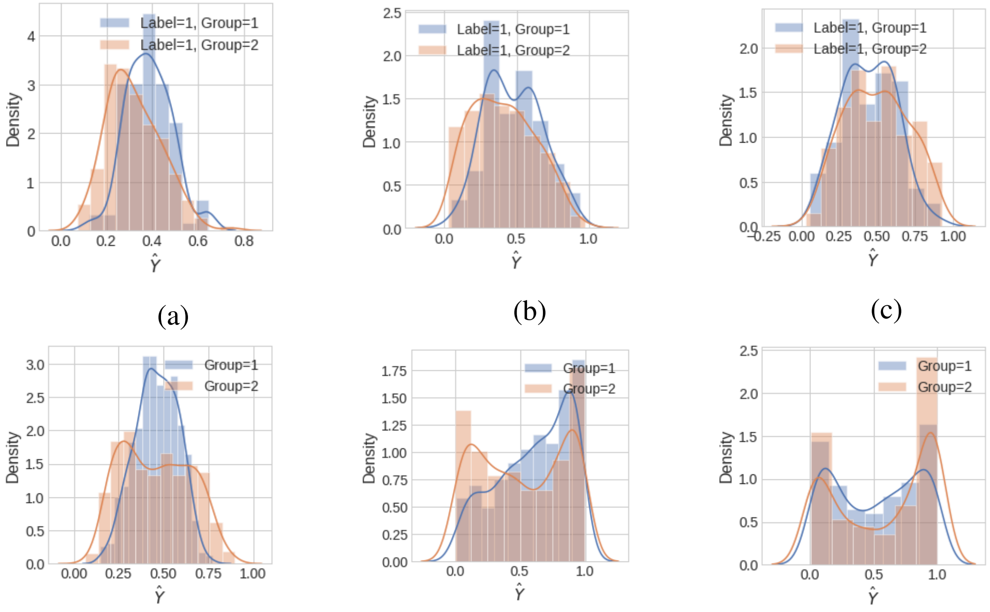
Fig. 5.2 Distribution of predictions for label 1 for different group memberships produced by logistic regression model. (a), (b), (c) correspond to models with varied levels of regularisation ($\lambda = 0, \lambda = 500, \lambda = 1000$, respectively).

5.1.2 Statistical Testing

In this section, we demonstrate how the (conditional) dependence measures on which Fair-COCCO is based on can be employed as a test statistic to perform statistical tests, leading to stronger guarantees and more transparency in practice.

Figure 5.3 shows the distributions of predictions under different fairness constraints. Notably, EO only requires statistical independence between predictions and sensitive attributes given true outcome whereas DP precludes sensitive attributes from having any predictive value by enforcing independence between predictions and attributes.

Table 5.1 reveals the accuracy-fairness trade-offs as well as p-values under different regulation strengths. The p-values indicate the likelihood of observing the data under the null hypothesis (conditional/unconditional independence), see §3.2. As we expect, stronger fairness regularisation leads to lower levels of unfairness as measured by DI and DEO, as well as stronger guarantees in statistical tests. For example, $\lambda = 10000$, we can say with 81% chance that predictions are conditionally independent of sensitive attributes (under EO) or 27% chance that predictions are independent of sensitive attributes (under DP).



(a) (b) (c)

(d) (e) (f)

Fig. 5.3 Top row: distribution of predictions for label 1 of different group memberships under EO. Bottom row: distribution of predictions for different group memberships under DP. Predictions are produced by regularised logistic regression model with $\lambda = 0, \lambda = 500, \lambda = 1000$, respectively across each row.

5.2 Experiments on One-Shot Decision Making

Having validated our method on synthetic data, we now evaluate the performance of Fair-COCCO and benchmarks methods on six datasets from the public UCI ML Repository (Dua and Graff, 2017), whose characteristics are summarized in Table 5.2. A full description of the datasets can be found in Appendix C. The Adult, Drugs, German and COMPAS datasets have binary attributes and outcome, which means we can compare against existing fair learners. The C&C and Students datasets have continuous outcomes with multiple protected attributes; a setting that none of the existing works can handle.

Table 5.1 Accuracy-fairness trade-offs under different fairness notions and corresponding statistical test results.

(a) Accuracy (ACC), DI, COCCO score (scaled by 1e−3), p-value under DP.

λ	ACC	DI	COCCO	p-value
0	78.71	1.08	7.83	0.00
100	78.00	0.82	3.99	0.06
500	75.21	0.85	3.13	0.09
1000	55.30	1.02	3.12	0.18
10000	52.36	0.99	0.37	0.27

(b) Accuracy (ACC), DEO, COCCO score (scaled by 1e−3), p-value under EO.

λ	ACC	DEO	COCCO	p-value
0	78.71	0.41	3.27	0.00
100	72.96	0.03	0.34	0.21
500	70.36	0.03	0.26	0.45
1000	53.54	0.02	0.29	0.57
10000	51.38	0.00	0.17	0.81

Model Details For all experiments, we train a two-layer neural network with ReLU-activated nodes. The number of nodes chosen is between 40∼100 depending on the complexity of the data. The network is trained with Cross Entropy or MSE Loss and is optimised using Adam (Kingma and Ba, 2014). The hyperparameters include batch size $\in \{64, 128, 256\}$, learning rate $\in \{1e{-}2, 1e{-}3, 1e{-}4\}$, and fairness penalty $\in \{10^0, 10^1, 10^2, 10^3\}$ and are chosen through cross-validation. For datasets without a defined test set, the data is split 60-20-20 into train, validation and test set and results are averaged over 10 runs. Experiments are run on either a CPU or NVIDIA Tesla K40C GPU, taking around an hour.

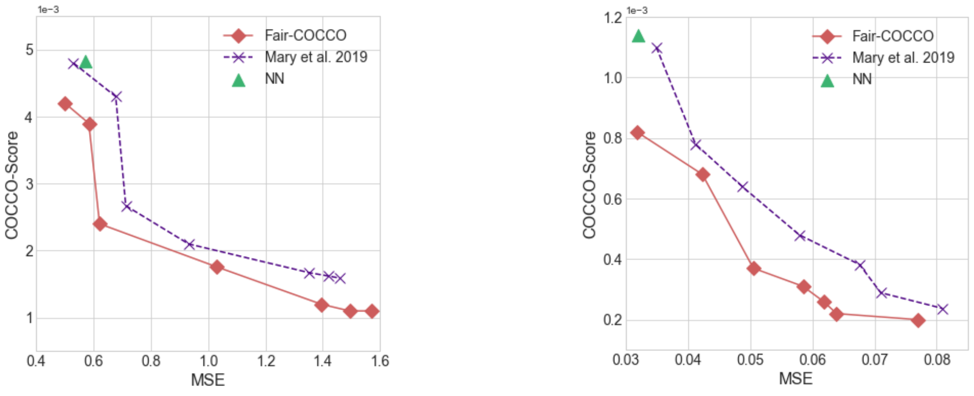
Table 5.2 Description of datasets. '-B' suffix indicates binary variables, '-C' indicates continuous variables.

