UNIVERSITY OF CALIFORNIA
SANTA CRUZ

MACHINE LEARNING AND THE MULTIAGENT ALIGNMENT PROBLEM

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE AND ENGINEERING

by

Reilly Raab

March 2024

The dissertation of Reilly Raab
is approved:

_____
Assistant Professor Yang Liu, Chair

_____
Professor Luca de Alfaro

_____
Professor Lise Getoor

_____
Professor Mingyan Liu

_____
Professor Emeritus Daniel Friedman

_____
Peter Biehl
Vice Provost and Dean of Graduate Studies

Contents

Abstract

Machine Learning and the Multiagent Alignment Problem

Reilly Raab

In the context of artificial intelligence (AI) or machine learning (ML), we speak of the "alignment" of an AI system's behavior with human goals, values, and ethical principles. "The alignment problem" has proven challenging, and as the capabilities and applications of AI rapidly advance, the shortcomings of standard solutions are increasingly consequential. This dissertation focuses on an often overlooked but critically important complication to the alignment problem: Socially-consequential AI systems affect their environment (involving, for example, human populations) and are therefore subject to dynamical feedback driven by other agents. We address three central questions:

(1) As intelligent agents adapt to each other, does a system aligned using current leading approaches remain aligned?
(2) Can we anticipate and utilize adaptive agents' reactions to data-driven policy to achieve aligned objectives dynamically?
(3) How can we guarantee alignment for AI systems that interact with complex, multi-agent environments that are difficult to model or predict?

We will address these questions using the theoretical framework and experimental tools of machine learning—integrating concepts from dynamical systems, evolutionary game theory, constrained optimization, and control theory. We hope to demonstrate that a dynamical systems approach to deployed AI is not only necessary but beneficial to the goal of alignment.

1 Introduction

1.1 The Rapid Development of AI

Over the last decade, machine learning (ML) techniques have enabled stunning technical advances in artificial intelligence (AI). With sufficient scale, models developed with ML have proven capable of accurately recreating (and therefore generating) rich, context-dependent examples of structured data such as text, images, and video (Vaswani et al., 2017; He et al., 2022; Bar-Tal et al., 2024). When trained in virtual environments, AI models have demonstrated super-human play in Atari, Go, and Chess without prior knowledge of game rules (Schrittwieser et al., 2020) and exhibited capabilities useful for autonomous driving and robotic control (Li et al., 2019; Amini et al., 2022; Hwangbo et al., 2019; Akkaya et al., 2019).

Such high-profile successes have driven significant, mutually reinforcing investments in ML that continue to advance state-of-the-art performance while encouraging new applications for AI. Novel neural network architectures, mature software ecosystems, optimizing compilers, highly parallel hardware, and large, high-quality data sets combine to lower the barriers to entry for further research and commercialization. In addition, recent evidence of so-called "scaling laws" (Kaplan et al., 2020; Wei et al., 2022; Hernandez et al., 2022; Chowdhery et al., 2023) have coincided with the construction of large-scale "foundation models", which are intended to provide Application Programming Interfaces for use on myriad uses of AI on downstream tasks (Bommasani et al., 2021).

At present, AI systems based on ML are used in a wide variety of socially-consequential tasks. Machine learning is currently used to predict loan default rates, calculate appropriate insurance premiums, anticipate the pricing of financial securities, automate

surveillance, and produce recommendations that affect individual access to credit, education, housing, employment, healthcare, and legal status (Hao, 2020; Metz and Satariano, 2020; Newton, 2021; Hernandez, 2021).

1.2 The Alignment Problem

When deployed in socially consequential tasks, decisions made with AI systems can have marked negative outcomes, not all of which are readily apparent to affected individuals. Recent public deployments of large language models have captured media attention due to their potential for nefarious use cases, such as the generation of misinformation, nonconsensual pornography, phishing attacks, or malicious code. In more benign situations, such models routinely "hallucinate" plausible but factually incorrect assertions.

Behind the scenes, however, AI models regularly exhibit bias, discrimination, and unauditable decisions, typically with little recourse or explanation. First, data is often biased and generated by social systems with legacies of inequity. Second, common models or the loss functions can induce inductive biases that encourage "shortcuts" or "simplifications", dismissing or ignoring minority groups. As a result, groups that are not well represented by data, exhibit more diverse behaviors, or are unfairly summarized by limited, unrepresentative features can be mistreated often without recourse.

This state of affairs has prompted action from world leaders, who have called for the alignment and regulation of AI and ML. In October 2023, the Biden Administration issued an executive order on the "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (Joseph R. Biden Jr., 2023). On the first day of 2024, the Vatican called for international efforts to regulate the use and development of AI to limit existing problems driving inequality and injustice, instead promoting efforts

focused on peace (Pope Francis, 2023).

Unfortunately, the policies generated by machine learning can be challenging to regulate and are not controllable to the same degree we expect from typical software programs. Beyond nascent research tools in "Explainable AI" (XAI) that target specific architectures, we cannot reliably understand, explain, interpret, or intercede on the black-box models that underpin AI decision-making. This gap between the desired control of AI outputs and what is currently possible is variously recognized as "the alignment problem", "the control problem", or "the safety problem".

1.3 Standard Interventions Fall Short

The standard approach to mitigating the misalignment of AI outputs is to attempt to express human goals as mathematical objectives that must be accounted for during model training (Christian, 2020). Sadly, it is generally difficult to craft optimization objectives that do this. First, humans regularly argue about ethical imperatives, and attempts to distill common objectives into mathematical objectives can be fraught and self-contradictory (Corbett-Davies and Goel, 2018). Second, a common observation in the training of AI models is that effective learning can benefit from "reward-shaping"—that is, an intentional misdirection of the model's objective away from our ultimate goals.

Even if we could precisely represent our current goals and objectives to the model, and even if such rewards were tractably learnable by AI methods, the approach of seeking alignment through singular training objectives faces an important complication in dynamical environments: In realistic settings, AI systems interact with complex, adaptive, multiagent systems—people, society, and other AI systems. Other agents typically have their own incentives, engaging in strategic actions that can, cumulatively,

shift the context and environment of the target system away from its training data. This *feedback* from other agents risks amplifying and exacerbating errors and misaligned behaviors.

When we account for dynamics and the multitude of human incentives that shape the use of AI systems, the realistic, immediate risk posed by AI is not a technological coup engineered by superintelligent agents. Instead, it is the amplification of latent incentives, distorted by machine interpretation, which threaten to destabilize human norms, institutions, and societal bonds. (Crawford and Calo, 2016; Chaney et al., 2018; Fuster et al., 2018; Ensign et al., 2018).

AI does not need to be integrated into all-encompassing surveillance systems, military drones, or global psyops campaigns to merit concern: AI provides a path to the realization of *homo economicus*—the hyper-rational, ethically unencumbered profit-maximizer.

1.4 Contributions of this Dissertation

This dissertation motivates the incorporation of multiagent dynamics into our understanding of the alignment problem and discusses technical approaches for making the resulting problem tractable.

Chapter 2 provides a brief background that reviews essential concepts and an overview of the field as represented by the research literature. Chapters 3 to 5 address each of the questions outlined in the abstract in turn. Finally, Chapter 6 draws conclusions from this cumulative body of work and its potential contributions toward informing our understanding of the alignment problem.

2 Preliminaries

2.1 Empirical, Black-Box Function Optimization

At its heart, machine learning (ML) is a set of techniques for finding a function that approximately minimizes a scalar objective. The function is a parametric function, denoted h_θ, which is described by a collection of numbers or *vector of parameters* θ in the same way that a hand saw may be described by a set of angles and dimensions that are expected to vary for different use cases (e.g., cutting steel rods, dry wood, tree branches, etc.). The objective, a functional of h_θ that quantifies how *bad* the current choice of θ is, is expressed using a "loss function" $\mathscr{L}$, which maps test cases z to a measure of badness (e.g., how long it takes to saw through a given object). The canonical goal is to minimize $\mathscr{L}$ in expectation over samples z.

$$\underset{\theta}{\text{minimize}} \quad \mathop{\mathrm{E}}_{Z}\left[\mathscr{L}(h_\theta, Z)\right]. \tag{2.1}$$

In general, h_θ can be any function that maps a set of inputs (e.g., observations, context) to a set of outputs (e.g., predictions, actions), and is sometimes called a *policy*. Meanwhile, the loss function $\mathscr{L}$ can be any measure of badness (such as the disagreement between the output of h_θ and desired values for a given example z) that ultimately depends on θ in a smooth (differentiable) manner.

There is a conceptual overlap between machine learning and evolution that we can exploit to develop intuitions for ML. Indeed, this correspondence can be made technically precise in a way that unifies their mathematics Appendix A. Nonetheless, there are also important practical differences that we aim to highlight in setting the stage for the exposition of this dissertation.

Table 2.1: Agent-specific variables forming a Markov chain.

Variable	Meaning	Domain	Realizations
G	group	$\mathcal{G} = \{1, 2, ..., n\}$	g, h, i, j
Y	qualification	$\{0, 1\}$ i.e., {unqualified, qualified}	y
X	feature	$(-\infty, \infty)$	x
$\hat{Y}$	classification	$\{0, 1\}$ i.e., {reject, accept}	$\hat{y}$
q_y	Probability density function of X given $Y = y$		

2.1.1 An Empirical Approximation

In Prob. (2.1), the "E" denotes an expectation value, indicating an average over random samples z. As we will soon see, the distribution $\mathcal{D}$ of examples z is critical to the policies trained by machine learning. In order to treat this distribution explicitly in the objective, we rewrite the canonical objective as

$$\underset{\theta}{\text{minimize}} \quad \mathcal{L}(\theta, \mathcal{D}), \tag{2.2}$$

where $\mathcal{D}$ represents a probability distribution for samples of z and

$$\mathcal{L}(\theta, \mathcal{D}) = \mathop{\mathrm{E}}_{Z}\left[\mathscr{L}(h_\theta, Z)\right] \coloneqq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}(h_\theta, z_i), \quad z_i \sim \mathcal{D}. \tag{2.3}$$

In this form, it is clear that the objective may be approximated empirically, using a finite number N of randomly drawn samples:

$$\mathcal{L}(\theta, \mathcal{D}) \approx \frac{1}{N} \sum_{i=1}^{N} \mathscr{L}(h_\theta, z_i), \quad z_i \sim \mathcal{D}. \tag{2.4}$$

Indeed, this is a key insight of standard machine learning techniques, where we use a large "dataset" of examples $\{z_1, z_2, ..., z_N\}$ on which $\mathscr{L}$ can be evaluated for any h_θ.

2.1.2 The Smooth Black-Box

To find a function that minimizes the desired objective, it is necessary to consider many functions. Formally, we desire a large, expressive *hypothesis class* $\mathcal{H}$ of possible functions $h_\theta \in \mathcal{H}$. To locate functions within a large hypothesis class requires many degrees of freedom, meaning that the vector space to which θ belongs is *high-dimensional*. In much the same way that life on Earth has settled on a common machinery built on the genetic base-pair sequences of nucleic acids to express a bewildering diversity of specialized organisms, applications of machine learning often recycle a set of "architectures" involving "artificial neural networks" to express large diversity of functions h_θ.

Like the mapping from DNA to organism traits, the mapping from parameters θ to h_θ is a "black-box": The correspondence between parameter values and the behavior of the function is difficult to interpret, explain, or intervene on through targeted modifications and engineering. Unlike the black-box used by evolution, however, the type of black-box useful for machine learning is *smooth*. By this, we mean that the values of θ are *continuous* (unlike discrete sequences of DNA codes), and we rely on $\mathscr{L}$ (thus $\mathcal{L}$) to vary *smoothly* (differentiably) with changes in θ, such that small changes to θ make small changes to h_θ. Thankfully, many familiar problems in engineering, optimization, and everyday experiences map to settings where ML applies.

As a familiar example of a problem involving a "smooth black-box", consider the need to manipulate a radio antenna in order to receive multiple television stations or channels. In this analogy, the antenna's position is described by a handful of angles comprising a parameter vector θ. These angles determine the directional gain of the antenna, represented by a function h_θ, that maps spacial directions and frequencies to an overall

level of amplification. The mapping from θ to h_θ is potentially complicated, ultimately governed by electrodynamics, but it *is* smooth. To complete the analogy, h_θ determines the ability of the antenna to resolve a given channel z. We may, therefore, quantify the antenna's *loss* of signal on channel z using a "loss function" $\mathscr{L}(h_\theta, z)$. ML may be used to find appropriate angles θ by repeatedly flicking through desired channels z, making small adjustments to θ that improve the reception in a process of *iterative refinement*.

2.1.3 Iterative Refinement

In the same way that you might adjust a radio antenna by making minor adjustments that improve the reception while repeatedly flicking between test channels, ML prescribes a procedure for updating θ in a series of small updates. These updates are, in general, solutions to local approximations of the original optimization problem (Prob. (2.2)):

$$\theta^{t+1} = \arg\min_{\theta} \quad \hat{\mathcal{L}}^t(\theta, \mathcal{D}) + \frac{1}{\eta} d^t(\theta, \theta^t). \tag{2.5}$$

In this equation, $\hat{\mathcal{L}}^t$ represents an approximation of $\mathcal{L}$ that is local to θ^t (i.e., has limited error for values of θ close to θ^t), and d^t is a metric that implies a non-negative "distance" between successive iterates of θ, thus penalizing large updates. In Eq. (2.5) and throughout our discussion, we adopt discrete-time semantics, where time takes on discrete values $t \in \{1, 2, ...\}$, and denote the value of quantities such as θ at time t with a superscript, as in θ^t.

There are many possible choices one can make regarding how to render the approximation $\hat{\mathcal{L}}^t$ or the metric d^t, but there is a canonical choice that uses a first-order Taylor

approximation and the Euclidean metric, respectively:

$$\hat{\mathcal{L}}(\theta, \mathcal{D}) = \mathcal{L}(\theta^t, \mathcal{D}) + \left\langle \theta - \theta^t, \nabla_\theta \mathcal{L}(\theta, \mathcal{D}) \Big|_{\theta=\theta^t} \right\rangle. \tag{2.6}$$

$$d(\theta, \theta^t) = \frac{1}{2}\|\theta - \theta^t\|_2^2 = \frac{1}{2}(\theta - \theta^t)^2. \tag{2.7}$$

These choices render a closed-form solution to Eq. (2.5) known as *gradient descent*:

$$\theta^{t+1} = \theta^t - \eta \nabla_\theta \mathcal{L}(\theta, \mathcal{D}) \Big|_{\theta=\theta^t} \tag{2.8}$$

When we combine these choices with the empirical approximation of Eq. (2.4), we get *stochastic gradient descent*, the workhorse algorithm of machine learning, which may be practically implemented on computers using an algorithm known as backpropagation.

$$\theta^{t+1} = \theta^t - \eta \nabla_\theta \frac{1}{N} \sum_{i=1}^{N} \mathscr{L}(h_\theta, z_i) \Big|_{\theta=\theta^t}, \quad z_i \sim \mathcal{D}. \tag{2.9}$$

In the above equation, η is known as a "learning rate" and N is known as the "batch size" used for each update.