Dataset	Examples	Features	Sensitive (A)	Outcome (Y)
Adult (Kohavi, 1996)	45222	12	Gender-B	Income-B
Drugs (Mirkes, 2015)	1885	11	Ethnicity-B	Drug use-B
German (Hoffman, 1994)	1700	20	Foreign-B	Income-B
COMPAS (Angwin et al., 2016)	6172	10	Ethnicity-B	Recidivism-B
C&C (Redmond, 2009)	1994	128	Ethnic Proportions-C	Crime Rate-C
Students (Cortez, 2014)	649	33	Age-C, Gender-B	Performance-C

Single Binary Attributes and Outcomes In order to prove Fair-COCCO is competitive with state-of-the-art on binary variables, we reproduce the experiments from (Donini et al., 2018; Mary et al., 2019) based on UCI's Drugs, German, Adult and COMPAS datasets. We also compare against (Hardt et al., 2016; Zafar et al., 2017) and a standard (unfair) neural network (NN).[1] As can be seen in Table 5.3[2]; notably, Fair-COCCO achieves higher levels

[1] None of the other EO-based related works have publicly available code or results for these datasets.

[2] All benchmark results are taken from (Donini et al., 2018) and (Mary et al., 2019). We have re-run their experiments with our own pre-processing pipeline. We were unable to reproduce the results reported by Mary et al. (2019) on COMPAS, and include their reported scores instead.



(a) (b)

Fig. 5.4 Fairness-accuracy trade-off curve for single sensitive attribute ‘racePctBlack’ on C&C (a) and ‘age’ on Students (b). Note, the optimum desiderata is at the origin, where both MSE and unfairness are minimised.

of fairness (lower DEO) while maintaining strong predictive accuracy on large and small datasets.

Table 5.3 Accuracy (ACC) and DEO on benchmark of relevant algorithms on datasets. Results are aggregated over 10 runs for datasets without a designated test set. *NN* is an unregularised neural network and *Fair-COCCO* is the NN with our proposed regularisation.

	COMPAS		German		Drug		Adult	
Method	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO
Zafar et al. (2017)	0.69 ± 0.02	0.10 ± 0.06	0.62 ± 0.09	0.13 ± 0.11	0.69 ± 0.03	0.02 ± 0.07	0.78	0.05
Hardt et al. (2016)	0.71 ± 0.01	0.08 ± 0.01	0.71 ± 0.03	0.11 ± 0.18	0.75 ± 0.11	0.14 ± 0.08	0.82	0.11
Donini et al. (2018)	0.73 ± 0.01	0.05 ± 0.03	0.73 ± 0.04	0.05 ± 0.03	0.80 ± 0.03	0.07 ± 0.05	0.81	0.01
Mary et al. (2019)	0.96 ± 0.00	0.00 ± 0.00	0.72 ± 0.03	0.21 ± 0.15	0.80 ± 0.04	0.00 ± 0.01	0.83	0.08
NN	0.90 ± 0.00	0.01 ± 0.00	0.74 ± 0.07	0.11 ± 0.35	0.80 ± 0.08	0.06 ± 0.12	0.84	0.19
Fair-COCCO	0.89 ± 0.00	0.00 ± 0.00	0.73 ± 0.10	0.02 ± 0.02	0.80 ± 0.06	0.01 ± 0.01	0.83	0.04

Single Continuous Attribute Our closest competitor is (Mary et al., 2019), which is not able to handle multiple attributes but does allow for single continuous attributes. We take the datasets C&C and Students, and only use protected attributes `racePctBlack` and `age` respectively. We plot the performance versus fairness by varying the fairness penalty, see Figure 5.4. Observe that Fair-COCCO obtains a better trade-off between fairness and MSE than Mary et al. (2019).

Multiple Attributes and Continuous Outcomes Many real world environments require prediction with multiple sensitive attributes, which require simultaneous protection. We illustrate a main contribution of our proposal, by evaluating Fair-COCCO on datasets with multiple sensitive attributes and continuous outcome variables.

For C&C, the sensitive attributes are ethnic proportions (i.e. `racePctBlack`, `racePctWhite`, `racePctAsian`, `racePctHisp`) which are given as continuous values. On Student, the attributes are `gender` (binary) or `age` (continuous). Both problems demand multiple sensitive attributes to be simultaneously protected, which is natural for Fair-COCCO but cannot be done by any existing EO-based method (Table 3.1). We compare Fair-COCCO against (unfair) linear regression and XGBoost models, demonstrating strong accuracy-fairness performance even on multi-dimensional variables (Table 5.4) We also demonstrate how Fair-COCCO scales as a regulariser with the number of sensitive attributes and attach those results in Appendix D.

Table 5.4 MSE and Fair-COCCO score when protecting multiple attributes. Results are aggregated over 10 runs and COCO scores are scaled by 1e−3.

	C&C		Student	
Method	MSE	COCCO	MSE	COCCO
Linear Regression	0.030	1.34	0.280	4.69
XGBoost	0.024	1.95	0.257	3.35
NN	0.031 ± 0.001	1.26 ± 0.12	0.254 ± 0.008	4.39 ± 0.13
Fair-COCCO	0.034 ± 0.003	0.55 ± 0.11	0.303 ± 0.013	2.32 ± 0.20

5.3 Experiments on Sequential Decision Making

Finally, we emphasise that Fair-COCCO is not limited to the one-shot supervised learning setup and demonstrate how our approach can be applied in a sequential decision making scenario for learning fairer policies in imitation learning. This is an area in which it is *especially* important to learn fairly – since the fundamental assumption is that the (usually human) demonstrator is acting *near-optimally*.

However, implicit biases can easily manifest in such settings, as an example it has been well-documented that implicit prejudices held by healthcare professionals affect treatment outcomes for minority groups (Hall et al., 2015). Under no circumstances should this be allowed to leak into a policy that will be used in practice down the line. As such we demonstrate our approach on the MIMIC-III ICU database (Johnson et al., 2016a), containing data routinely collected from adult patients in the United States.

We note that we do not compare Fair-PoLe with other known baselines. The reason for this is simple - there simply do not exist comparable methods for fair policy learning in the offline setting. The only methods that can be reasonably applied are behavioural cloning options, and we introduce multiple variations in an ablation study to demonstrate the potential of our proposed method.

5.3.1 Learning Fair Policies for Sepsis Treatment

Sepsis is one of the leading causes of mortality in intensive care units (ICUs) (Singer et al., 2016) and while efforts have been made to provide guidelines for treatments, physicians at bedside largely rely on experience giving rise to possible variations in fair treatments.

Experimental Setup We analyse the decisions made by clinicians to treat sepsis, using a patient cohort fulfilling the Sepsis-3 criteria, delineated by Komorowski et al. (2018) from the MIMIC-III ICU Database (Johnson et al., 2016a). The key characteristics of the Sepsis-3 Cohort are described in Table 5.5. The dataset contains trajectories for $19,585$ patients in the ICU, where each trajectory contains states (clinical state of patient) and actions (medical interventions) recorded at 4 hour intervals. The trajectories terminate when the patients are discharged or succumb to sepsis, with patients spending on average 53 hours in the ICU, accruing an average of 13 measurements.

For each patient, we have a 45-dimensional feature vector $\tilde{S} \in \mathbb{R}^{45}$ containing static patient information, lab values, vitals measurements and intake/output recordings, recorded at 4 hour resolution. There also exists $\mathcal{Z} \in \{0,1\}$ a binary sensitive attribute corresponding to patient gender. The dataset records the actions taken by clinicians at 4 hour intervals, $\mathcal{A} \in \{0,1,\dots,24\}$, corresponding to 25 discrete interventions based on intravenous (IV) fluid and vasopressor dosage within the 4 hour window. Ground-truth treatment outcomes $\mathcal{Y} \in \{0,1\}$ are recorded at the end of the trajectory for each patient, corresponding to 90-day mortality after discharge. There are no true treatment outcomes for each individual action, and we compute action-level outcomes using methods described by Raghu et al. (2017). For the complete problem setup, refer to Appendix E.

Methods We consider several methods to compare against to demonstrate the potential of Fair-PoLe:

- Behavioural cloning without regularisation (BC). The expert's demonstrations $\mathcal{D}$ are divided into *i.i.d.* state-action pairs,

Table 5.5 Summary of MIMIC-III Sepsis-3 Cohort dataset. Trajectory count is based on the number of unique patients and trajectory lengths reports the average number of measurements (average time) spent in the ICU.

		Trajectory Count	**Trajectory Lengths**	**Total Measurements**
Total		19,585	13.06 ± 4.95	255,755
Gender	Male	10,901	13.10 ± 4.93	142,819
	Female	86,84	13.01 ± 4.98	112,936
Mortality	Survival	17,712	12.87 ± 4.88	228,023
	Death	1,873	14.81 ± 5.28	27,732

- Behavioural cloning with reward sparsity regularisation (BC.RR). This is based on the work of Piot et al. (2014); Reddy et al. (2019), which demonstrated that implicit rewards regularisation can effectively incorporate subsequent dynamics,
- Fair-PoLe with fairness imposed at an action level (Fair-PoLe.A), ensuring individual actions are fair with respect to the protected attribute,
- Fair-PoLe with fairness imposed at a policy level and MC sampling of 3 samples to approximate subsequent states (Fair-PoLe M=3),
- Fair-PoLe with fairness imposed at policy level but using observed next actions as approximation to expectation (Fair-PoLe M=1).

Model Details For all methods, we approximate the Q-value network $Q_\theta(\cdot)$ with a two hidden layer neural network (NN) with ReLU activation. The networks are trained with Cross Entropy Loss and optimised using Adam (Kingma and Ba, 2014). The transition network $T_\phi(\cdot,\cdot)$ is likewise approximated with a three hidden layer NN with ReLU activations. A Gaussian output layer is employed that learns the mean and standard deviations for the transition prediction. The local reparameterisation trick (Kingma et al., 2015), MSE Loss, and Adam are employed to train the transition network. The hyperparameters include batch size $\in \{64, 128, 256\}$, learning rate $\{1\text{e}{-}2, 1\text{e}{-}3, 1\text{e}{-}4\}$ and fairness penalty $\in \{10, 25, 50, 100\}$ and are tuned through cross-validation. We use $\gamma = 0.9$ as the discount factor. The dataset is split into train, validation and test sets such that 60%, 20% and 20% of the *trajectories* are assigned to each set. The results are averaged over 10 runs. Experiments are tun on NVIDIA Tesla K40C GPUs, taking between a couple of hours to complete.

Results Table 5.6 describes the performance-fairness trade-off achieved by different models. We use the same evaluation metrics as detailed in our previous set of experiments

(§5.2) to evaluate accuracy-fairness trade-off. However, there are subtle distinctions—here, accuracy describes the *action matching* performance and COCCO scores are measured at the *trajectory* level (i.e., using Equation 4.7) instead of action-level as in the one-shot setting. We see that the most simplistic behavioural cloning model already achieves strong accuracy, likely as the distribution between train and test sets are quite similar, thus the lack of generalisability of BC methods are less pronounced. Behavioural cloning with reward sparsity regularisation, by incorporating temporal dynamics through reward sparsity, achieves superior accuracy, although this also comes at a higher cost of fairness.

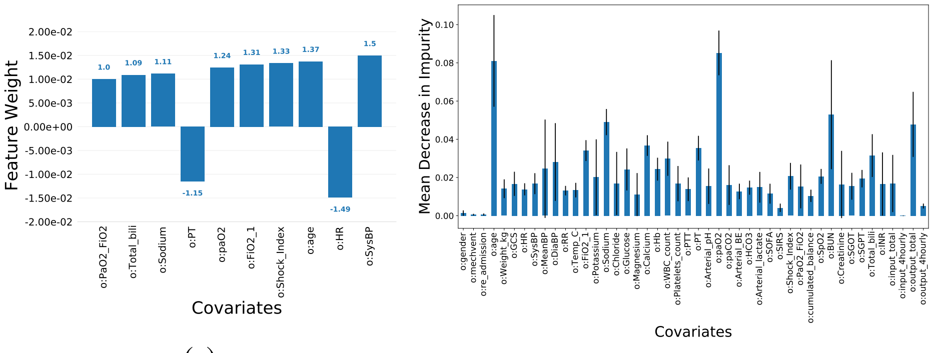
Additionally, we see that while Fair-PoLe.A (imposes fairness at the action level) does not suffer a major decrease in action matching accuracy, does incur the highest fairness penalty. This echoes the findings of D'Amour et al. (2020); Liu et al. (2018), which highlighted the dangers of naïvely imposing fairness constraints without long-term planning. Lastly, we see that the two variations of Fair-PoLe with fairness imposed at the policy level achieve similar trade-off characteristics. Fair-PoLe where the rewards are computed using the observed next state performs better with the lowest fairness penalty, likely as the expectation over subsequent states is with respect to the true transition dynamics, resulting in more accurate estimatinos.

Table 5.6 Accuracy and Fair-COCCO score of different inverse reinforcement learning methods. Results are aggregated over 10 runs and COCCO scores are scaled by $1e-3$.

Model	Accuracy	COCCO
BC	84.18 ± 1.62	2.97 ± 2.12
BC.RR	85.33 ± 1.93	3.12 ± 4.99
Fair-PoLe.A	81.28 ± 1.70	3.23 ± 2.57
Fair-PoLe M=3	78.07 ± 1.93	2.76 ± 2.18
Fair-PoLe M=1	79.44 ± 1.95	2.65 ± 1.69

Interpretation Next, we demonstrate how the implicit reward learning approach adopted by Fair-PoLe allows opens up for a variety of interpretation that is decoupled from the exact parameterisation of $R(\cdot,\cdot)$. We illustrate this by applying linear regression and Random Forest methods to interpret clinical decision making when 1) no intervention is prescribed and 2) when the highest doses of IV and Vasopressors are prescribed.

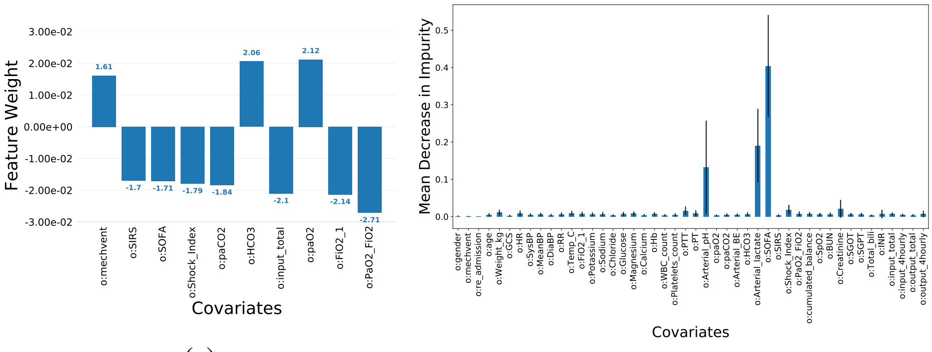
We leave formal analysis and interpretation to domain experts but highlight some notable points. When no treatment is prescribed, which can reasonably be assumed to mean the patient's sepsis condition is under control, more aspects of the patient's condition are taken into consideration (Figure 5.5). However, when the highest doses of drugs are prescribed, key indicators, namely lactate levels and SOFA scores, dominate consideration (Figure 5.6). These correspond the most important mortality indicators for clinicians during sepsis treatment (Liu et al., 2019).