With the development of expressive, smooth, black-box ML architectures (such as neural networks) and the backpropagation algorithm for efficiently calculating gradients (as used in Eq. (2.9)) machine learning has found widespread use for optimization tasks that rely on extrapolating from an existing dataset $\{z_1, z_2, ...\}$ on which $\mathscr{L}$ can be evaluated for any h_θ.

2.1.4 Example: Binary Classification

A classic example of how ML techniques can be applied to practical tasks is provided by *binary classification*. Imagine using ML to learn a function h_θ that maps a vector of an

individual's *features* x (such as age, income, education, net assets, etc.), to a continuous value $h_\theta(x) \in (-\infty, \infty)$ from which we derive a binary decision $\hat{y}_\theta(x) \in \{-1, 1\}$ (such as whether the person is qualified for a loan).

$$\hat{y}_\theta(x) \coloneqq \begin{cases} -1 & h_\theta(x) \leq 0 \\ 1 & \text{otherwise} \end{cases}. \tag{2.10}$$

Given a set of values $z = (x, y)$ consisting of pairs of feature vectors and known "ground truth" or "labels" $y \in \{-1, 1\}$ (e.g., a bank's records of previous loan applications and whether the applicant was truly qualified), we would like to use ML to train a policy that will extrapolate from these examples by minimizing the "zero-one" loss, which penalizes disagreement between $\hat{y}_\theta(x)$ and y:

$$\mathscr{L}_{0/1}(h_\theta, z) = \begin{cases} 1 & \hat{y}_\theta(x) \neq y \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1 & -yh_\theta(x) > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{2.11}$$

A standard result in ML is that $\mathscr{L}_{0/1}$ can be minimized indirectly if we instead minimize a "margin-based loss function" such as *square-loss*

$$\mathscr{L}(h_\theta, z) = -yh_\theta(x) + \frac{1}{2}h_\theta^2(x), \tag{2.12}$$

or the variously-named (log, logistic, cross-entropy, etc.) information-theoretic loss,

$$\begin{aligned} \mathscr{L}(h_\theta, z) &= \begin{cases} -\log\left[p_\theta(x)\right] & y = 1 \\ -\log\left[1 - p_\theta(x)\right] & y = -1 \end{cases}, \quad p_\theta(x) \coloneqq \frac{1}{1 + e^{-h_\theta(x)}} \\ &= -yh_\theta(x) + 2\log\left[e^{h_\theta(x)/2} + e^{-h_\theta(x)/2}\right]. \end{aligned} \tag{2.13}$$

By minimizing the error rate of h_θ, and therefore $\hat{y}_\theta(x)$, over our historical data set, we can use the learned policy on *new* examples in order to automate, assist, or otherwise accelerate loan the approval for the bank... or so the thinking goes...

2.2 Present Normative Interventions

The pipeline for machine learning is built around an optimization problem for a given distribution of examples. Adopting this mindset, what do you do when a policy you've optimized using machine learning ends up being biased against a gender, a disability status, or a racial category?

Such issues are not hypothetical: This is exactly what happened in the Florida criminal system (Angwin et al., 2020; Metz and Satariano, 2020), with Google Images (Hern, 2018; Grant and Hill, 2023; Grant, 2024), in Amazon's hiring (Dastin, 2018), and in myriad other examples that negatively affect some groups of people more than others.

More consequentially, what do you do when your ML policy breaks the law by resulting in "disparate impact" when conditioned on race, sex, gender, religion, etc.? First, you retract the deployed model (quietly or with a public apology, depending on the flavor of hot water you're in). Next, you research how the issue occurred and task your engineers to fix the issue as quickly as possible. Where possible, you fund research that might help you avoid the same problem in the future—or at least certify that you're using "best-practices".

Research will tell you that there are several mechanisms by which such problems can occur in the first place. That is, there are many ways that ML can generate biased and discriminatory outcomes; The data can be biased; The hypothesis class can be biased; The objective can be biased.

Unfortunately, the black-box architectures that make ML so widely applicable also render the resulting model difficult to interpret, explain, or correct. Without effective tools for interpreting and interceding on AI decision-making, what can you do?

Where data augmentation and output filtering do not suffice, leading approaches to the alignment problem intervene on the *objective itself*. That is, the proposed solution is to constrain the policy to respect some (potentially implicit) threshold of undesirable or "alignment-violating" behavior quantified by a function $\mathcal{V}$:

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \mathcal{L}(\theta, \mathcal{D}) \\ \text{subject to} \quad & \mathcal{V}(\theta, \mathcal{D}) \leq \varepsilon. \end{aligned} \tag{2.14}$$

The "slack variable" ε is introduced in this equation to allows some parametric violation of the fairness condition while allowing a straight-forward unification with "regularized" approaches to fairness (See Appendix B).

What should $\mathcal{V}$ look like?

2.2.1 Group Fairness

The need to constrain ML models to respect quantitative boundaries aligned with intuitive notions of fairness has spawned the field of "algorithmic fairness". The basic idea has been to formalize different *fairness measures*, often in terms of statistical (in)consistencies across groups (Dwork et al., 2012; Zemel et al., 2013; Hardt et al., 2016b; Zafar et al., 2017a; Chouldechova, 2017; Feldman et al., 2015; Kleinberg et al., 2016) or between similar individuals (Dwork et al., 2012; Zemel et al., 2013), so-called "preference guarantees" (Zafar et al., 2017b; Ustun et al., 2019a; Kusner et al., 2017), or causal considerations Kusner et al. (2017); Kasy and Abebe (2021).

While such measures are not always mutually compatible (Corbett-Davies and Goel, 2018), such definitions have become standard in the research literature and are frequently treated as proxies de jure, representing baselines that may or may not be used in future legal defense of suspect ML policies. We briefly outline a few fairness measures that we will repeatedly reference throughout this dissertation.

The most salient aspect of group fairness is the use of pre-defined, fixed "groups" to which people belong. While the standard boundaries between groups in the research literature are fixed by demographic race, sex, country of origin, etc.., this dissertation acknowledges efforts to define meaningful groups in terms of social network clusters or shared adaptive behaviors. We will attempt to avoid considering groups purely in terms of narrow demographic categories.

Some fairness definitions classically consider only two groups (even going so far as to pre-specify them as "advantaged" and "disadvantaged"). To remain as agnostic to setting as possible, we will consider countably many groups $g \in \mathcal{G} = \{1, 2, ..., n\}$, though we will frequently run experiments with only two groups, in part because these are the easiest to visualize in two dimensions and thus interpret. Likewise, many fairness definitions are also specialized to the binary classification setting, which involves using ML to make a binary prediction $y \in \{-1, 1\}$ for an individual belonging to group g with features x. Parts of our subsequent exposition will adopt this setting for simplicity, but many results generalize to regression tasks, etc.

Unless otherwise stated, we will adopt a convention that defines the *violation* of fairness, also called "disparity", as the group-wise variance of a some (possibly vector-valued)

Table 2.2: Random Variables for Binary Classification

Variable	Meaning	Domain	Realizations
G	group	$\mathcal{G} = \{1, 2, ..., n\}$	g
Y	qualification	$\{-1, 1\}$ i.e., {unqualified, qualified}	y
X	feature vector	$\mathbf{R}^d$	x

quantity ξ. That is, we define

$$\begin{aligned}\mathcal{V}(\theta, \mathcal{D}) = \operatorname*{Var}_g \left[\xi_g(\theta, \mathcal{D})\right] &\coloneqq \frac{1}{n}\sum_{g=1}^{n} \left\|\xi_g(\theta, \mathcal{D}) - \bar{\xi}(\theta, \mathcal{D})\right\|^2 \\ &= \left(\frac{1}{n}\sum_{g=1}^{n} \left\|\xi_g(\theta, \mathcal{D})\right\|^2\right) - \left(\left\|\bar{\xi}(\theta, \mathcal{D})\right\|^2\right),\end{aligned} \tag{2.15}$$

where ξ measures different quantities in different contexts, and we introduce the bar-notation $\bar{(\cdot)}$ to denote an average, as in

$$\bar{\xi}(\theta, \mathcal{D}) \coloneqq \frac{1}{n}\sum_{g=1}^{n} \xi_g(\theta, \mathcal{D}). \tag{2.16}$$

As another convention, throughout this dissertation, we use the capitalized letters $Z = (X, Y, \hat{Y}, G)$ to represent the properties of a randomly drawn sample $Z \sim \mathcal{D}$

Loss Parity is perhaps the simplest fairness definition when one assumes that the loss incurred by a policy is *incentive-compatible*, i.e., equally undesirable to affected individuals as to the policy-designer. In this case, we treat ξ_g and group-wise loss as synonymous:

$$\xi_g(\theta, \mathcal{D}) \coloneqq \operatorname*{E}_{Z\sim\mathcal{D}}\Big[\mathscr{L}(h_\theta, Z) \mid G{=}g\Big]. \tag{2.17}$$

Demographic Parity, also known as "statistical parity" (Dwork et al., 2012; Zemel et al., 2013; Feldman et al., 2015), is perhaps the most frequently considered fairness intervention. For binary classification tasks, it prescribes an equal positive predictive rate

(e.g. loan-approval rate) across groups. We can measure the *violation* of demographic parity in terms of Eq. (2.15) by specifying

$$\xi_g(\theta, \mathcal{D}) \coloneqq \Pr_{X,G\sim\mathcal{D}}\Big(\hat{y}_\theta(X) = 1 \;\Big|\; G = g\Big). \tag{2.18}$$

Equal Opportunity is much like demographic parity but requires equal positive classification rates for *qualified* (i.e., y=1) individuals across groups.

$$\xi_g(\theta, \mathcal{D}) \coloneqq \Pr_{X,Y,G\sim\mathcal{D}}\Big(\hat{y}_\theta(X) = 1 \;\Big|\; Y = 1, G = g\Big). \tag{2.19}$$

Note that equal opportunity, which requires access to ground truth labels Y, may not always be measurable in practice.

Equalized Odds, considered by Hardt et al. (2016b); Zafar et al. (2017a); Chouldechova (2017), like equal opportunity, requires that a classifier has equally accurate classification rates across groups for qualified individuals, but also demands the same for *unqualified* individuals. Note, in this case, that we treat ξ_g as a vector quantity indexed by $y \in \{-1, 1\}$.

$$\xi_{g,y}(\theta, \mathcal{D}) \coloneqq \Pr_{X,Y,G\sim\mathcal{D}}\Big(\hat{y}_\theta(X) = 1 \;\Big|\; Y = y, G = g\Big). \tag{2.20}$$

2.2.2 Ethics Inferred from Examples

Rather than attempting to mathematically define fairness, a recent line of work has proposed using machine learning to infer whether *individual* policy outputs are aligned, using labeled examples. That is, reconsidering standard ML setting, if we are given a dataset of examples $\widetilde{z} = (\widetilde{x}, \widetilde{y})$ where $\widetilde{x}$ represents a policy behavior (i.e., an input-

output pair (x, y)) and $\widetilde{y}$ labels whether this behavior demonstrates *alignment*, we can train a policy v_ϕ, with parameters ϕ, to accurately recognize whether individual policy outputs are "aligned". Using a standard classification loss function $\mathscr{L}$, this idea may be represented as

$$\mathcal{V}(\theta, \mathcal{D}) = \mathop{\mathrm{E}}_{z\sim\mathcal{D}}[v_{\phi^\star}(h_\theta, z)]. \tag{2.21}$$

$$\phi^\star = \arg\min_{\phi} \mathop{\mathrm{E}}_{\widetilde{z}\sim\widetilde{\mathcal{D}}}[\mathscr{L}(v_\phi, \widetilde{z})]. \tag{2.22}$$

This approach is currently a leading proposal for regulating the outputs of so-called "large language models" (LLMs), where the range of undesirable behaviors is far broader than can be accounted for by simple statistics as in the case of binary classification tasks. In particular, recent work has explored using "weak" LLMs to evaluate and flag problematic outputs of "larger" LLMs in order to automate this process and accelerate the work of humans who label examples.

2.2.3 The Tacit Assumption

Whether derived from formal group-fairness definitions or learned from labeled examples of (un)ethical behavior, the function $\mathcal{V}$ is frequently used as in Prob. (2.14) to regulate a policy h_θ with the tacit assumption that $\mathcal{D}$ will remain *fixed*. We term this class of interventions "present normative" insofar as they prescribe a norm or standard that is specific to a present context or environment, represented by $\mathcal{D}$.

This tacit assumption is not unique to issues of fairness but is baked into the standard optimization problem at the heart of machine learning. In reality, especially when ML is deployed for consequential decision-making, $\mathcal{D}$ is liable to *change*.

2.3 Past Performance is no Guarantee of Future Results

The standard machine learning paradigm assumes that a single distribution $\mathcal{D}$ represents both the set of *training examples* and the set of examples that are actually encountered during *deployment*. In practice, the distribution of examples we use to train a policy may differ from the distribution of examples when we use the policy. Adopting standard nomenclature, throughout this paper, we will make use of the symbol $\mathcal{S}$ to represent the "source" distribution on which an ML policy is trained. We will use the symbol $\mathcal{T}$ to represent the "target" distribution on which the ML policy is actually deployed.

In general, the source and target distribution are distinct: $\mathcal{S} \neq \mathcal{T}$. In the research literature, this phenomenon is known as *distribution shift*, and strategies to reduce these problems are known as *transfer learning* or *domain adaptation*: Such literature often focuses on the fact that the source-trained policy may be suboptimal on the target distribution. i.e.,

$$\theta^\star = \arg\min_\theta \quad \mathcal{L}(\theta, \mathcal{S}) \implies \theta^\star = \arg\min_\theta \quad \mathcal{L}(\theta, \mathcal{T}). \tag{2.23}$$

This behavior can have ramifications for misaligned behavior in dynamical environments, as explored in Section 3.4. For present normative interventions, however, it is also worth highlighting that the source-trained policy may violate the alignment constraint:

$$\begin{aligned} \theta^\star = \arg\min_\theta \quad & \mathcal{L}(\theta, \mathcal{S}) \\ \text{subject to} \quad & \mathcal{V}(\theta, \mathcal{S}) \leq \varepsilon. \end{aligned} \implies \mathcal{V}(\theta^\star, \mathcal{T}) \leq \varepsilon. \tag{2.24}$$

There are many causes of distribution shift in practice. The most commonly considered is the error resulting from the empirical approximation of (Eq. (2.4)) used for training.

That is, even when $\mathcal{S} = \mathcal{T}$, the *true* source distribution used for training is a finite approximation of the intended distribution. Another source of distribution shift is caused by exogenous changes or events that occur in the environment: For example, if we were to train an ML model to predict individuals' creditworthiness from a set of features represented in loan applications, overall changes in the distribution of applicants might be affected by interest rates, GDP growth, employment trends, inter-state immigration, etc.. As a result, applications that would likely be approved in one set of circumstances (e.g., on which our model was trained) may need to be denied in different settings to limit credit risk while respecting alignment constraints, but our model could be oblivious to this shift.

In this dissertation, we are primarily concerned with distribution shift caused by the *policy itself*.

2.3.1 Feedback and Strategic Response

This dissertation considers problems of alignment when a machine learning policy participates in a multiagent system that *adapts* to its actions, thereby shifting the distribution of examples it faces *interactively*. Rather than occupying a theoretical niche, such a situation is fundamental to the real-world issues we face as a society with the advent of AI.