(a)

(b)

Fig. 5.5 Different reward interpretations in scenarios where clinicians did not prescribe any treatment. (a) shows the 10 most important feature weights learned by a linear regression model and (b) ranks feature importance based on Mean Decrease in Impurity returned by a Random Forest Regressor.



(a)

(b)

Fig. 5.6 Different reward interpretations in scenarios where clinicians prescribed the strongest possible combination of IV and Vasopressors. (a) shows the 10 most important feature weights learned by a linear regression model and (b) ranks feature importance based on Mean Decrease in Impurity returned by a Random Forest Regressor.

Chapter 6

Conclusion

In this thesis, we introduced methods to audit fairness of decisions made by humans and algorithms alike. We also focused on evaluating fairness in sequences of decisions, which are modelled as MDPs. Last, we developed fair-policy learning, a regularisation based learning method that is modular and can be incorporated into any gradient optimisation procedure and illustrated learning on one-shot and sequential problems.

In Chapter 3, we proposed Fair-COCCO, a kernel-based fairness measure that quantifies unfairness in the presence of multiple sensitive attributes as well as continuous or discrete outcomes. We made clear the connection between Fair-COCCO and other strong measures of fairness in literature and illustrate a statistical testing method that can lead to stronger guarantees in practice. We empirically demonstrate the ability of Fair-COCCO to capture fairness through simulated studies and demonstrate superior performance in balancing fairness-prediction trade-off on a range of benchmarks.

Then in Chapter 4, we tackle the more challenging problem of algorithmic fairness in sequential decision making, becoming the first work to address this problem. We propose definitions of fairness in sequential settings modelled as MDPs, where fairness is defined on the expected value of a policy for different sub-groups. We introduce a Fair-PoLe, a novel IRL algorithm that is completely offline, avoids costly inner-loop computations and allows for functionally expressive reward function approximators to be learned. This is achieved through implicit reward learning and fairness regularisation placed on fairness at a policy level. In our experiments, we demonstrate the potential of Fair-PoLe to learn fair policies in complex real-world environments on the task of sepsis treatment.

Limitations The main limitation of Fair-COCCO is also the source of its key strength—the matrix operations required to kernalise the data and embed it in the RKHS has complexity $\mathcal{O}(N^3)$. We propose two directions to alleviate this, by using i) a low-rank approximation of the kernel matrix and ii) subsets of minibatches to evaluate the Fair-COCCO score. Empirically, these computational tricks do not noticeably impact performance. As Fair-COCCO is not a convex function, incorporating it was regularisation on rewards in Fair-PoLe cannot fully resolve the degeneracy issues common to IRL problems. We proposed covariance measures on the original feature space as a alternative means of convex regularisation.

Future Works One interesting direction for future work is to evaluate performance of Fair-COCCO on larger datasets or improving its computational characteristics by incorporating the kernel operation speed ups proposed by Zhang et al. (2012). One assumption that we made Fair-PoLe is that the environment is fully-observable and Markovian, an assumption that rarely holds in real-world problems such as healthcare. An interesting project for the future would be to model policies that depend on histories in partially observable environments.

Societal Impact It is encouraging to see that the ML community is paying more attention for algorithmic fairness. To that end, we humbly hope our proposed methods can address some of the challenges currently facing deployment of fair algorithms. Nonetheless, we caution against using our proposed methods as a ‘certificate‘ of fairness. As Corbett-Davies et al. (2017) rightfully emphasise, fairness measures do not rule out blatantly unfair practices. Additionally, future works should focus on interpretable fairness quantification and fairness-aware training to provide better understanding of root causes of unfairness in prediction settings.

References

Abbeel, P. and Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. there's software used across the country to predict future criminals. and it's biased against blacks.

Avery, R. B., Brevoort, K. P., and Canner, G. (2009). Credit scoring and its effects on the availability and affordability of credit. *Journal of Consumer Affairs*, 43(3):516–537.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48.

Baharlouei, S., Nouiehed, M., Beirami, A., and Razaviyayn, M. (2019). R\'enyi fair inference. *arXiv preprint arXiv:1906.12005*.

Bain, M. and Sammut, C. (1995). A framework for behavioural cloning. In *Machine Intelligence 15*, pages 103–129.

Baker, C. R. (1973). Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289.

Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1:2017.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104:671.

Bechavod, Y. and Ligett, K. (2017). Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*.

Bergsma, W. P. (2004). *Testing conditional independence for continuous random variables*. Citeseer.

Bica, I., Jarrett, D., Hüyük, A., and van der Schaar, M. (2020). Learning" what-if" explanations for sequential decision-making. *arXiv preprint arXiv:2007.13531*.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al. (2016). End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Brown, D. S. and Niekum, S. (2019). Deep bayesian reward learning from preferences. *arXiv preprint arXiv:1912.04472*.

Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X. (2013). Controlling attribute effect in linear regression. In *2013 IEEE 13th international conference on data mining*, pages 71–80. IEEE.

Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.

Chan, A. J. and van der Schaar, M. (2021). Scalable bayesian inverse reinforcement learning. *arXiv preprint arXiv:2102.06483*.

Cho, J., Hwang, G., and Suh, C. (2020). A fair classifier using kernel density estimation. *Advances in Neural Information Processing Systems*, 33:15088–15099.

Choi, J. and Kim, K.-E. (2011a). Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research*, 12:691–730.

Choi, J. and Kim, K.-E. (2011b). Map inference for bayesian inverse reinforcement learning. In *NIPS*, pages 1989–1997.

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

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806.

Cortez, P. (2014). UCI machine learning repository.

Creager, E., Madras, D., Pitassi, T., and Zemel, R. (2020). Causal modeling for fairness in dynamical systems. In *International Conference on Machine Learning*, pages 2185–2195. PMLR.

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., and 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*, pages 525–534.

Daskalaki, E., Diem, P., and Mougiakakou, S. G. (2016). Model-free machine learning in biomedicine: Feasibility study in type 1 diabetes. *PloS one*, 11(7):e0158722.

Daudin, J. (1980). Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590.

del Barrio, E., Gordaliza, P., and Loubes, J.-M. (2020). Review of mathematical frameworks for fairness in machine learning. *arXiv preprint arXiv:2005.13755*.

Dewey, D. (2014). Reinforcement learning and the reward engineering principle. In *2014 AAAI Spring Symposium Series*.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. (2018). Empirical risk minimization under fairness constraints. *arXiv preprint arXiv:1802.08626*.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

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

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

Finn, C., Levine, S., and Abbeel, P. (2016). Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*, pages 49–58. PMLR.

Fleisher, W. (2021). What's fair about individual fairness? *Available at SSRN 3819799*.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007). Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

Geist, M., Scherrer, B., and Pietquin, O. (2019). A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR.

Giusti, A., Guzzi, J., Cireşan, D. C., He, F.-L., Rodríguez, J. P., Fontana, F., Faessler, M., Forster, C., Schmidhuber, J., Di Caro, G., et al. (2015). A machine learning approach to visual perception of forest trails for mobile robots. *IEEE Robotics and Automation Letters*, 1(2):661–667.

Goel, N., Yaghini, M., and Faltings, B. (2018). Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., Smola, A. J., et al. (2007). A kernel statistical test of independence. In *Nips*, volume 20, pages 585–592. Citeseer.

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2.

Hacker, P. (2018). Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under eu law. *Common Market Law Review*, 55(4).

Hall, W. J., Chapman, M. V., Lee, K. M., Merino, Y. M., Thomas, T. W., Payne, B. K., Eng, E., Day, S. H., and Coyne-Beasley, T. (2015). Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. *American journal of public health*, 105(12):e60–e76.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. (2018). Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Hirschfeld, H. O. (1935). A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31, pages 520–524. Cambridge University Press.

Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. *Advances in neural information processing systems*, 29:4565–4573.

Hoffman, H. (1994). UCI machine learning repository.

Hoffman, K. M., Trawalter, S., Axt, J. R., and Oliver, M. N. (2016). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences*, 113(16):4296–4301.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *The annals of statistics*, 36(3):1171–1220.

Hussein, A., Gaber, M. M., Elyan, E., and Jayne, C. (2017). Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35.

Jarrett, D., Bica, I., and van der Schaar, M. (2020). Strictly batch imitation learning by energy-based distribution matching. *arXiv preprint arXiv:2006.14154*.

Jeon, W., Su, C.-Y., Barde, P., Doan, T., Nowrouzezahrai, D., and Pineau, J. (2020). Regularized inverse reinforcement learning. *arXiv preprint arXiv:2010.03691*.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. (2020). Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pages 862–872. PMLR.

Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016a). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Johnson, K. D., Foster, D. P., and Stine, R. A. (2016b). Impartial predictive modeling: Ensuring fairness in arbitrary models. *Statistical Science*, page 1.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., Salimans, T., and Welling, M. (2015). Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28:2575–2583.

Klein, E., Geist, M., and Pietquin, O. (2011). Batch, off-policy and model-free apprenticeship learning. In *European Workshop on Reinforcement Learning*, pages 285–296. Springer.

Klein, E., Geist, M., Piot, B., and Pietquin, O. (2012). Inverse reinforcement learning through structured classification. In *NIPS 2012*, pages 1–9.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kohavi, R. (1996). UCI machine learning repository.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720.

Lee, D., Srinivasan, S., and Doshi-Velez, F. (2019). Truly batch apprenticeship learning with deep successor features. *arXiv preprint arXiv:1903.10077*.

Levine, S. (2018). Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*.

Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. (2018). Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR.

Liu, Z., Meng, Z., Li, Y., Zhao, J., Wu, S., Gou, S., and Wu, H. (2019). Prognostic accuracy of the serum lactate level, the sofa score and the qsofa score for mortality among adults with sepsis. *Scandinavian journal of trauma, resuscitation and emergency medicine*, 27(1):1–10.

Lowy, A., Pavan, R., Baharlouei, S., Razaviyayn, M., and Beirami, A. (2021). Fermi: Fair empirical risk minimization via exponential r\'enyi mutual information. *arXiv preprint arXiv:2102.12586*.

Markham, A. and Others (2012). Ethical decision-making and Internet research: Version 2.0. *Association of Internet Researchers*.

Mary, J., Calauzenes, C., and El Karoui, N. (2019). Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pages 4382–4391. PMLR.

Mirkes, E. (2015). UCI machine learning repository.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Ng, A. Y., Russell, S. J., et al. (2000). Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2.

Niekum, S., Osentoski, S., Konidaris, G., Chitta, S., Marthi, B., and Barto, A. G. (2015). Learning grounded finite-state representations from unstructured demonstrations. *The International Journal of Robotics Research*, 34(2):131–157.

Parbhoo, S., Bogojeska, J., Zazzi, M., Roth, V., and Doshi-Velez, F. (2017). Combining kernel and model based learning for hiv therapy selection. *AMIA Summits on Translational Science Proceedings*, 2017:239.

Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., and Camps-Valls, G. (2017). Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 339–355. Springer.

Piot, B., Geist, M., and Pietquin, O. (2014). Boosted and reward-regularized classification for apprenticeship learning. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1249–1256.

Piot, B., Geist, M., and Pietquin, O. (2016). Bridging the gap between imitation learning and inverse reinforcement learning. *IEEE transactions on neural networks and learning systems*, 28(8):1814–1826.

Podesta, J. and Others (2014). Big data: Seizing opportunities, preserving values. *Executive Office of the President*.

Pomerleau, D. A. (1989). Alvinn: An autonomous land vehicle in a neural network. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA ARTIFICIAL INTELLIGENCE AND PSYCHOLOGY ....

Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P., and Ghassemi, M. (2017). Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*.

Ramachandran, D. and Amir, E. (2007). Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pages 2586–2591.

Reddy, S., Dragan, A. D., and Levine, S. (2019). Sqil: Imitation learning via reinforcement learning with sparse rewards. *arXiv preprint arXiv:1905.11108*.

Redmond, M. (2009). UCI machine learning repository.

Reed, M. and Simon, B. (1980). Functional analysis. revised and enlarged edition. *Methods of Modern Mathematical Physics, Academic Press*.

Rényi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3-4):441–451.

Rogers, W. and Ballantyne, A. (2008). Gender and trust in medicine: vulnerabilities, abuses, and remedies. *IJFAB: International Journal of Feminist Approaches to Bioethics*, 1(1):48–66.

Ross, S. and Bagnell, D. (2010). Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 661–668. JMLR Workshop and Conference Proceedings.

Schölkopf, B., Smola, A. J., Bach, F., et al. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., et al. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810.

Steinberg, D., Reid, A., and O'Callaghan, S. (2020a). Fairness measures for regression via probabilistic classification. *arXiv preprint arXiv:2001.06089*.

Steinberg, D., Reid, A., O'Callaghan, S., Lattimore, F., McCalman, L., and Caetano, T. (2020b). Fast fair regression via efficient approximations of mutual information. *arXiv preprint arXiv:2002.06200*.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tikhonov, A. N. (1963). On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, volume 151, pages 501–504. Russian Academy of Sciences.

Tillman, R. E., Gretton, A., and Spirtes, P. (2009). Nonlinear directed acyclic structure learning with weakly additive noise models. In *NIPS*, pages 1847–1855. Citeseer.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.

Weng, W.-H., Gao, M., He, Z., Yan, S., and Szolovits, P. (2017). Representation and reinforcement learning for personalized glycemic control in septic patients. *arXiv preprint arXiv:1712.00654*.

Witsenhausen, H. S. (1975). On sequences of pairs of dependent random variables. *SIAM Journal on Applied Mathematics*, 28(1):100–113.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*.

Ziebart, B. D. (2010). *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., et al. (2008). Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA.

Appendix A

Connection to Other Fairness Measures

We discuss the theoretical connections between Fair-COCCO and Mutual Information (MI) and Maximal Correlation Coefficients (MCC) in depth here. There exists several main approaches to characterising conditional dependence, here we sketch the relevant ones:

- Daudin (1980) characterises dependence as the correlations of functions in certain L^2 spaces,
- Kernel-theory based measures (our work). This characterisation is related to Daudin's in that instead of considering all functions in L^2, this measure exploits RKHSs corresponding to some characteristic kernels, which are much smaller. Alternatively, if we restrict the functions considered in the L^2 to those belonging to the RKHS, Daudin's approach reduces to the kernel-based measure,
- Conditional Mutual Information (CMI), which can be concisely expressed mathematically as $X \perp Y|Z \iff I(X;Y|Z) = 0$.