Let us consider a simple example of how agents can react to policy, thereby shifting the environment. Imagine a population of strategic individuals who submit applications for schools, jobs, or loans to be evaluated by an ML model: Such individuals have potent incentives to embellish certain parts of their applications while underplaying others. The may try to establish (or avoid) purchasing habits that correlate with positive or negative classification. The may also try to establish features that serve primarily to

enhance their application, such as extracurricular activities, certifications, or awards. Finally, they might also decide to abstain from future application altogether. Each of these courses of action can shift the distribution of examples assumed by the ML model, violating any alignment constraints (Hardt et al., 2016a; Ustun et al., 2019b; Zhang et al., 2020).

From this motivating example, it is clear that the use of AI for consequential decision-making subjects it to the complex system of dynamics and feedback that characterize human society. At stake are the risks of runaway dynamics (e.g., "the rich get richer") that threaten the stability of societal institutions and demography, the dynamics of power, class divisions, and meritocratic norms. It is perhaps less clear what those consequences may be in different situations, nor what can be done about them.

In some ways, the problem is not new to AI: *human* intelligence and agency faces the same considerations. Nonetheless, the formulation of intelligent behavior as described in the framework of machine learning, when married to perspectives of control theory, dynamical systems, and game theory, provides us with powerful tools to systematically address the dynamical, multiagent alignment problem. This dissertation thus builds on a nascent body of work that addresses the alignment of AI while accounting for such dynamics.

2.4 Related Work

This dissertation contributes to a growing body of literature that addresses the alignment problem in contexts with dynamical feedback. In this section, we outline prior and contemporary work that has considered the intersection of fairness and distribution shift, dynamics caused by machine learning policies, or the need for online policies that adapt to unanticipated dynamics.

Language in this section has been adapted from previously published work (Raab and Liu, 2021; Chen et al., 2022; Yin et al., 2023; Raab et al., 2024).

2.4.1 Fairness Subject to Distribution Shift

Prior literature has explored issues of algorithmic fairness subject to distribution shift. In particular, one line of work has considered how distribution shift (without also considering reactionary policy updates) can degrade or counter-act typical present normative interventions (Liu et al., 2018; Hu et al., 2019; Hu and Zhang, 2022).

Complementary to the issue of how distribution shift can affect fairness guarantees, several recent studies have considered specific examples of fairness *transferability* subject to distribution shift (Schumann et al., 2019; Coston et al., 2019; Singh et al., 2021; Rezaei et al., 2021; Kang et al., 2022). That is, the task at hand is to prove that an "aligned" ML model, trained distribution $\mathcal{S}$ will be aligned on another distribution $\mathcal{T}$, within some bound.

Within this problem domain, Schumann et al. (2019) examine *equality of opportunity* and *equalized odds* as definitions of group fairness subject to distribution shifts quantified by an H-divergence function. Coston et al. (2019) consider *demographic parity* subject to a *covariate shift* assumption while group identification remains unavailable to the classifier. Singh et al. (2021) focus on common group fairness definitions for binary classifiers subject to a class of distribution shift that generalizes covariate shift and label shift by preserving some conditional probability between variables; and Rezaei et al. (2021) similarly consider common binary classification fairness definitions such as equalized odds subject to covariate shift.

Related to the task of certifying fairness with anticipated distribution shift is the setting in which a target distribution is known or can be sampled from. As an example, a model

might assume covariate shift and the availability of some unlabeled target data (Coston et al., 2019; Singh et al., 2021; Rezaei et al., 2021). In this task, Singh et al. (2021) focus on learning stable models that will preserve prediction accuracy and fairness, utilizing a causal graph to describe anticipated distribution shifts. Rezaei et al. (2021) takes a robust optimization approach, and Coston et al. (2019) develops prevalence-constrained and target-fair method for building a model robust to covariate shift.

2.4.2 Modelled Dynamics of Fairness Interventions

As highlighted by D'Amour et al. (2020), analysis of fairness subject to distribution shift cannot ignore mutual feedback and sustained interaction and hope to capture essential dynamics. In order to model a dynamical, multiagent system in sustained interaction with an ML policy, we require good models. As in any domain, models must always compromise between realism, tractability, generality, and usefulness.

Efforts to model population responses to algorithmic policy and the autonomous dynamical systems that result from myopically updating policies have been performed by Coate and Loury (1993); D'Amour et al. (2020); Zhang et al. (2020); Heidari et al. (2019); Wen et al. (2019); Liu et al. (2020); Hu and Chen (2018); Mouzannar et al. (2019); Williams and Kolter (2019); Perdomo et al. (2020); Hu and Zhang (2022); Zhang et al. (2019); Dean et al. (2022); Hashimoto et al. (2018). Mouzannar et al. (2019) address stateful dynamical transitions of the type that we consider in Raab and Liu (2021) and address in Sections 3.5, 4.2.3, and 5.4. Perdomo et al. (2020); Hu and Zhang (2022) consider stateless dynamical transitions of the type we consider in Raab et al. (2024) and address in Sections 3.4 and 4.1. We organize different classes of models based on their relationship to state in Section 3.2.

Most work in this domain has highlighted the failure modes or suitability of present

normative interventions (Section 2.2). Hashimoto et al. (2018) also considers distributionally robust policies and Morik et al. (2020) identify interventions for myopic optimization by adopting a feedback control mechanism specific to raking on online matching platforms.

2.4.3 Safe Reinforcement and Online Learning

Beyond case-by-case treatments of specific dynamical models, more recent work has considered ML models that adapt to *unknown* dynamical environments. In Chapter 5, we argue that such settings fundamentally require an *online* approach, which includes some reinforcement learning (RL) methods. The key difficulty is in attempting to give guarantees regarding alignment or constraints.

In *model-based* approaches, the algorithm learns an explicit dynamical model of the environment (Efroni et al., 2020; Singh et al., 2020; Brantley et al., 2020; Zheng and Ratliff, 2020; Kalagarla et al., 2021; Liu et al., 2021; Ding et al., 2021a). In *model-free* approaches, the algorithm learns a policy that is implicitly refined according to the dynamics, but the dynamical transitions are not explicitly modelled. In practice, model-free algorithms often require the use of a *simulator* that allows exploration over arbitrary state-action pairs (Xu et al., 2021; Ding et al., 2020; Bai et al., 2022), which is unrealistic for real-world unknown dynamics.

In general, it is difficult to establish safety guarantees for online methods without severe assumptions on the dynamics or practicality of the method. In Yin et al. (2023), we adapt online learning methods to provide probabilistic bounds on cumulative regret and disparity.

3 Alignment is not Static

As intelligent agents adapt to each other, does a system aligned by present normative intervention remain aligned?

As alluded to in Sections 1.3 and 2.3.1, this question is of great importance in determining whether current approaches to AI alignment are sufficient for mitigating the societal risks.

In general, the answer to the stated question is "it depends": It depends on the incentives of other agents in the system, the dynamics of the environment, and how much agents' actions affect each other. This being said, there are clear examples of multiagent dynamics that we can model for in which the answer is "no". Given this negative answer, we will develop a general framework we can use to quantify how bad the misalignment can be.

We will address approaches to alignment in dynamical contexts in Chapter 4. In this chapter, we discuss different models for adaptive dynamics and their ramifications for our motivating question.

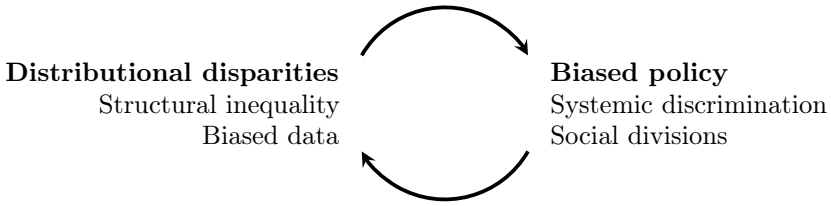

Figure 3.1: The mutual interactions of ML policy with its environment can lead to pernicious feedback loops.

3.1 Mechanisms for Distribution Shift

There are many ways that distributions can change, and modelling how populations of humans react to change is a difficult task. A common approach to the challenge of summarizing distribution shift in a way that is theoretically tractable is to imagine that the distribution of examples $\mathcal{D}$ may be summarized by a few key variables, and to propose feasible mechanisms by which these variables can change.

We now introduce a few different (but non-exhaustive) ways of summarizing a distribution $\mathcal{D}$ of individuals, belonging to different groups $g \in \{1, 2, ..., n\}$, who are subject to binary predictions by an ML classifier. For intuition, we will again consider an example setting in which ML is used for a binary classification task, where Z denotes a randomly selected applicant, X represents his or her vector of features, Y represents his or her true (unknown) label or "qualification", and G represents his or her group.

For a binary classification setting (Section 2.1.4), the distribution $\mathcal{D}$ corresponds to a joint probability distribution over the joint variables $Z = X, Y, G$. We will allow ourselves to use the notation

$$\mathcal{D} \equiv \Pr(X, Y, G). \tag{3.1}$$

There are many ways that this probability distribution can be decomposed.

3.1.1 Label Shift

One way to decompose the distribution in Eq. (3.1) is as

$$\Pr(X{=}x, Y{=}y, G{=}g) = \underbrace{\Pr(G{=}g)}_{\text{Assume constant}} \underbrace{\Pr(Y{=}y \mid G{=}g)}_{\text{Model}} \underbrace{\Pr\{X{=}x \mid Y{=}y, G{=}g\}}_{\text{Assume constant}} . \tag{3.2}$$

As indicated by the braces under the factors that comprise the expansion on the right hand side of Eq. (3.2), *label shift* assumes that the type of distribution shift encountered in a given problem can be well-approximated by a shifting conditional probability (density) function $\Pr(Y{=}y \mid G{=}g)$, which can be determined according to some model. $\Pr(G{=}g)$ and $\Pr(X{=}x \mid Y{=}y, G{=}g)$, meanwhile, are assumed to be static.

As an example of what the label shift assumption entails, imagine that our binary classification task was to predict whether an individual has contracted SARS-CoV-2 ($Y{=}1$) using blood measurements X. Suppose we have robust models for how specific protein densities, blood cell counts, etc., statistically depend on Y. If only population infections rates were to change over time, we would expect, absent virus mutation, the statistical effect that Y has on X for a randomly selected individual (i.e., $\Pr(X \mid Y, G)$ to remain fixed. In such a situation, the overall, joint distribution of X, Y, and G would be well-modelled by the label-shift assumption.

Meaningful values for possible groups G in this setting might include types of occupation, such as health-care workers, in-person staff, and remote workers: These groups would likely have different rates of infection. Geographical groupings or distinctions based on risks to Covid infection, such as whether someone is immunocompromised, are also appropriate as motivating examples.

We will introduce a detailed example of Eq. (3.2) in Section 3.5.1.

3.1.2 Covariate Shift

Another way to decompose the distribution in Eq. (3.1) is as

$$\Pr(X{=}x, Y{=}y, G{=}g) = \underbrace{\Pr(G{=}g)}_{\text{Assume constant}} \underbrace{\Pr(X{=}x \mid G{=}g)}_{\text{Model}} \underbrace{\Pr\{Y{=}Y \mid X{=}x, G{=}g\}}_{\text{Assume constant}}. \tag{3.3}$$

As in label shift, only a single factor is assumed to be wholly responsible for the type of distribution shift assumed by *covariate shift*: In this case, the conditional probability (density) function $\Pr(X{=}y \mid G{=}g)$ is modelled as a variable while $\Pr(G{=}g)$ and $\Pr(Y{=}x \mid X{=}y, G{=}g)$ are assumed to be static.

Consider again the example of classifying SARS-CoV-2 infection Y using data X obtained from blood. Now imagine that measurement values X have the same predictive relationship to SARS-CoV-2 infection Y (i.e. ($\Pr(Y \mid X)$ remains constant), but policy changes have shifted such that you no longer test *asymptomatic* individuals, which by and large occupy a certain segment of the feature space (possible X values). The distribution $\Pr(X)$ has now shifted in isolation, as assumed by the covariate shift assumption.

3.1.3 Participation Rates

A third way to decompose the distribution in Eq. (3.1) is as

$$\Pr(X{=}x, Y{=}y, G{=}g) = \underbrace{\Pr(G{=}g)}_{\text{Model}} \underbrace{\Pr(Y{=}y, X{=}x \mid G{=}g)}_{\text{Assume constant}}. \tag{3.4}$$

In this case, the conditional probability (density) function $\Pr(Y{=}y, X{=}y \mid G{=}g)$ is assumed to remain constant while $\Pr(G{=}g)$ is updated according to model.

An example of shifting participation rates is easy to imagine when users may elect whether to "participate" in the ML classification we wish to consider. For example, supposed we grouped users by geographic region and occupation— variables that can strongly correlate with political affiliation or attitudes towards public health mandates. If certain groups of users were to suddenly boycott Covid-19 testing, independent of feature values or Covid-19 status, the distribution shift might be well-modelled by changing group participation rates alone.

We considered a model based on Eq. (3.4) in Section 3.4.1.

We next consider how these simplified representations of distributions can be modelled in time.

3.2 Models of Dynamics

We consider two categories of policy-induced distribution shift driven by multiagent dynamics: The first is a tractable simplification of policy-induced distribution shift in which the distribution is a fixed function of policy and certain consequence of policies within the available hypothesis class $\mathcal{H}$ are known a priori. This first model is useful for demonstrating the hazards of ignoring such dynamics. The second makes use of evolutionary game theory to model a large population of affected agents that respond to an ML policy and considers the mean-field population dynamics of such responses. This second model allows a richer class of dynamical transitions while still remaining tractable enough to be useful and avoiding common problems faced by prior literature.

As introduced in Section 2.1.3, we continue to use discrete-time semantics, wherein

the parameters θ^t and distribution $\mathcal{D}^t$) evolve in time $t \in \{1, 2, ...\}$ through repeated interactions. To disambiguate the sequence of updates, however, we specify that θ^{t+1} responds to $\mathcal{D}^t$, and that $\mathcal{D}^t$ responds to θ^t. While this is perhaps an unnatural time-indexing when focusing on how $\mathcal{D}$ evolves in response to θ, our later discussion will be more concerned with how θ adapts to $\mathcal{D}$.

$$... \mapsto \theta^t \mapsto \mathcal{D}^t \mapsto \theta^{t+1} \mapsto \mathcal{D}^{t+1} \mapsto ... \tag{3.5}$$

Given this ordering, to use the language of Section 2.3, $\mathcal{D}^t$ is the "source" distribution for θ^{t+1}, while each θ^{t+1} induces the "target" distribution $\mathcal{D}^{t+1}$. When θ and $\mathcal{D}$ alternately update in this fashion, we describe an autonomous dynamical system that evolves in time.

3.2.1 Distributions as Function of Policy

Our first model is built on a simplistic assumption of how ML policy affects its environment: For discrete-time updates, this assumption is the existence of an (a priori *unknown*) function S for which

$$\theta^t \mapsto \mathcal{D}^t, \quad \text{e.g.,} \quad \mathcal{D}^t = \mathrm{S}(\theta^t). \tag{3.6}$$

Note that, because the ML policy h_θ is parametric, we express the induced distribution as a function of the parameters θ, for simplicity.