Having established the general approaches, we elaborate on the exact connection between the measure we use and others:

Maximum Correlation Coefficient At a high level, both Fair-COCCO and the maximum correlation coefficient (MCC) fall within the framework set out by Rényi (Rényi, 1959), namely that for sufficiently rich function classes, the functional correlation (or, alternatively, the cross-covariance) can be used to measure (conditional) dependence.

MCC is defined as the supremum of Pearson's correlation coefficient ρ over all Borel-measurable functions of finite variance. Formally:

Definition A.0.1 (Maximal Correlation Coefficient (MCC)) *Given two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, the Maximal Correlation Coefficient is defined as follows:*

$$MCC(X,Y) = \sup_{f,g} \rho(f(X), g(Y)) \tag{A.1}$$

where ρ is the Pearson's correlation coefficient, f and g are measurable functions with $\mathbb{E}[f^2(X)], \mathbb{E}[g^2(Y)] < \infty$.

Evidently, MCC fits the characterisation of dependendence provided in Daudin (1980). However, as the supremum is computed over an infinite-dimensional space, MCC is not directly computable. Baharlouei et al. (2019); Mary et al. (2019) employ Witsenhausen's characterisation (Witsenhausen, 1975), which calculates MCC as the second largest singular value of a stochastic matrix Q. As estimating the singular values is difficult, Mary et al. (2019) instead employ the χ^2 divergence as the upper-bound of MCC:

$$MCC(X,Y)^2 \leq \int \frac{\pi(x,y)}{\pi(x)\pi(y)} dx dy = \chi^2(\pi(x,y), \pi(x) \otimes \pi(y)) \tag{A.2}$$

To avoid cumbersome notation, we let $\pi(x,y)$ be shorthand for the joint distribution $\pi_{XY}(x,y)$ and $\pi(x)$, $\pi(y)$ are the corresponding marginals $\pi_X(x)$ and $\pi_Y(y)$. This upper bound is the same as the kernel-free integral expression (in the limit of infinite data) for $||V_{YX}||^2_{HS}$ (Theorem 4 in Fukumizu et al. (2007)):

$$||V_{YX}||^2_{HS} = \int\int \left(\frac{\pi(x,y)}{\pi(x)\pi(y)} - 1 \right)^2 \pi(x)\pi(y)\, dx dy \tag{A.3}$$

The kernel dependence measure is thus equivalent to the χ^2 divergence, which is also well-known as *mean square contingency* (Hirschfeld, 1935). A similar relationship holds between the conditional cross-covariance operator $||V_{YX|Z}||^2_{HS}$ and the conditional mean square contingency.

Thus, our work and other algorithmic fairness works using MCC (Baharlouei et al., 2019; Mary et al., 2019) rely on the same ideal regulariser but employ a different empirical estimator, which is where the advantages of our proposed methods lie:

- Our measures deliver kernel estimators for the χ^2-divergence (with known consistency, see Theorem 5, Fukumizu et al. (2007)). However, it is straightforward to compute as only Gram matrix operations (trace, product) are involved,
- While Witsenhausen's characterisation of MCC is an easy approximation for discrete cases, it is not easily extended to continuous features and rely on strong assumptions (matrix Q is viewed as kernel of a linear operator on $L^2(d\pi_Y\pi_X)$, see Witsenhausen (1975)),
- There are two layers of difficulty in empirically computing the MCC for continuous variables. 1) It involves estimating the joint distribution and marginal distributions. Mary et al. (2019) employs KDE to estimate the distributions, which inherits the difficulty in choice of kernels. Additionally, direct estimation of PDF is infeasible if the joint space even has a moderate number of dimensions, restricting the scalability. 2) the marginalisation over Q is intractable for continuous values. To make this tractable, Mary et al. (2019) compute the estimation of the density on a regular square grid. This partitioning introduces its own set of problems, namely thresholding effects and that enough samples exist in each partition for low variance estimates.

To the best of our knowledge, no practical estimators of MCC exist yet for multi-dimensional, continuous variables. Instead, continuous variables are treated the same way as discrete ones (see another example of this Baharlouei et al. (2019)).

(Conditional) Mutual Information (Conditional) Mutual Information (CMI) is the best known dependence measure, but its finite sample empirical estimate is not straight-forward, especially for continuous variables. Formally, MI is defined:

$$MI(X,Y) = \int\int \pi(x,y) \log\left(\frac{\pi(x,y)}{\pi(x)\pi(y)}\right) dx dy \tag{A.4}$$

Compared to Equation A.3 and noting that $\log(z) \leq z - 1$, the inequality can be derived: $MI(X,Y) \leq ||V_{YX}||^2_{HS}$, which holds under the same assumptions as Lemma 3.2.1. Thus, Fair-COCCO is an upper bound on MI.

The main challenge with employing MI in practise is the difficulty in estimating it for continuous variables. This includes estimating densities as well as marginalising over the random variables. Due to these difficulties, there haven't been many works on employing CMI with continuous variables in the algorithmic fairness literature. To the best of our knowledge, Steinberg et al. (2020a) is the only work that employs CMI as a regulariser, but only works when the sensitive attributes are categorical and involves training a separate classifier in the inner loop.

Appendix B

Experiments using Demographic Parity

To highlight Fair-COCCO's compatibility with different fairness definitions other than EO, we apply it to demographic parity (DP). This requires statistical independence between predictions and attributes. *Disparate impact* (DI) is a metric frequently used to evaluate DP (Feldman et al., 2015):

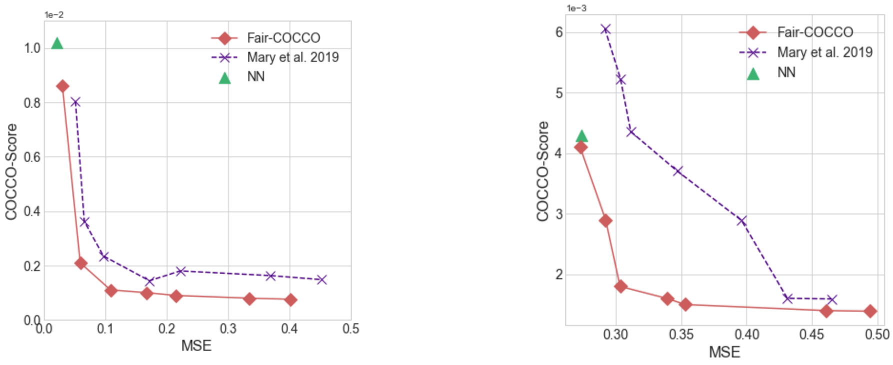
$$DI = \frac{P(\hat{Y} = 1 | A = 1)}{P(\hat{Y} = 1 | A = 0)} \tag{B.1}$$

where $A = 1$ and $A = 0$ denote respectively the discriminated and non-discriminated groups. The US Equal Employment Opportunity Commission Recommendation advocates that DI should not be below 80%, commonly known as the 80%-rule.[1]

We perform the same experiments on i) binary classification tasks (Table B.1), ii) regression task with multiple sensitive attributes (Table B.2), and iii) regression with a single sensitive attribute (Figure B.1).[2] The experiments are performed using the same procedures described in §4.2. Note, DI closer to 1 corresponds to lower levels of disparate impacts across population subgroups. We demonstrate that Fair-COCCO can balance strong predictive accuracy with fairness across multiple tasks and datasets, thus demonstrating the flexibility of our proposed method.