This dynamical assumption has come to be recognized under the name of "performative prediction" in prior work that has considered the effects of distribution shift on loss (Perdomo et al., 2020) and in multiagent settings (Narang et al., 2022).

Policy-Induced Participation Rates

In general, Eq. (3.6) may be combined with any summary representation of the shifting distribution, such as those listed in Section 3.1. As an example, consider how participation rate (Section 3.1.3) might be directly induced by an ML-trained policy: Suppose every individual who might interact with an ML maintains a personal threshold for model performance above which they will become an active user of the ML model, and that members of a given group $g \in \{1, 2, ..., n\}$ all observe the same model performance. As a consequence, user participation rates will monotonically increase with the performance of the ML model on that group. When performance maps to negative loss, we assume a fixed function φ_g for each group g such that

$$\rho_g^t = \varphi_g(\ell_g^t) \quad \text{and} \quad \frac{\mathrm{d}\varphi_g}{\mathrm{d}\ell_g} < 0. \tag{3.7}$$

To make this example more concrete, imagine the ML model as a recommendation service (for products, videos, internet search results, etc.) run by a corporation or firm. To develop intuition, we may group users based on shared interests and preferences (e.g., their favorite film genre). We may imagine how the firm will lose users from group g as the quality of recommendations for that group deteriorates.

We use the model of Eq. (3.7) in Sections 3.4 and 4.1.

3.2.2 State Dependence

A major limitation of Eq. (3.6) is that the induced target distribution is assumed to be a function of the deployed policy alone. Given our prior discussion about the myriad causes of distribution shift (Section 2.3), this may not be realistic.

An alternative is to model $\mathcal{D}^t$ as an aspect of *state* s^t, which depends on θ^{t-1} as well as

its prior value:

$$\begin{pmatrix} \theta^t \\ \mathbf{s}^{t-1} \end{pmatrix} \mapsto \mathbf{s}^t \mapsto \mathcal{D}^t, \quad \text{e.g.,} \quad \begin{aligned} \mathcal{D}^t &= \mathrm{M}(\mathbf{s}^t) \\ \mathbf{s}^t &= \mathrm{U}(\theta^t, \mathbf{s}^{t-1}). \end{aligned} \tag{3.8}$$

In this framing s represents the environment or *state* (which entails $\mathcal{D}$), while θ may be interpreted as an *action*.

In general, many contributing factors may comprise state relevant to the dynamics, including prior history and exogenous variables. The dependence of $\mathcal{D}$ on state may also allow simplifications, however, in which the evolving state is identified with a specific shifting summary of the distribution, as discussed in Section 3.1.

Replicating Qualifications

Many possible simplifications inherit from specific mechanisms of distribution shift (Section 3.1). In such cases, the evolving state may admit a simpler (i.e., lower-dimensional) representation than $\mathcal{D}^t$ itself.

For a concrete example, consider the case of label shift and let us represent the group "qualification rate" as:

$$s_g \coloneqq \Pr(Y{=}1 \mid G{=}g). \tag{3.9}$$

We may propose models for how s_g evolves in time that are founded on evolutionary game theory. In particular, if Q_g represents the utility of qualification ($Y{=}1$) to an individual in group g and N_g represents the utility of non-qualification ($Y{=}0$), then we may model

$$s_g^t \coloneqq s_g^{t-1} \frac{Q_g(\theta^t)}{s_g^{t-1} Q_g(\theta^t) + (1 - s_g^{t-1}) N_g(\theta^t)}. \tag{3.10}$$

Interpreting this equation, when $Q_g(\theta^t) > N_g(\theta^t)$, then $Q_g(\theta^t) > (s_g^{t-1}Q_g(\theta^t) + (1 - s_g^{t-1})N_g(\theta^t))$, and the qualification rate s_g modelled by Eq. (3.10) will increase by a multiplicative factor. When $N_g > Q_g$, s_g will decrease. Intuitively, $Q_g(\theta^t)$ and $N_g(\theta^t)$, as the expected utility of qualification or non-qualification respectively, are mapped by Eq. (3.10) to *replication rates* for two competing straties (whether or not to seek qualification). These utilities thus determine the probability that a strategy (become (un)qualified) will spread, in a viral manner, within each group g. For a discussion of the underlying model for Eq. (3.10), inspired by evolutionary game theory, please refer to Appendix A.

This model is notable for plausibly explaining persistent disparities under group-independent prediction policies—i.e., those that do not discriminate on the basis of group membership—without assuming a setting that is structurally imbalanced between groups (Raab and Liu, 2021).

In particular, by assuming that utilities associated with each possible outcome $(Y, \hat{Y})$ are universal across groups, such a model is able to endogenizes disparities by appealing to unequal *initial conditions* rather than by assuming inherent differences between groups of people, as done by previous work.

We use the model of Eq. (3.10) in Sections 3.5, 4.2.3, and 5.4.

3.2.3 Markov Transitions

A generalization of Eq. (3.8) is to allow *stochastic transitions*, thus defining Markovian dynamics .

$$\begin{pmatrix} \theta^t \\ \mathbf{s}^{t-1} \end{pmatrix} \mapsto \Pr\big(\mathbf{s}^t\big), \quad \text{e.g.,} \quad \mathbf{s}^t \sim \mathbf{P}(\theta^t, \mathbf{s}^{t-1}). \tag{3.11}$$

Such a framing is the standard setting for *reinforcement learning*, in which an AI policy must select the optimal value of θ in a given state to maximize some long-term reward, possibly by learning to explicitly model the stochastic transitions indicated by Eq. (3.11).

Linear Markov Transitions

A further refinement of Eq. (3.11) are so-called "linear" Markov dynamics, in which $\mathbf{P}$ is *linear* in θ and $\mathbf{s}$ when both quantities are represented in appropriately chosen coordinates.

We represent the Linear MDP assumption with respect to a "feature map" ϕ: To be technically precise, the feature map represents the *reproducing kernel Hilbert space* (RHKS) in which the system dynamics are linear. The formal assumption then is that there exists, for each state s' an appropriate vector $\mu[s']$ such that, for all s and θ,

$$\Pr(s' \mid s, \theta) = \big\langle \phi(s, \theta), \mu[s'] \big\rangle \tag{3.12}$$

While this may seem like a significant restriction, it is well-founded when $\mathbf{s}^t$ belongs to the infinite space of all possible observables of a system, as supported by Koopman operator theory (Brunton et al., 2021) Unfortunately, we seldom have the ability to represent infinitely many dimensions, or even to know the best finite-dimensional representation of $\mathbf{s}^t$ in practice.

3.3 Baseline ML Polices

As discussed in Section 2.2, the standard paradigm in machine learning is an optimization problem that assumes a static distribution. Where necessary to address issues of algorithmic alignment, it is standard to impose a constraint on the policy informed by

the current distribution, as in

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \mathcal{L}(\theta, \mathcal{D}) \\ \text{subject to} \quad & \mathcal{V}(\theta, \mathcal{D}) \leq \varepsilon. \end{aligned} \tag{3.13}$$

In other words, the baseline response to issues with machine learning are to intervene on the objective, given that the distribution of examples — the data that makes ML work — is difficult to change.

Given this standard approach to algorithmic alignment, we now focus on how this approach fares when the distribution evolves in time.

We first consider the case in which Eq. (3.13) fully prescribes the response of θ^{t+1} to $\mathcal{D}^t$, a method known as "Repeated Risk Minimization" (RRM). Next, we consider the case in which the process of iterative refinement discussed in Section 2.1.3 is run concurrently with updates to $\mathcal{D}$. Finally, we outline a common alteration to Eq. (3.13) that seeks to encode a notion of *robustness* to distribution shift.

3.3.1 Repeated Risk Minimization

"Repeated risk minimization" (RRM) names an approach to dealing with distribution shift that, essentially, ignores it:

$$\begin{aligned} \theta^{t+1} = \arg\min_{\theta} \quad & \mathcal{L}(\theta, \mathcal{D}^t) \\ \text{subject to} \quad & \mathcal{V}(\theta, \mathcal{D}^t) \leq \varepsilon. \end{aligned} \tag{3.14}$$

Intuitively, RRM updates θ^t in response to $\mathcal{D}^t$, at each time t, by choosing the myopically optimal policy without regard to how $\mathcal{D}$ may change in time. This method as been considered by Perdomo et al. (2020), who show that the method, in practice, can

achieve small performance gaps relative to optimal models subject to the assumption of Section 3.2.1. As we will see, this method can also go horribly wrong, even subject to similar assumptions: We consider RRM as a baseline policy in Sections 3.5.2 and 5.4.2.

3.3.2 Repeated Gradient Descent

As discussed in Section 2.1.3, problems in ML such as Eq. (3.13) are addressed in practice by iteratively refining an approximate solution using a sequence of constrained optimization subproblems that each approximate the true objective. When the distribution $\mathcal{D}$ evolves in time, a simple adaptation this approach is to iterate a single step of this sequence of subproblems to iteratively refine θ, anticipating that $\mathcal{D}$ may likewise shift between iterates. Such a technique is known as "Repeated Gradient Descent" (RGD) (Perdomo et al., 2020).

$$\begin{aligned} \theta^{t+1} = \arg\min_{\theta} \quad & \mathcal{L}(\theta, \mathcal{D}^t) + \frac{1}{2\eta}(\theta - \theta^t)^2 \\ \text{subject to} \quad & \mathcal{V}(\theta, \mathcal{D}^t) \leq 0. \end{aligned} \tag{3.15}$$

It should be noted that, frequently, the only algorithmic distinction between Eq. (3.14), implemented via iterative refinement, and Eq. (3.15) is the number of updates to θ that are made prior to considering how $\mathcal{D}$ may have changed in the interim.

RGD is also considered by Perdomo et al. (2020), who note qualitatively similar behavior to the performance of RRM on the problems they considered. We consider RGD as a baseline policy in Sections 3.4.2 and 4.1.3.

3.3.3 Distributional Robustness

A final baseline that we may consider, which straddles the boundary between normative present fairness and policies that account for the

possibility of error between the source distribution $\mathcal{S} = \mathcal{D}^t$ and the target distribution $\mathcal{T} = \mathcal{D}^{t+1}$. Given that such distribution-shift can occur even without dynamics, simply as a result of approximation errors from empirical sampling, there are a few method that are intended to be "distributionally robust".

While formulated in terms of adversarial Stackelberg games in some approaches, we map this to the notation of constrained optimization represented by a primal Lagrangian problem: Essentially, we imagine that we must choose a policy to optimize for the worst-case scenario that may occur as a result of distribution-shift. We specify that the target distribution $\mathcal{D}$ lies within a bounded distance of the nominal distribution $\mathcal{D}^t$, as measured by d.

$$\begin{aligned} \theta^{t+1} = \arg\min_{\theta} \Bigg[\max_{\mathcal{D}} \quad & \mathcal{L}(\theta, \mathcal{D}) \Bigg] \\ \text{subject to} \quad & \mathcal{V}(\theta, \mathcal{D}) \leq \varepsilon, \\ & d(\mathcal{D}, \mathcal{D}^t) \leq \delta. \end{aligned} \tag{3.16}$$

Often, the framework of Eq. (3.16) requires simplifications about *how* distributions may shift, as in Section 3.1, in order to be tractable. We will discuss such assumptions within the context of adversarial distribution shifts more fully in Section 3.6.

Modifications of this basic idea may also be applied to the constraint itself, as in group-fairness. That is, rather than defining $\mathcal{V}$ a measure of diverging treatments between groups (Eq. (2.15)), one can reformulate fairness as

$$\mathcal{V}(\theta, \mathcal{D}) = \max_{g} \Big[\xi_g(\theta, \mathcal{D}) \Big] \tag{3.17}$$

in the case that ξ_g is *undesirable*, as is the case for incentive-compatible loss.

An example of this form of present normative intervention is provided by Hashimoto et al. (2018) and considered as an additional baseline by Yin et al. (2023) for the work

represented in Section 5.4.

3.4 The Failure of Myopia

Parts of this section are adapted from previously published work by Raab et al. (2024).

Even when present normative interventions do not restrict the behavior of ML policies, the phenomenon of mutual feedback between dynamical environments and *ex post* adaptive policy responses can exhibit misaligned behavior that drives undesirable long-term outcomes.

In this section, we provide an intuitive demonstration of this sort of misalignment. First, we discuss a setting involving policy-induced participation rates, as in (Section 3.2.1) relevant to recommendation systems. Second, we interpret how a common baseline, repeated gradient descent assuming a fixed distribution, can result in *increasing* loss in this system, which we can visualized geometrically.

3.4.1 A Recommendation System Example

An online service run by a *firm* that algorithmically recommends user-generated content (e.g., fixed-length videos) to other users based on inferred preferences. Let us assume that users will only engage with the service if a sufficiently high percentage of recommended videos are personally interesting. As outlined in Section 3.2.1, we say that the proportion of prospective users in each group that will interact with the service, denoted ρ_g, monotonically decreases as a function of the group-specific loss ℓ_g, where

Assumption 3.4.1 (Participation Decreases with Loss)**.**

$$\ell_g^t \coloneqq \mathop{\mathrm{E}}_{Z\sim\mathcal{D}^t}\Big[\mathscr{L}(h_\theta, Z) \mid G{=}g\Big]; \tag{3.18}$$

$$\rho_g^{t+1} = \varphi_g(\ell_g^t) \quad \text{where} \quad \frac{\mathrm{d}\varphi_g}{\mathrm{d}\ell_g} < 0. \tag{3.19}$$

The firm seeks to maximize user engagement (e.g., the total daily number of videos watched on the platform) in order to drive advertising revenue. We represent this objective as

$$\underset{\boldsymbol{\ell}\in\mathcal{A}}{\text{minimize}} \quad \langle \boldsymbol{\ell}, \varphi(\boldsymbol{\ell})\rangle, \tag{3.20}$$

where we introduce the group-indexed vectors

$$\text{Group-specific losses} \quad \boldsymbol{\ell} \coloneqq (\ell_1, \ell_2, ..., \ell_n), \tag{3.21}$$

$$\text{Group-specific participation rates} \quad \boldsymbol{\rho} \coloneqq (\rho_1, \rho_2, ..., \rho_n), \tag{3.22}$$

and where the decision variable $\boldsymbol{\ell} \in \mathcal{A}$ is a consequence of an assumption that any policy parameterized by θ, consistently maps to the same set of achievable group-wise losses $\mathcal{A}$.

Assumption 3.4.2 (Static Set of Feasible Losses)**.**

$$\forall t, \quad \boldsymbol{\ell}^t \in \mathcal{A}. \tag{3.23}$$

This simplification would apply, for example, when a user's choice of whether to engage with the service, in group g, is statistically independent of other group-specific preferences relevant to the recommendation service.

θ	parameter value.
$\mathcal{D}$	distribution of users.
$\mathcal{L}$	the objective function (total loss).
g	discrete group index.
ℓ_g	average loss for group g.
ρ_g	participation rate for group g.
$\mathcal{A}$	set of achievable losses.
φ_g	map from ℓ_g to ρ_g.

Table 3.1: Choice of notation

For this problem, repeated gradient descent (RGD; Section 3.3.2) is given by

$$\boldsymbol{\ell}^{t+1} = \arg\min_{\boldsymbol{\ell} \in \mathcal{A}} \ \mathcal{L} \coloneqq \langle \boldsymbol{\rho}^t, \boldsymbol{\ell} \rangle. \tag{3.24}$$

We consider the set of achievable losses $\mathcal{A}$ to be convex. The convexity of this set is well justified by the ability of the firm to adopt mixed policies; that is, for any two loss vectors $a, b \in \mathcal{A}$, we assume that the firm is free to deploy a stochastic mixture of the policies that resulted in a and b, implying that $\mathcal{A}$ is closed under convex combinations.

Without loss of generality, we fix zero loss for each group to correspond to zero participation:

$$\boldsymbol{\ell} \preceq 0 \ \text{ and } \ \varphi(0) = 0. \tag{3.25}$$

We do this to restrict interpretations of Eq. (3.24) to situations in which the firm has incentives to realize high participation rates (as opposed to eliminating users that are universally costly).

3.4.2 A Geometric Picture of Misalignment

The primary utility of Section 3.4.1 is that it offers a literal, geometric picture of present-normative misalignment in a dynamical context: The vector of participation rates $\boldsymbol{\rho}$

may be interpreted as a *dual vector* to $\boldsymbol{\ell}$, such that the overall objective in Eq. (3.24) is the dot-product "alignment" between the two: In Fig. 3.2, $\boldsymbol{\ell}$ and $\boldsymbol{\rho}$ should ideally point in opposite directions as far as possible.

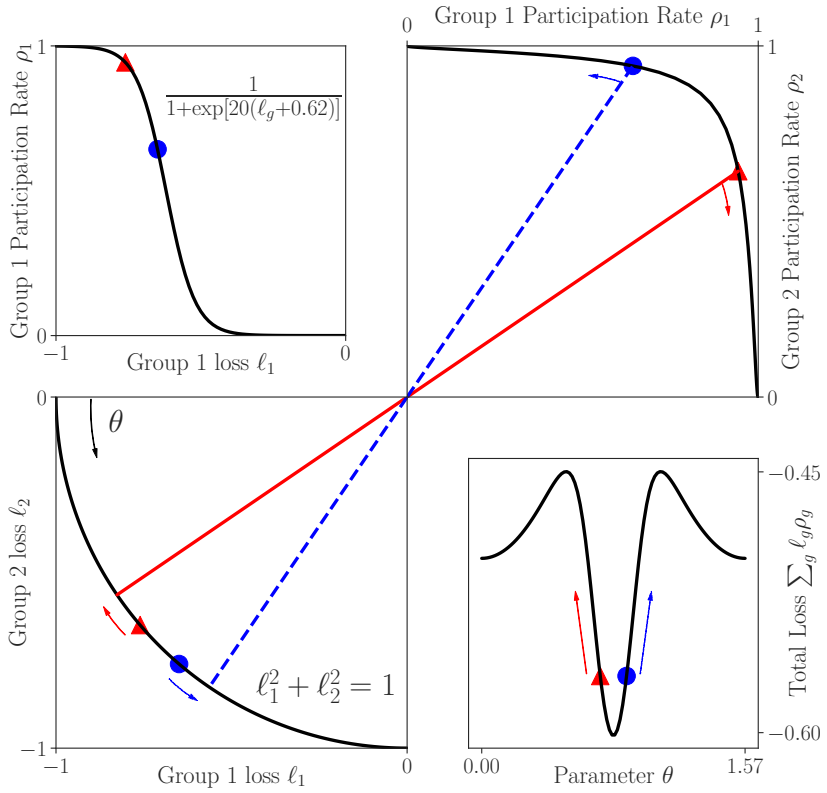

Figure 3.2: Failures of myopic optimization. In this figure, we consider two policies, represented by the red triangle and the blue circle, which correspond to combinations of achievable group-specific losses $\boldsymbol{\ell}$. In this example, the set $\mathcal{A}$ is defined by a quadrant of the unit disk (lower left). These group-losses induce corresponding group participation rates (upper right) via $\rho_1 = \varphi(\ell_1)$ (upper left). Because total loss has the form of an inner product, we may interpret $\boldsymbol{\rho}$ as vector in the dual space of $\boldsymbol{\ell}$, where we wish to increase the relative alignment of corresponding $\boldsymbol{\rho}$ and $\boldsymbol{\ell}$ vectors.

In Fig. 3.2, we see that there exist alternative values of $\boldsymbol{\ell}$ which would reduce loss *if* $\boldsymbol{\rho}$ *were fixed* (i.e., under the assumption made by RGD). Choosing these alternative values, however, causes $\boldsymbol{\rho}$ to *shift* in reality, which again induces $\boldsymbol{\ell}$ to change in a direction that continues the feedback cycle. The true form of the performative loss (lower right of

the figure) when the relationship of $\boldsymbol{\rho}$ on ℓ is accounted for indicates that the policies selected by RGD will actually increase loss over time.

The root cause of this misalignment is modifying ℓ without considering how this changes ℓ. If the policy is repeatedly optimized for the engagement of the current user base, as prescribed by present-normative paradigm, the resulting dynamics may enter a feedback-cycle that exclusively targets over-represented users (Fig. 3.2), increasing overall loss.

We will revisit this setting in Section 4.1, where we suggests ways of accounting for the induced distribution shift and compare our suggestions to the myopic baseline represented above. First, we will show in Section 3.5 that present-normative constraints are also subject to dynamical misalignment.

3.5 The Potential Harm of Present-Normative Interventions

Parts of this section are adapted from previously published work by Raab and Liu (2021).

As shown in Section 3.4, myopic solutions to inherently dynamical problems can lead to misaligned dynamics, even without additional constraints to address bias. It is perhaps unsurprising, therefore, that the incorporation of present-normative interventions (which do not account for induced dynamics; Section 2.2) in the machine learning pipeline does not necessarily prevent undesired outcomes: In some cases, such interventions can exacerbate the problems they seek to address.

3.5.1 A Setting for Unintended Selection

In this section, we provide a setting in which present-normative interventions can backfire, driving widening distributional inequalities in society. Our setting, published

in Raab and Liu (2021), is closely related to that of Coate and Loury (1993) but treats n groups and a more granular classifier utility function. To ground our intuition for this example, we imagine a binary classification task (Section 2.1.4) in which an ML-trained policy must approve or deny loan applications. We picture this setting in Fig. 3.3. Each individual in the population possesses a group membership $G \in \{1, 2, ...\}$, a feature profile X, and a binary label Y that he or she may *choose* at each time t. Given the existence of this choice, we will frequently refer to individuals in the population as *agents*.

Groups: We will model groups as isolated subpopulations that differ only in size and initial proportions of qualified individuals. That is, we are careful to attribute no inherent disparities between groups to underlying, inherent differences: We will model all agents identically, across groups. In each group, individuals compare qualifications and resulting classifier outcomes, then *choose* whether it is in their best interest to invest in becoming qualified or unqualified in the future.

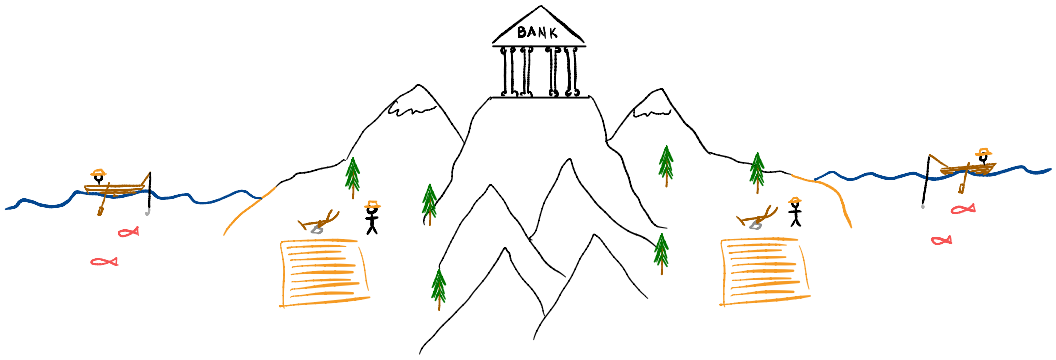

Figure 3.3: A graphical representation of a setting in which an ML policy is used by a lending institution to directly approve or deny loan applications. In this setting, the population is divided into isolated groups, identical other than initial conditions, that compare strategies for qualification and resulting outcomes among themselves.

Table 3.2: Agent-specific variables in this model form a Markov chain.

Variable	Meaning	Domain	Realizations
G	group	$\mathcal{G} = \{1, 2, ..., n\}$	g, h, i, j
Y	qualification	$\{-1, 1\}$ i.e., {unqualified, qualified}	y
X	feature	$(-\infty, \infty)$	x
$\hat{Y}$	classification	$\{-1, 1\}$ i.e., {reject, accept}	$\hat{y}$

Features: In this example, we think of each agent's feature profile X as a scalar-valued summary of attributes reported to the ML classifier in a loan application, such as age, education, assets, etc., which we interpret as an overall *credit score*. Because we wish to avoid assuming inherent differences between groups, we consider the label-conditioned feature distribution $\Pr(X \mid Y = y)$ as group-independent. That is, an (un)qualified individual in one group looks statistically the same as an (un)qualified individual in any other group (groups only differ, statistically, in initial qualification rates). Let us define the fixed, differentiable probability density functions

$$q_y(x) \coloneqq \Pr(x \mid Y{=}y); \quad y \in \{-1, 1\}. \tag{3.26}$$

Assumption 3.5.1. For tractability of our model, we assume that $q_y(x)$ is nowhere zero and that the ratio $q_1(x)/q_{-1}(x)$ is strictly increasing in x. That is,

$$\forall x, y,\ q_y(x) \in (0, \infty); \quad \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{q_1(x)}{q_{-1}(x)}\right) > 0 \tag{3.27}$$

This assumption ensures that the feature X is statistically "well-behaved": As credit scores increase as x, the odds that individuals with that credit score x will pay off the prospective loan also increases.

Predictions: Having assumed that q_1/q_{-1} is monotonically increasing in x (Asm. 3.5.1), we model the ML-trained model as a *threshold* classifier, which will approve loans above a certain threshold credit-score and deny applicants below the threshold.

$$h_\theta(x, g) \coloneqq \begin{cases} 1 & x \geq \theta_g \\ -1 & x < \theta_g \end{cases} \tag{3.28}$$

To point out a failure mode of myopic optimization, we assume that θ^{t+1} is chosen to maximize profit on the *currently observed distribution* $\mathcal{D}^t = \Pr_t(Y, X, G)$ as an instance of repeated risk minimization (RRM; Section 3.3.1), subject to imposed group-fairness conditions (Section 2.2.1). We model the profit of decision $\hat{y}$ conditioned on true qualification y as a constant entry in a matrix $C_{y,\hat{y}}$, such that the classifier is incentivized to make the correct prediction ($C_{y=\hat{y}} > C_{y\neq\hat{y}}$). In summary, we specify

$$\theta^{t+1} = \arg\min_\theta \quad -\sum_g \sum_{y,\hat{y}=\{-1,1\}} C_{y,\hat{y}} \Pr_t(Y{=}y, h(x, g){=}\hat{y} \mid X{=}x, G = g) \tag{3.29}$$

$$\text{subject to} \quad \mathcal{V}(\theta, \mathcal{D}^t) \leq 0, \tag{3.30}$$

for various possible constrained disparity measures $\mathcal{V}$.

Table 3.3: Variables of state.

Variable	Meaning
s_g	Qualification rate for group g
θ_g	Feature classification threshold for g

Evolving Labels: The binary label $Y \in \{1, 1\}$ of an agent indicates his or her qualification for the loan, which is unknown to the classifier (and must be predicted from X and G). Intuitively, a qualified ($Y = 1$) individual will repay a loan in full

if accepted ($\hat{Y} = 1$). We interpret qualification as a desirable attribute at a societal level (one which public policy would promote if it could), and note that different group qualification rates are the sole cause of disparity in this model.

Importantly, in this model, individuals are free to *choose* their qualification at each round t. Anticipating algorithmic classification, how should agents decide whether to become qualified? We imagine that there is a cost associated with the ability to pay back loans, and that the rationality of doing so depends on what an individual knows about the classification policy they will be subjected to (and potential outcomes, which they estimate from incomplete information provided by peer examples).

To deal with uncertainties that agent may possess, we model this decision for updating personal qualification by the mechanism of *imitating* the strategies of others based on popularity and observed utility. For example, if your friend chose to become qualified for a loan and now runs a small business, the net profit of the transaction may induce you to seek qualification yourself by first building credit history. If instead many of your neighbors receive loans despite being unqualified and manage to live lavishly, you may infer that qualification is a waste of resources.

Björnerstedt and Weibull (1994) have shown that *imitation* in this form, whereby agents stochastically update to strategies weighted by utility and popularity yields the replicator equation (Section 3.2.2 and Appendix A), which we use in its discrete time form, as detailed by Friedman and Sinervo (2016). We therefore use the replicator equation to model the replication or spread of successful strategies (and the abandonment are unsuccessful strategies). We thus consider binary labels $Y \in \{-1, 1\}$ as competing evolutionary *strategies* in the population, with fitness determined by the average decisions (and error rates) of the classifier on the group.

Table 3.4: Variables representing notions of utility.

Variable	Meaning	Indices
C	utility to classifier	$y, \hat{y}$
A	utility to agent	$y, \hat{y}$
Q	utility of qualification to agent	g
N	utility of non-qualification to agent	g

Let us represent the *qualification rate* in group g as s_g:

$$s_g^t \coloneqq \Pr_t\Big(Y = 1 \;\Big|\; G = g\Big). \tag{3.31}$$

The replicator equation (Eq. (3.10)) in this context yields

$$s_g^t = \frac{s_g^{t-1} Q_g(\theta^t)}{s_g^{t-1} Q_g(\theta^t) + (1 - s_g^{t-1}) N_g(\theta^t, t)} \tag{3.32}$$

where

$$Q_g(\theta) = \sum_{\hat{y} \in \{-1,1\}} A_{1,\hat{y}} \Pr_t(h_\theta(X, g) = \hat{y} | Y = 1) \tag{3.33}$$

$$N_g(\theta) = \sum_{\hat{y} \in \{-1,1\}} A_{-1,\hat{y}} \Pr_t(h_\theta(X, g) = \hat{y} | Y = -1) \tag{3.34}$$

This model assumes that the fitness of qualification, Q_g, and the fitness of non-qualification, N_g, are dependent on the probability that the classifier will correctly classify the individual in group g. This probability is influenced by group membership, the current policy θ^t, and the inherent desirability (net cost) associated with each possible outcome, represented by the matrix $A_{y,\hat{y}}$ for $y, \hat{y} \in \{-1, 1\}$.

Assumption 3.5.2. We assume that $A_{-1,1} \neq A_{-1,-1}$ and $A_{1,1} > A_{1,-1}$. That is, agents care about the decision of the classifier, and qualified agents prefer to get the loan.