[1] www.uniformguidelines.com.

[2] None of the other methods in Table 5.3 have available code or results offor DP-regularised learning. We compare against results of Mary et al. (2019) re-run on our pre-processing pipeline.



(a)

(b)

Fig. B.1 Fairness-accuracy trade-off curve under DP for single sensitive attribute ‘racePct-Black’ on C&C (a) and ‘age’ on Students (b). Note, the optimum desiderata is at the origin, where both MSE and unfairness are minimised.

Table B.1 Accuracy (ACC) and DI under DP on benchmark of relevant algorithms on datasets. Results are aggregated over 10 runs for datasets without a designated test set. *NN* is an unregularised neural network and *Fair-COCCO* is the NN with our proposed regularisation. Note that the optimal DI value is 1.0.

	COMPAS		German		Drug		Adult	
Method	ACC	DI	ACC	DI	ACC	DI	ACC	DI
Mary et al. (2019)	0.87 ± 0.04	0.76 ± 0.07	0.71 ± 0.08	0.96 ± 0.25	0.80 ± 0.06	0.73 ± 0.17	0.79	0.83
NN	0.90 ± 0.00	0.39 ± 0.32	0.74 ± 0.07	1.26 ± 0.54	0.80 ± 0.08	0.42 ± 0.22	0.84	0.22
Fair-COCCO	0.88 ± 0.03	0.90 ± 0.06	0.73 ± 0.06	1.02 ± 0.19	0.78 ± 0.02	0.84 ± 0.07	0.83	0.97

Table B.2 MSE and Fair-COCCO score under DP when protecting multiple attributes. Results are aggregated over 10 runs and COCO scores are scaled by 1e−2.

	C&C		Student	
Method	MSE	COCCO	MSE	COCCO
Linear Regression	0.024	3.59	0.280	0.57
XGBoost	0.024	3.09	0.247	0.50
NN	0.023 ± 0.002	3.66 ± 0.04	0.224 ± 0.051	0.55 ± 0.03
Fair-COCCO	0.027 ± 0.007	1.36 ± 0.89	0.263 ± 0.060	0.31 ± 0.12

Appendix C

Description of Datasets

1. **Adult** (Kohavi, 1996). The task on the Adult dataset is to classify whether an individual's income exceeded $50K/year based on census data. There are 48842 training instances and 14 attributes, 4 of which are sensitive attributes (`age`, `race`, `sex`, `native-country`). Here, the sensitive attribute is chosen to be `sex`, which can be either female or male.

2. **Drug Consumption (Drugs)** (Mirkes, 2015). The classification problem is whether an individual consumed drugs based on personality traits. The dataset contains 1885 respondents and 12 personality measurements. Respondents are questioned on drug use on 18 drugs, including a fictitious drug `Semeron` to identify over-claimers. Here, we focus on `Heroin` use, drop the respondents who claimed to use `Semeron` and transform the categorical response into a binary outcome: "Never Used" versus "Used". The binary sensitive attribute is `Ethnicity`.

3. **South German Credit (German)** (Hoffman, 1994). The German dataset contains 1000 instances with 20 predictor variables of a debtor's financial history and demographic information, which are used to predict binary credit risk (i.e. complied with credit contract or not). The sensitive attribute is a binary variable indicating whether the debtor is of foreign nationality.

4. **COMPAS** (Angwin et al., 2016). COMPAS is a commercial software commonly used by judges and parole officers for scoring criminal defendant's likelihood of recidivism. The dataset contains 6172 instances with 10 features. The outcome is a binary variable corresponding to whether violent recidivism occurred (`is_violent_recid`) and the

sensitive attribute is `race`, which is binarised into “Caucasian” and “Non-Caucasian” defendants.

5. **Communities and Crime (C&C)** (Redmond, 2009). C&C contains socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey and crime data from 1995 FBI UCR. It contains 1994 instances of communities with 128 attributes. The outcome of the regression problem is crime rate within each community `ViolentCrimesPerPop`, which is a continuous value. There are three sensitive attributes, corresponding to ethnic proportions in the community—`racePctBlack`, `racePctWhite`, `racePctAsian`.

6. **Student Performance (Students)** (Cortez, 2014). The Students dataset predicts academic performance in the last year of high school. There are 649 instances with 33 attributes, including past academic information and student demographics. The response variable is a continuous variable corresponding to final grade and the sensitive attributes are `age` (continuous value from 15-22) and `sex` (‘F’-female, ‘M’-male).

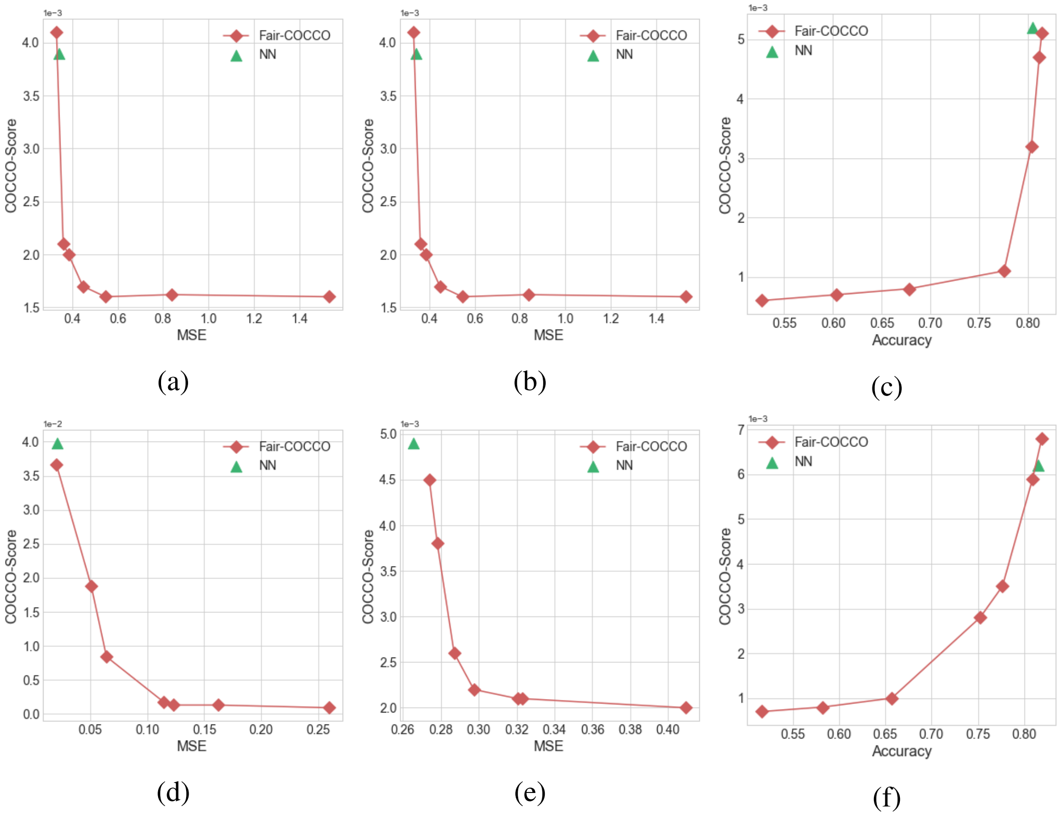
Appendix D

Accuracy-Fairness Trade-offs: Multiple Attributes

One of the key contributions of this study is the introduction of a differentiable fairness penalty that can naturally extend to multiple sensitive attributes. In this section, we generate the frontier of possible values on three experiments to better evaluate the accuracy-fairness trade-offs in different tasks:

1. Regression task on C&C with sensitive attributes: `racePctBlack`, `racePctAsian`, `racePctWhite`, and `racePctHisp`,
2. Regression task on Students with two sensitive attributes: `age` and `gender`,
3. Binary classification task on Drugs with three sensitive attributes: `age`, `gender`, and `ethnicity`.