Model Commentary: Our choice to model distribution shift using replicator dynamics in qualification rates provides several advantages. First, it is free from structurally asymmetric assumptions (e.g., assuming inherent, immutable advantages or disadvantages of certain groups) that are common in related research literature; Second, it is capable of explaining persistent disparities under Bayes-optimal, group-independent policies. Third, it derives from plausible, localized information exchange between individuals.

This modelling choice also restricts our interpretation of groups, which we model as functionally impermeable to the exchange of qualification strategies: i.e., they must be closed and insular. Given that "sensitive attributes" such as race, sex, color, etc. may not correspond to meaningful divisions between people (which depend on social context), this modelling assumption gives us a functional working definition: groups are defined by the extent to which they satisfy this assumption of mutual independence with regard to the exchange of information and strategies.

3.5.2 Simulation Results

In this section, we consider the setting outlined in Section 3.5.1 subject to demographic parity, equalized odds (Section 2.2), and the absence of constraints (which we term "laissez-faire"). We present a subset of simulations presented in Raab and Liu (2021).

In Fig. 3.4, we model two groups of equal size subject to the classifier utilities $C_{y,\hat{y}}$ and agent utilities $A_{y,\hat{y}}$ given by

$$\begin{bmatrix} A_{-1,-1} = 0.1 & A_{-1,1} = 5.5 \\ A_{1,-1} = 0.5 & A_{1,1} = 1.0 \end{bmatrix} \qquad \begin{bmatrix} C_{-1,-1} = 0.5 & C_{-1,1} = -0.5 \\ C_{1,-1} = -0.25 & C_{1,1} = 1.0 \end{bmatrix} \tag{3.35}$$

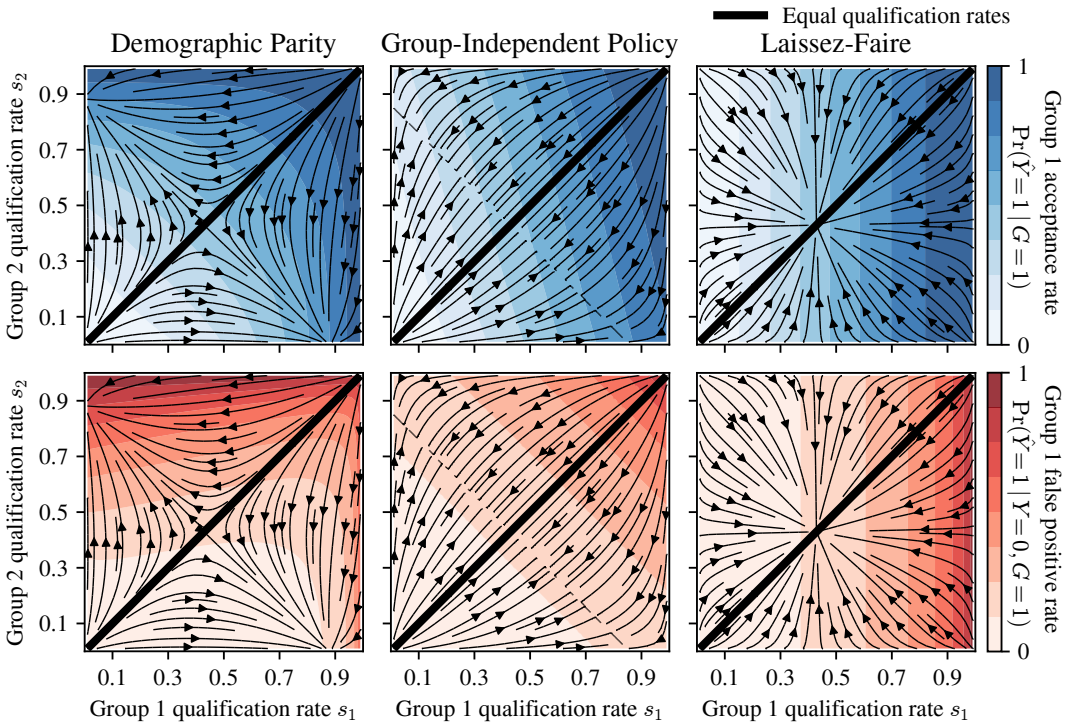

Figure 3.4: A "phase diagram" of the dynamics of the model presented in Section 3.5.1. The state of the system is provided by the qualification rates of two equal-sized groups (i.e., s_1 and s_2), represented by the two axes of each pane. At each time step, the classifier performs repeated risk minimization (RRM; Section 3.3.1; Eq. (3.29)) subject to demographic parity or equalized odds (Section 2.2), or no intervention at all, which we refer to as "laissez-faire". The population then responds, as modelled by Eq. (3.32), by updating qualification rates. Through the mutual recursion of these updates, the system state evolves in time, in directions depicted by the streamlines in each pane. Blue shading represents the acceptance rate for group one (i.e., $\Pr\Big(\hat{Y}{=}1 \mid G{=}1\Big)$). Orange shading likewise depicts the classifier's false positive rate for group one. By the symmetry of the setting, these vales apply equally to group two when reflected under exchange of the pane axes.

As demonstrated by Fig. 3.4, both demographic parity and equalized odds generally fail to establish qualification rate parity between the two groups (the set of states along the dark black line from lower left to upper right), despite the fact that demographic parity achieves equal nearly equal acceptance rates, equalized odds achieves nearly equal false positive rates (and true positive rates) across both groups, and qualification rate

parity is naturally established by the *lack* of intervention. As a result, it is clear that underlying disparities between the groups *persist* as a consequence of the intervention.

For demographic parity, the intervention requires "subsidizing" the under-qualified group at each time step, thereby "unintentionally selecting" for non-qualification in that group: Moreover, these subsidies grow larger with time as the system reaches an extreme state of near-universal qualification in one group and near-universal non-qualification in the other.

Likewise for equalized odds, the intervention arguably does more harm than good by preventing the system from reaching an equitable state that eliminates all statistical distinctions between groups, which occurs under laissez-faire dynamics.

We will provide more rigorous statements about the failure modes of these interventions when we perform a dynamical analysis of this example system in Section 4.2.

3.6 Adversarial Bounds on Short-Term Alignment Violations

Parts of this section are adapted from previously published work by Chen et al. (2022).

In the previous section, we established that misalignment can be a problem in the long-term, when distribution shift interacts with myopic optimization, but we relied on simplified models of distribution shift to do so. A more general approach for quantifying how distribution-shift can violate alignment constraints, at least in the short-term, may be developed within a simple adversarial framework. In this section, we consider an approach for quantifying the degree to which present normative interventions are (in)appropriate for alignment in dynamical contexts. Rather than assume specific dynamical reactions to policy, we consider worst-case or adversarial scenarios subject to map bounds distribution shift to bounds on constraint violations.

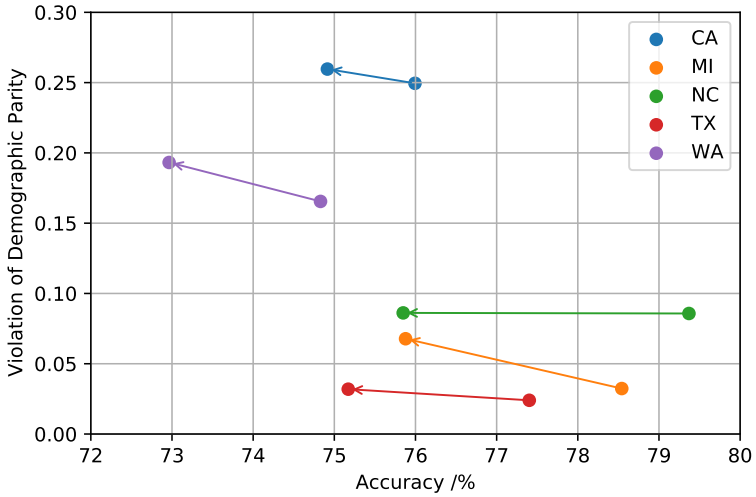

Figure 3.5: Decreases in accuracy and increases in fairness violations when a classifier initially trained to predict income on data specific to individual US states is reused on 2018 data—an example of distribution-shift over time (Chen et al., 2022).

Since it can be difficult in general to predict *how* the distribution will shift, we define a *vector-valued premetric* to quantify distribution-shift, then we suppose a bound on the maximum shift we may expect to observe. Subject to this bound, we then address the question of how much might we violate our fairness constraints *in the worst case*.

Definition 3.6.1 (Statistical Divergence)**.** Given any two distributions p and q, a *divergence* D is any real-valued function of the two distributions that satisfies *non-negativity* and the *identity of indiscernibles*. That is,

$$D(p \parallel q) \geq 0 \quad ; \quad D(p \parallel q) = 0 \iff p = q \tag{3.36}$$

We will refer to $D(p \parallel q)$ as the "*divergence* from q to p".

The identity of indiscernibles implies reflexive nullity (i.e., there is no measurable difference between two copies of the same object) but also requires that two indistinguishable distributions must, in fact, be the same distribution. Typical examples of a

statistical divergences are the Kullback-Leibler divergence and various f-divergences.

Definition 3.6.2 (Vector-Valued Distribution Shift)**.** Define the group-vectorized distribution shift $\mathbf{D}$ for a classification problem as

$$\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \coloneqq \sum_{g} \mathbf{e}_g D\Big(\Pr_{\mathcal{T}}(X, Y \mid G = g) \,\Big\|\, \Pr_{\mathcal{S}}(X, Y \mid G = g) \Big) \tag{3.37}$$

where $\mathbf{e}_g$ represents a unit vector indexed by g.

Intuitively, the vector-valued quantity $\mathbf{D}(\mathcal{T} \parallel \mathcal{S})$ measures how much the relevant variables (features and labels, for a classification problem) shift within each group.

Next, we suppose a bound on the group-specific shifts that might occur in the worst-case:

Assumption 3.6.1 (Bounded Distribution Shift)**.** The group-vectorized distribution shift from the source distribution $\mathcal{S}$ to the target distribution $\mathcal{T}$ is element-wise bounded by a non-negative vector $\mathbf{B}$, i.e.,

$$\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \mathbf{B} \quad ; \quad \mathbf{B} \succeq 0 \tag{3.38}$$

where $\preceq$ and $\succeq$ denote element-wise inequalities.

Finally, subject to the given bound Asm. 3.6.1, we define the maximal violation of the fairness constraint or alignment intervention:

Definition 3.6.3 (Supremal Disparity)**.** Define the supremal disparity of policy given by policy with parameters θ within distribution shift $\mathbf{B}$ of $\mathcal{S}$ as

$$\forall \mathbf{b} \succeq 0, \quad \mathcal{V}^{\star}_{\mathbf{b}}(\theta, \mathcal{S}) \coloneqq \sup_{\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \mathbf{b}} \mathcal{V}(\theta, \mathcal{T}) \tag{3.39}$$

The form of the corresponding bound assured by Asm. 3.6.1 is

$$\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \mathbf{B} \implies \mathcal{V}(\theta, \mathcal{T}) \leq \mathcal{V}^{\star}_{\mathbf{B}}(\theta, \mathcal{S}) \tag{3.40}$$

While additional bounds are explored by Chen et al. (2022), let us give the flavor of the possible results of this approach:

3.6.1 A Lipschitz Bound

For a fixed policy with parameters θ optimized for the source distribution $\mathcal{S}$, the supremal disparity achievable by distribution-shift within bound $\mathbf{b}$, i.e. $\mathcal{V}^{\star}_{\mathbf{b}}(\theta, \mathcal{S})$, defines a scalar field in the non-negative cone of $\mathbf{b} \succeq 0$. By treating $\mathcal{V}^{\star}_{\mathbf{b}}$ as a scalar potential, we choose to define the conservative vector field $\mathbf{F}$ such that

$$\mathbf{F}(\mathbf{b}) \coloneqq -\nabla_{\mathbf{b}} \mathcal{V}^{\star}_{\mathbf{b}} \tag{3.41}$$

Such a construction ensures the path-independence of the line integral of $\mathbf{F}$ along any curve C from 0 to $\mathbf{B}$. That is,

$$-\int_{C} \left\langle \mathbf{F}(\mathbf{b}), \mathrm{d}\mathbf{b} \right\rangle = \mathcal{V}^{\star}_{\mathbf{b}}(\theta, \mathcal{S}) - \mathcal{V}^{\star}_{0}(\theta, \mathcal{S}) \tag{3.42}$$

$$= \mathcal{V}^{\star}_{\mathbf{b}}(\theta, \mathcal{S}) - \mathcal{V}(\theta, \mathcal{S}) \tag{3.43}$$

Theorem 3.6.1 (Lipschitz Upper Bound for a Curve)**.** *Given an element-wise Lipschitz bound for* $\mathbf{F}$ *along any curve* C *with endpoints* 0 *and* $\mathbf{B}$*, i.e. when there exists some finite* $\mathbf{L} \succeq 0$ *such that*

$$\forall \mathbf{b} \in C, \quad \left|\mathbf{F}(\mathbf{b})\right| \preceq \mathbf{L}, \tag{3.44}$$

we may conclude

$$\mathcal{V}^{\star}_{\mathbf{B}}(\theta, \mathcal{S}) = \mathcal{V}(\theta, \mathcal{S}) - \int_{C} \left\langle \mathbf{F}(\mathbf{b}), \mathrm{d}\mathbf{b} \right\rangle \tag{3.45a}$$

$$\leq \mathcal{V}(\theta, \mathcal{S}) + \left\langle \mathbf{L}, \mathbf{B} \right\rangle \tag{3.45b}$$

Alternatively phrased,

$$\mathbf{D}(\mathcal{T} \parallel \mathcal{S}) \preceq \mathbf{B} \implies \mathcal{V}(\theta, \mathcal{T}) \leq \mathcal{V}(\theta, \mathcal{S}) + \mathbf{L} \cdot \mathbf{B} \tag{3.46}$$

To visualize this result, we depict the meaning of the Lipschitz bound for two groups $\{g, h\}$ in Fig. 3.6.

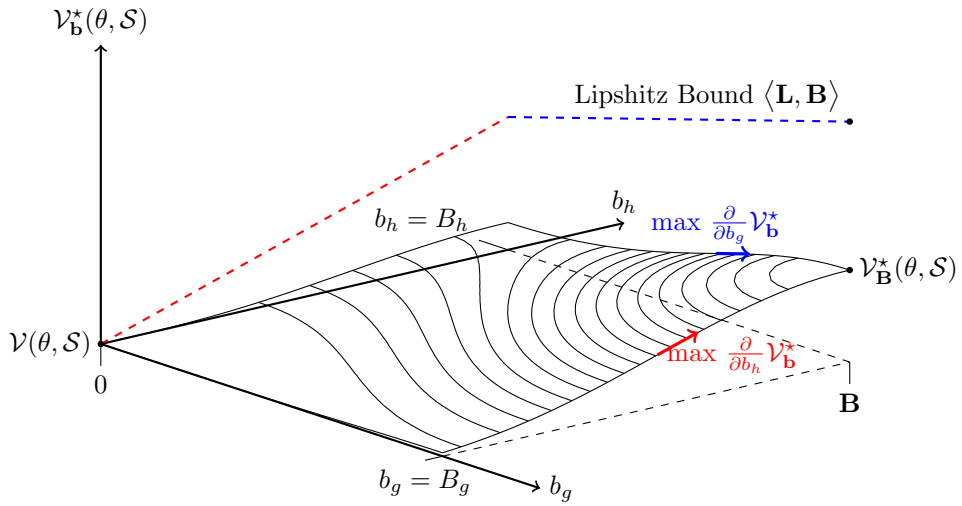

Figure 3.6: A Lipschitz bound in the form of Eq. (3.46) subject to Asm. 3.6.1, for two groups $\{g, h\}$ implies that that the supremal disparity $\mathcal{V}^{\star}_{\mathbf{B}}$ cannot exceed the path-integral of the maximum growth rate.

As we demonstrate in Section 3.6.2, Thm. 3.6.1 is not vacuous, and may be applied natural to certain settings introduced Section 3.1

3.6.2 Demographic Parity subject to Label Shift

Label Shift We may apply Thm. 3.6.1 in a setting in which distribution-shift result from a label-shift mechanism (Section 3.1.1) with known bounds.

For this example, let us measure group-specific distribution shifts between $\mathcal{S}$ and $\mathcal{T}$ by the change in proportion of true positive labels ($Y{=}1$) in each group, corresponding to a change in *qualification rates* s_g, as in

$$s_g(\mathcal{T}) \coloneqq \Pr_{\mathcal{T}}(Y{=}1 \mid G{=}g). \tag{3.47}$$

$$D_g(\mathcal{S} \parallel \mathcal{T}) \coloneqq \big|s_g(\mathcal{S}) - s_g(\mathcal{T})\big|. \tag{3.48}$$

Consistent with Asm. 3.6.1, we will consider such distribution shifts, within an element-wise measure bounded by $\mathbf{B}$.

Let us next denote the true positive rate and false positive rate of a policy with parameters θ on distribution $\mathcal{T}$ within group g as (respectively),

$$\beta_g^+(\theta, \mathcal{T}) \coloneqq \Pr_{\theta,\mathcal{T}}(\hat{Y}{=}1 \mid Y{=}1, G{=}g). \tag{3.49}$$

$$\beta_g^-(\theta, \mathcal{T}) \coloneqq \Pr_{\theta,\mathcal{T}}(\hat{Y}{=}1 \mid Y{=}{-}1, G{=}g). \tag{3.50}$$

Because β_g^+ and β_g^- are conditioned on the value of Y, they are invariant under label shift given a constant policy θ. Assuming a fixed, universal source distribution $\mathcal{S}$, we will hereafter elide the dependencies of these quantities, writing merely β_g^+ and β_g^-.

In terms of these quantities, we may write

$$\Pr_{\theta,\mathcal{T}}(\hat{Y}{=}1 \mid G{=}g) = \beta^+ s_g(\mathcal{T}) + \beta^-\big(1 - s_g(\mathcal{T})\big). \tag{3.51}$$

from which it follows that

$$\frac{\partial}{\partial s_g(\mathcal{T})} \Pr_{\theta,\mathcal{T}}(\hat{Y}{=}1 \mid G{=}g) = \beta_g^+ - \beta_g^-. \tag{3.52}$$

Demographic Parity Next, departing from the convention established in Section 2.2.1 for this example (for theoretical simplicity), let us choose to measure demographic parity as

$$\mathcal{V}(\theta, \mathcal{T}) \coloneqq \sum_{g,h \in \mathcal{G}} \left| \Pr_{\mathcal{T}}(\hat{Y}{=}1 \mid G{=}g) - \Pr_{\mathcal{T}}(\hat{Y}{=}1 \mid G{=}h) \right|. \tag{3.53}$$

Differentiating Eq. (3.53), we obtain a Lipschitz condition:

Lemma 3.6.2 (Disparity Rates)**.** *The supremal disparity $\mathcal{V}^\star_{\mathbf{B}}$ is subject to a Lipschitz condition given by*

$$\forall g, \quad \frac{\partial}{\partial b_g} \mathcal{V}^\star_{\mathbf{b}}(\theta, \mathcal{S}) \leq L_g = (|\mathcal{G}| - 1) \left| \beta_g^+ - \beta_g^- \right|. \tag{3.54}$$

It follows that such a measure of group-wise disparity is subject to a Lipschitz bound, as in Thm. 3.6.1:

Theorem 3.6.3 (Demographic Parity under Label Shift)**.** *For demographic parity (Eq.* (3.53) *subject to label-shift (Eq.* (3.51)*),*

$$\mathcal{V}(\theta, \mathcal{T}) - \mathcal{V}(\theta, \mathcal{S}) \leq (|\mathcal{G}| - 1) \sum_g \left| s_g(\mathcal{S}) - s_g(\mathcal{T}) \right| \left| \beta_g^+ - \beta_g^- \right|. \tag{3.55}$$

Interpreting this result, when β_g^+ is close to β_g^-, the policy looks like a random classifier, and label shift has limited effect on statistical group disparity. When $|\beta_g^+ - \beta_g^-|$ is large, indicating high classifier accuracy, our bound exposes a direct trade-off between

accuracy and fairness transferability guarantees.

3.6.3 Numerical Validation

We may evaluate Thm. 3.6.3 in the system described by Section 3.5, a binary classification setting with an analytical model of population response based on replicator dynamics: We compare the our theoretical result in Thm. 3.6.3 to the simulated dynamics of this system subject to a policy constrained by demographic parity.

As in Section 3.5.2, we graphically represent all possible states of the example system by the state vector of qualification rates for two equal-sized groups. We assume that the classifier in each state is trained by (RRM; Section 3.3.1; Eq. (3.29)) subject to demographic parity.

Subject to the dynamics prescribed by Eq. (3.32), we depict the local evolution of the state, using streamlines, as well as the *rate of change* of the violation of fairness—i.e. prior to retraining.

We compare this latter quantity to our the theoretical bound.

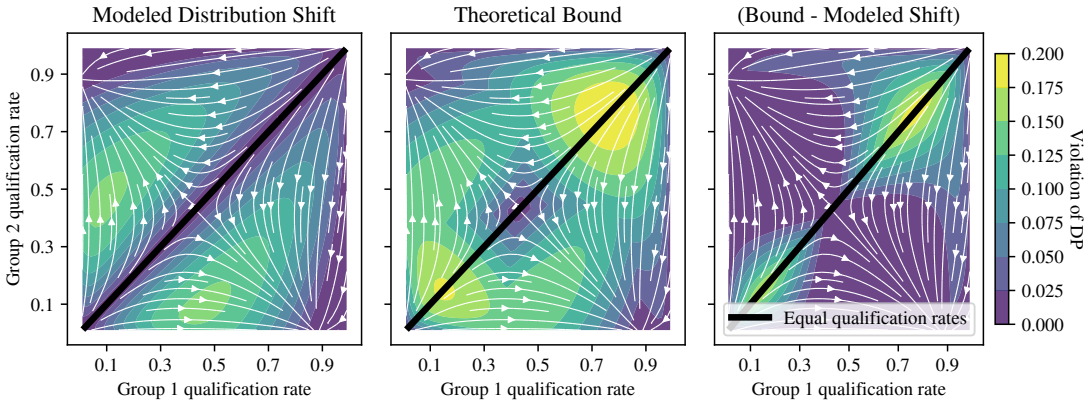

Figure 3.7: A policy satisfying demographic parity is subject to distribution shift prescribed by replicator dynamics. The change in qualification rates subject to this policy follows the streamlines, while the change in $\mathcal{V}$ prior to retraining is represented with color and compared to the theoretical bound (Thm. 3.6.3) (Thm. 3.6.3).

Interpreting our results, we note that the bound lacks information about the *relative directions* of the change in acceptance rates for each group. It thus over-approximates possible fairness violations when group acceptance rates adapt with the same sign. When group acceptance rates move in opposing directions, however, the bound gives excellent agreement with the modelled replicator dynamics.

4 Alignment with Dynamics

Can we anticipate and utilize adaptive agents' reactions to data-driven policy to achieve aligned objectives dynamically?

In Section 2.3, we commented on the fact that, in real-world deployments, machine learning policies frequently encounter *distribution-shift*. Moreover, we highlighted how ML policies can *induce* such distribution-shift directly, as intelligent agents react in their own best interest to implemented policies.

In Chapter 3 we observed that the standard reaction to such distribution-shift, i.e., simply retraining or fine-tuning the policy to the shifted distribution, can have undesirable consequences: Even without constraints, retraining for distribution-shift *ex-post* can lead to increasing loss over time, as explained and demonstrated in Section 3.4. When present normative constraints are incorporated to ensure "alignment", as considered in Section 3.5, the situation can also degrade.

These observations indicate that dealing with alignment in a dynamical context requires *anticipation* of policy-induced distribution shift and other agents' reactions to policy. In this section, we ask how we can incorporate such anticipation into ML policies such that we can realize aligned objectives in the long-term, over multiple policy iterations and interactions with a multiagent environment.

By focusing on alignment as an inherently dynamical problem shaped by the interaction of multiple intelligent agents, our perspectives and tools intersect with economics, game-theory, and control: To an economist or a game theorist, recognizing the organization of human society as a multiagent system, the question is whether an ML agent can induce transitions between or avoid undesirable *Nash equilibria*. To a control theorist, the question is whether an ML agent can exert *control* over the system to steer

it to desirable states and rewards.

We explore two techniques for addressing our questions: The first maps techniques from constrained optimization to sequential policies based on local information about policy-induced distribution-shift. The second considers more general dynamical models and is based on feedback control. We conclude the chapter by considering how an explicitly dynamical treatment of alignment opens new possibilities for explicit alignment objectives.

4.1 Sequential Policies as Optimization Programs

Parts of this section are adapted from previously published work by Raab et al. (2024).

How should a policy-maker anticipate and algorithmically address policy-induced distributions shift?

One way to do this, as outlined in Section 3.2.1 is to model the distribution as a function of policy, such that

$$\mathcal{D}^t = \mathrm{S}(\theta^t). \tag{4.1}$$

Subject to this assumption, the constrained optimization problem we wish to solve is

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \mathcal{L}(\theta, \mathrm{S}(\theta)) \\ \text{subject to} \quad & \mathcal{V}(\theta, \mathrm{S}(\theta)) \leq 0\,. \end{aligned} \tag{4.2}$$

In general the function S may be *unknown* or difficult to model, but (assuming its existence) this doesn't prevent us from using the available information to make iterative improvements.

Even when S is unknown in general, the idea of using limited or *local* (to the currently

deployed policy) information about an objective function and a constraint function is well-established by the use of *sequential programs* in constrained optimization (i.e., in a style similar to that of Section 2.1.3 and expanded on in Appendix B). Inspired by this connection, we may adapt standard, first-order numerical algorithms (belonging to the same family as repeated gradient descent Section 3.3.2) from sequential programming to Prob. (4.2).

To use such methods, we require not only the current values of $\mathcal{L}$ and $\mathcal{V}$ on θ^t and $\mathcal{D}^t$, but their *gradients*. For compactness, we denote

$$\nabla \mathcal{L}^t \equiv \nabla_\theta \mathcal{L}(\theta)\big|_{\theta=\theta^t}. \tag{4.3}$$

$$\nabla \mathcal{V}^t \equiv \nabla_\theta \mathcal{V}(\theta)\big|_{\theta=\theta^t}. \tag{4.4}$$

Assumption 4.1.1 (Gradients of Deployed Policy)**.** At each time t, the policy-maker is able to observe $\nabla \mathcal{L}^t$ and $\nabla \mathcal{V}^t$, i.e., the policy-maker has knowledge of the first-order dependence of $\mathcal{L}$ and $\mathcal{V}$ on θ at the currently deployed policy θ^t.

Asm. 4.1.1 is reasonable when $\mathcal{L}$ corresponds to empirical risk and $\mathcal{V}$ measures disparities between sets in the population, as in Section 2.2.1: With small, random perturbations to policies over the set of individuals in the population, first-order statistics can provide an estimate of the local dependence of $\mathcal{L}$ and $\mathcal{V}$ on θ (i.e., via (conditional) correlations between policy perturbations and outcomes). Conceptually, the policy-maker can estimate gradients from A/B testing, where slightly different policies are deployed on statistically independent subsets of the population: The local dependence of group participation rates on group losses may be estimated by the finite difference method, using the differences in quantities across policies.

4.1.1 Constrained Projected Gradient

To address Prob. (4.2) in an environment satisfying Asm. 4.1.1, we propose a method related to Fletcher's smooth exact penalty function (Fletcher, 1973; Conn et al., 2000). For further background, we also refer the reader to Nocedal and Wright (1999). Our method involves solving a sequential quadratic program parameterized by step size $\eta > 0$ and a scale factor $\alpha > 0$. We refer to this method as "Constrained Projected Gradient" (CPG), shown below.

$$\begin{aligned} \theta^{t+1} = \arg\min_{\theta \in \mathcal{A}} \quad & \left\langle \theta, \nabla \mathcal{L}^t \right\rangle + \frac{1}{2\eta}(\theta - \theta^t)^2. \\ \text{subject to} \quad & \left\langle \theta - \theta^t, -\nabla \mathcal{V}^t \right\rangle \geq \alpha \mathcal{V}^t. \end{aligned} \tag{CPG}$$

This method provably solves Prob. (4.2) in the limit of small η, subject to the following two assumptions:

Assumption 4.1.2 (Feasibility)**.** The fairness constraint is feasible. That is, $\exists \theta^\star \in \mathcal{A}$ such that $\mathcal{V}(\theta^\star) \leq 0$. Furthermore, the subproblem in Eq. (CPG) is feasible at each time step t.

The second stipulation of Asm. 4.1.2 eliminates the possibility that, for example, $\nabla \mathcal{V}^t = 0$ and $\mathcal{V}^t > 0$.

Assumption 4.1.3 (Properties of Disparity)**.** $\mathcal{V}$ is an invex function; that is, every critical point of $\mathcal{V}$ is a global minimum.

Asm. 4.1.3 is easily satisfied by choosing a suitable $\mathcal{V}$ that maintains the required zero-level set (e.g., as in Eq. (4.7)).

Theorem 4.1.1 (Asymptotic Convergence)**.** *Subject to Asms. 4.1.1 to 4.1.3, as* $(t \to \infty)$*, constrained projected gradient (Eq.* (CPG)*) converges to a feasible local optimum of the*

objective when the step size η is sufficiently small.

We provide a proof *sketch* below. We defer rigorous proof to Appendix C and briefly outline how CPG relates to Fletcher's smooth exact penalty function subject to convex constraints in Appendix B.

Proof Sketch. Our proof relies on establishing that CPG first achieves fairness, then converges to a critical point of the objective function. First, note that $\mathcal{V}^t > 0 \implies \left\langle \theta^{t+1} - \theta^t, -\nabla \mathcal{V}^t \right\rangle > 0$, subject to the constraints imposed by $\mathcal{A}$, by the fairness constraint of Eq. (CPG) and Asm. 4.1.2. That is, when the current policy is *unfair*, the algorithm makes progress towards fairness by decreasing disparity. Second, we show that $\mathcal{V}^t \leq 0 \implies \left\langle \theta^{t+1} - \theta^t, -\nabla \mathcal{L}^t \right\rangle > 0$. That is, once the current policy is *fair*, the algorithm decreases loss. This second fact follows from the fact that $\mathcal{V}^t \leq 0$ implies that the sign of $\left\langle \theta^{t+1} - \theta^t, -\nabla \mathcal{V}^t \right\rangle$ is unconstrained, and minimization of the objective will naturally ensure $\left\langle \theta^{t+1} - \theta^t, \nabla \mathcal{L}^t \right\rangle < 0$ subject to the constraints of $\mathcal{A}$.