We emphasise that, using our method, fairness w.r.t. multiple sensitive attributes can be tuned through a single penalty parameter whereas Mary et al. (2019), Pérez-Suay et al. (2017) will have to introduce multiple penalty terms, which presents optimisation challenges and additional hyperparameters. As Figure D.1 illustrates, fair outcomes can be achieved with minimal loss in MSE or accuracy.



(a) (b) (c)

(d) (e) (f)

Fig. D.1 Top row: trade-offs under EO. Bottom row: trade-offs under DP. (a), (d): C&C dataset with four sensitive attributes. (b), (e): students dataset with two sensitive attributes. (c), (f): drugs dataset with three sensitive attributes.

Appendix E

Problem Setup: Sepsis Treatment

The data used to develop and evaluate our experiment on fair imitation learning is extracted from the MIMIC-III ICU database (Johnson et al., 2016a), based on the Sepsis-3 cohort defined by Komorowski et al. (2018).

Discrimination in Healthcare Sepsis is one of the leading causes of mortality in intensive care units (Singer et al., 2016), and while efforts have been made to provide clinical guidelines for treatment, physicians at the bedside largely rely on experience, giving rise to possible variations in fair treatments. Prejudice in healthcare has been reported in many instances—for example, healthcare professionals are more likely to downplay women's health concerns (Rogers and Ballantyne, 2008) and racial biases affect pain assessment and treatment prescribed (Hoffman et al., 2016). Thus, it is critical, when learning to imitate expert policy, that no underlying prejudices are leaked into the learned policy.

Problem Setup We have access to a set of expert trajectories $\mathcal{D} = \{\tau_1, ..., \tau_N\}$, where each trajectory is a sequence of state-action pairs $\{(s_1, a_1), ..., (s_T, a_T)\}$. Each trajectories correspond to the state of a unique patient and the actions taken by the clinician to treat sepsis.

Data We obtain data from MIMIC-III and use the pre-processing scripts provided by Komorowski et al. (2018) to extract patients satisfying the Sepsis-3 criteria.[1] The description of the Sepsis-3 Cohort is provided in Table E.1.

[1] https://github.com/microsoft/mimic_sepsis.

Table E.1 Summary of MIMIC-III Sepsis-3 Cohort dataset. Trajectory count is based on the number of unique patients and trajectory lengths reports the average number of measurements (average time) spent in the ICU.

		Trajectory Count	**Trajectory Lengths**	**Total Measurements**
Total		19,585	13.06 ± 4.95	255,755
Gender	Male	10,901	13.10 ± 4.93	142,819
	Female	86,84	13.01 ± 4.98	112,936
Mortality	Survival	17,712	12.87 ± 4.88	228,023
	Death	1,873	14.81 ± 5.28	27,732

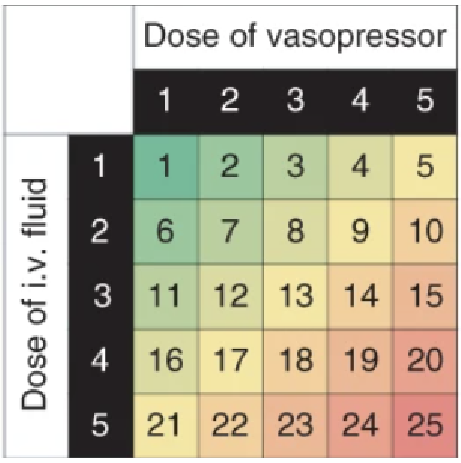
State Space The pre-processing yields 46 × 1 feature vectors for each patient at each timestep, which are summarised in Table E.2. Of the 46 features, there are 43 features corresponding to physiological parameters, lab values, vital signs and intake/output recordings. There are three sensitive attributes, namely gender, weight and age. In our experiments, we consider age as the sole sensitive attribute. Data are aggregated into 4 hour windows.

Table E.2 Description of patient features recorded at four hour intervals.

Feature Type	Features
Demographic	Gender, Age, Weight (kg),
Static	Re-admission, Glasgow Coma Scale (GCS), Sequential Organ Failure Assessment (SOFA), Systematic Inflammatory Response Syndrome (SIRS), Shock Index,
Lab Values	Potassium, Sodium, Chloride, Glucose, Magnesium, Calcium, White Blood Cell Count, Platelets Count, Bicarbonate, Hemoglobin, Partial Thromboplastin Time (PTT), Prothrombin Time (PT), Arterial pH, Arterial Blood Gas, Arterial Lactate, Blood Urea Nitrogen (BUN), Creatinine, Serum Glutamic-Oxaloacetic Transaminase (SGOT), Serum Glutamic-Pyruvic Transaminase (SGPT), Total Bilirubin, International Normalized Ratio (INR),
Vitals	Heart Rate, Systolic Blood Pressure, Mean Blood Pressure, Diastolic Blood Pressure, Respiratory Rate, Temperature (Celsius), FiO2, PaO2, PaCO2, PaO2/FiO2 ratio, SpO2,
Intake/Output	Mechanical Ventilation, Fluid Intake (4 hourly), Fluid Intake (Total), Fluid Output (4 hourly), Fluid Output (Total)

Action Space The actions taken by the clinician correspond to volume of intravenous (IV) fluids and maximum vasopressor (VP) dosage prescribed in a given 4 hour window. The action space is discretised into 25 values, with larger values corresponding to higher combined dosage. See Figure E.1 (provided in Komorowski et al. (2018)) for more on the action discretisation.

Treatment Outcome For our experiments, we consider two types of rewards: 1) defined on an action level and 2) defined on the trajectory level. Action level rewards are not readily available in the dataset and we leverage the estimation methods described by Raghu et al. (2017) to calcualte the true treatment outcome. The ground truth treatment outcome in each timestep is evaluated using SOFA (measuring organ failure) and the arterial lactate levels



		Dose of vasopressor				
		1	2	3	4	5
Dose of i.v. fluid	1	1	2	3	4	5
	2	6	7	8	9	10
	3	11	12	13	14	15
	4	16	17	18	19	20
	5	21	22	23	24	25

Fig. E.1 Discretisation of clinician actions. The dose of intravenous fluids and vasopressors were discretised into 25 possible actions. Graphic taken from Komorowski et al. (2018)

(higher in septic patients). Specifically, the treatment outcome penalises high SOFA scores and increases in SOFA and lactate levels from the previous timestep:

$$y_t = -0.025\mathbb{1}(s_{t+1}^{SOFA} = s_t^{SOFA} \cdot s_{t+1}^{SOFA} > 0) - 0.125(s_{t+1}^{SOFA} - s_T^{SOFA}) - 2\tanh(s_{t+1}^{lactate} - s_t^{lactate}) \tag{E.1}$$

At the trajectory level, true treatment outcomes are defined as the 90-day mortality after ICU admission, which are included in the dataset.