4.1.2 Application to Fair Participation

In this section, we detail how CPG can be applied to the setting explored in Section 3.4.1, with an added constraint. To briefly outline the setting again, we consider a recommendation task in which a policy-maker or *firm* is free to choose between policies that map to a static set of achievable group-losses $\mathcal{A}$ at each time step. We therefore consider the vector of group losses $\boldsymbol{\ell} \in \mathcal{A}$ as the decision variable ($\theta \equiv \boldsymbol{\ell}$). In each group, the loss ℓ_g maps to a monotonically decreasing participation rate $\rho_g = \varphi_g(\ell_g)$.

Because the distribution (summarized by $\boldsymbol{\rho}$ as discussed in Section 3.2.1) is a fixed function of the decision variable, we may express the loss function for the firm as a

function of $\boldsymbol{\ell}$ alone:

$$\mathcal{L}(\boldsymbol{\ell}) \coloneqq \langle \boldsymbol{\ell}, \boldsymbol{\rho} \rangle. \tag{4.5}$$

$$= \sum_g \ell_g \varphi_g(\ell_g). \tag{4.6}$$

To this objective, we now introduce a *fairness constraint* with slack variable $\varepsilon \geq 0$:

$$\mathcal{V}(\boldsymbol{\ell}) \coloneqq \operatorname*{Var}_g[\rho_g] + \varepsilon. \tag{4.7}$$

This constraint represents the imperative for the firm to maintain a demographically representative user-base, which can serve, for example, to mitigate negative public relations and diversify advertising revenue. We will comment on fundamental differences between the form of this constraint and those typically considered as "present normative" (Section 2.2) in Section 4.3.

The overall objective of the firm is therefore

$$\begin{aligned} \underset{\boldsymbol{\ell}}{\text{minimize}} \quad & \mathcal{L}(\boldsymbol{\ell}) \coloneqq \langle \boldsymbol{\ell}, \boldsymbol{\rho} \rangle. \\ \text{subject to} \quad & \mathcal{V}(\boldsymbol{\ell}) \leq 0. \end{aligned} \tag{4.8}$$

This objective has a form that is amenable with Prob. (4.2) and may be solved by a sequence of policies rendered by CPG.

4.1.3 Experiments

In this section, we evaluated CPG in multiple semi-synthetic settings to compare it to repeated gradient descent (RRM; Section 3.3.2) as well as "myopic projected gradient" (MPG), an amended form of repeated gradient descent that accounts for distribution

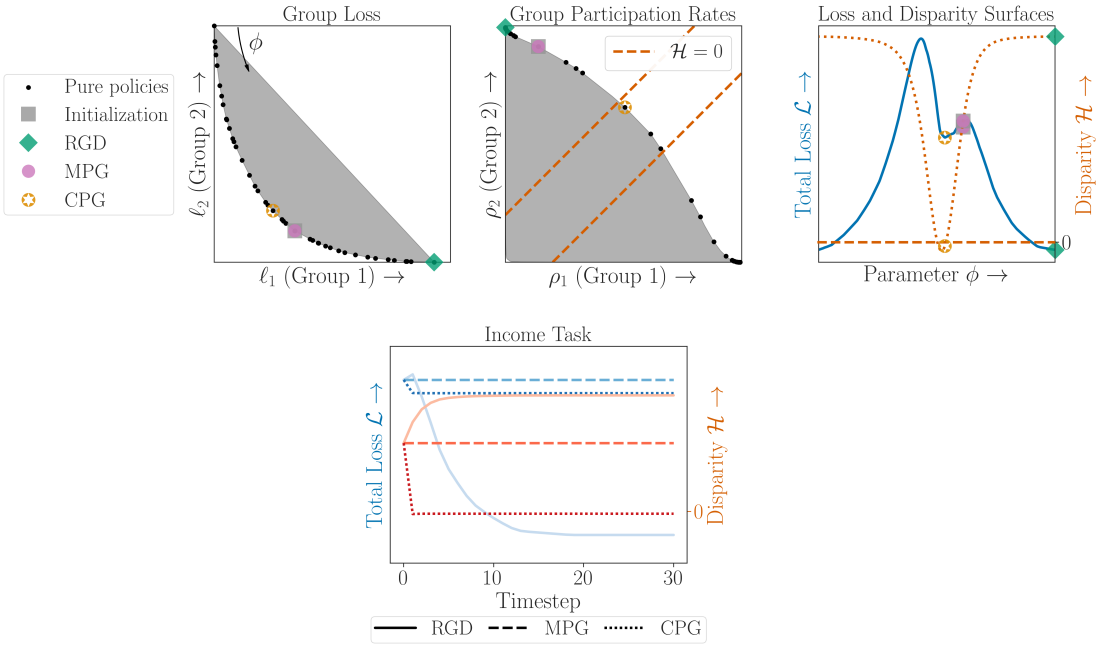

Figure 4.1: Income task. This setting has a highly non-convex loss surface, shown in the third pane, and demonstrates a situation in which CPG converges to the unique solution, MPG gets stuck in an unfair local minimum of the utility function, and RRM diverges to the highest disparity.

shift but does not incorporate constraints.

Concisely, MPG incorporates the gradient of ρ into the calculation of $\nabla\mathcal{L}$, but does not address the fairness constraint $\mathcal{V} \leq 0$.

$$\boldsymbol{\ell}^{t+1} = \arg\min_{\boldsymbol{\ell} \in \mathcal{A}} \quad \langle \boldsymbol{\ell}, \nabla\mathcal{L}^t \rangle + \frac{1}{2\eta}(\boldsymbol{\ell} - \boldsymbol{\ell}^t)^2, \tag{MPG}$$

We use MPG to demonstrate that the constraint $\mathcal{V} \leq 0$ in Prob. (4.2) is not automatically satisfied by accounting for policy-induced distribution shift as in Eq. (MPG).

Datasets Our settings derive from binary classification tasks on the American Community Survey Public Use Microdata Sample (ACS PUMS) dataset[1], as introduced by

[1] `https://github.com/socialfoundations/folktables`

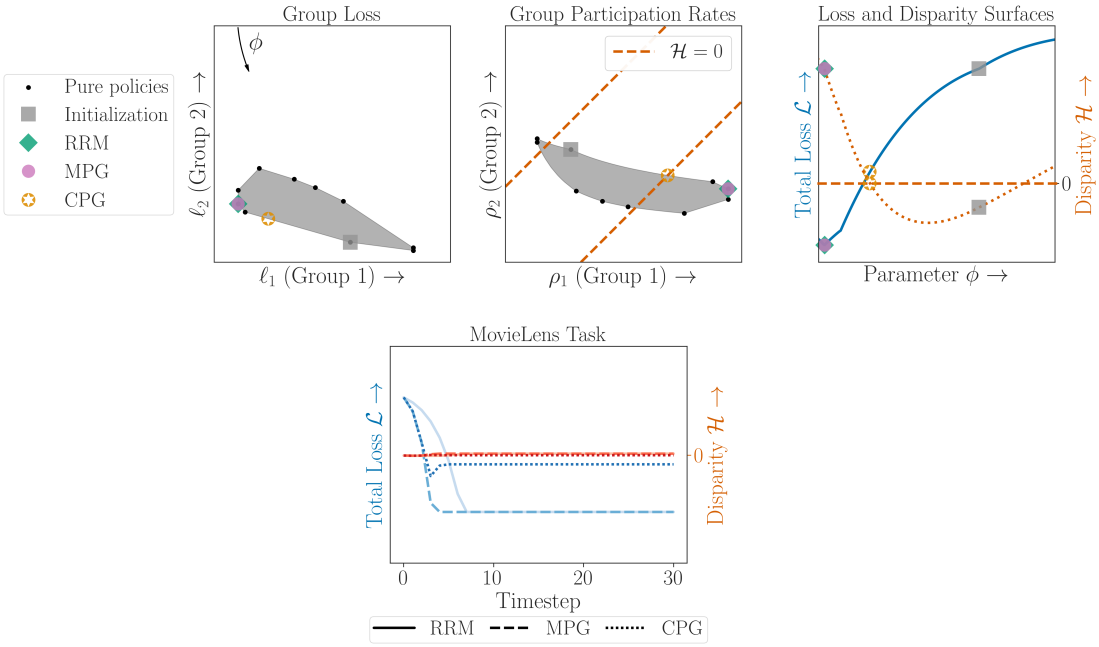

Figure 4.2: MovieLens Task. This particular setting demonstrates application of our method to a recommendation task. Only our proposed method, CPG, satisfies the fairness constraint, finding a solution to Prob. (4.2) at the boundary of the feasible set, where $\langle \nabla \mathcal{L}, \nabla \mathcal{V} \rangle < 0$. Both RRM and MPG locally optimize local utility at the expense of fairness.

Ding et al. (2021b), for specific US states in 2018, or a recommendation task on movie preferences using data (MovieLens) collected by Harper and Konstan (2015). Each task gives samples of joint feature (X), label ($Y \in \{0, 1\}$), and group ($G \in [k]$) variables, the joint distribution of which we summarize by writing $\mathcal{S}$.

Model Class and Achievable Losses For each task, we first define a set of achievable losses $\mathcal{A}$. We generate $n = 100$ different binary classifiers and record the vector of group-specific losses $\boldsymbol{\ell}$ achieved by the predicted labels $\hat{Y}$ for each classifier, where we define ℓ_g as the negative binary prediction accuracy conditioned on group g:

$$\ell_g = - \mathop{\mathrm{E}}_{X,Y,G \sim \mathcal{S}} \left[\hat{Y} = Y \mid G = g \right]. \tag{4.9}$$

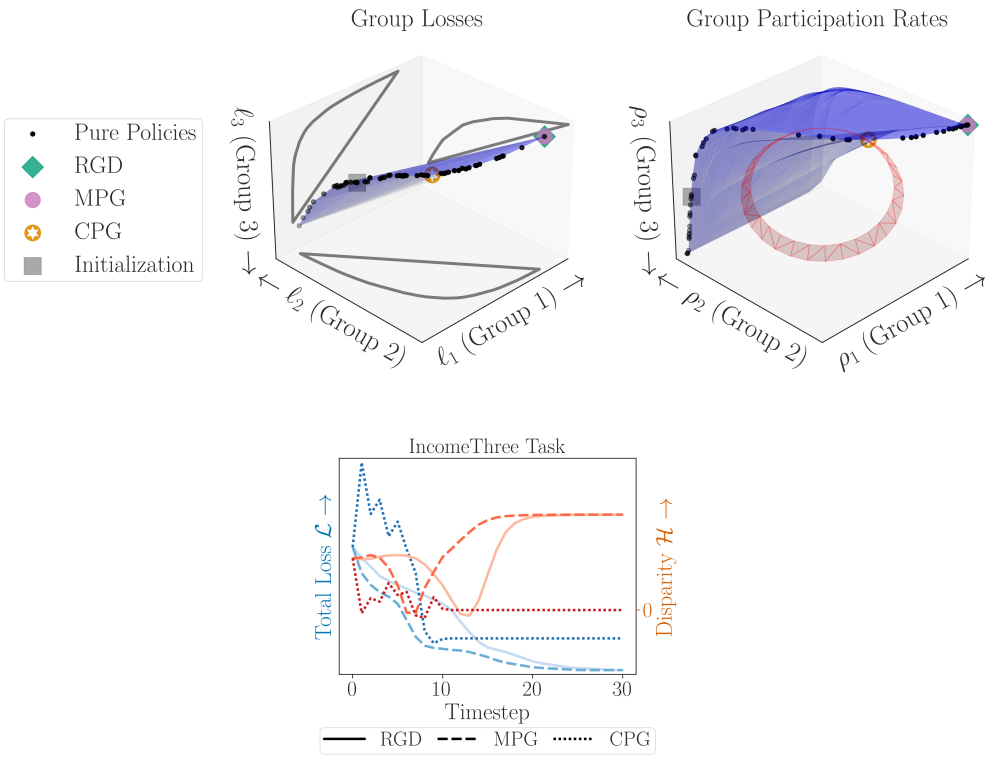

Figure 4.3: IncomeThree Task. This graphic is intended to showcase that Prob. (4.8) and CPG are not restricted to only two groups, and in general allow very large numbers of groups, though the resulting sets of achievable group losses and participation rates are difficult to visualize. Inspecting the time-series, we see that CPG increases disparity at intermediate time-steps, despite starting outside of the feasible (fair) set; we attribute this to the large initial step size, which is not guaranteed to eliminate non-linear behaviors of $\mathcal{L}$ and $\mathcal{V}$ within the linear approximation trust-region.

The set of achievable loss vectors $\mathcal{A}$ for the task is defined by the convex hull of these samples $\mathcal{A} = \text{Hull}(\{\boldsymbol{\ell}_i : i \in [n]\})$.

In our experiments, we consider logistic classifiers trained on different weighted logistic loss functions for the binary classification task, intended to simulate variable participation rates. For each classifier $i \in [n]$, we sample a vector of objective function term weights $\boldsymbol{\beta}$ uniformly at random from the $(k-1)$-simplex ($\sum_{g=1}^{k} \beta_g = 1$) and

solve the regularized logistic classification task

$$\begin{aligned}
&\min_{\boldsymbol{w}} \quad \sum_{g=1}^{k} l_g(\boldsymbol{w})\beta_g + \frac{1}{2}|\boldsymbol{w}|^2; \\
&l_g(\boldsymbol{w}) = \mathop{\mathrm{E}}_{X,Y,G\sim\mathcal{S}}\left[H_Y(h_{\boldsymbol{w}}(X)) \mid G = g\right]; \\
&H_p(q) = -p\log q - (1-p)\log(1-q); \\
&h_{\boldsymbol{w}}(X) = \frac{1}{1+e^{-\langle X,\boldsymbol{w}\rangle}}.
\end{aligned} \tag{4.10}$$

We train each possible classifier using the limited-memory method of Broyden, Fletcher, Goldfarb, and Shanno (LBFGS) (Liu and Nocedal, 1989), as implemented by scikit-learn (Pedregosa et al., 2011).

Synthetic Distribution Shift We model the function f that maps group loss to group participation rate as a reversed logistic function parameterized by bias and sensitivity parameters $b \in (-1, 0)$ and $s > 0$, respectively, and clipped to the interval $[0, 1]$. That is, we model φ as

$$\begin{aligned}
&\varphi(x) = \max\left[0, \min[1, \varrho(x)]\right]; \\
&\varrho(x) = \frac{1}{1+e^{s(x-b)}}; \quad x \in [-1, 0].
\end{aligned} \tag{4.11}$$

For intuition, this same function is used in the upper-left panel of Fig. 3.2 with parameters ($s = 20, b = -0.62$).

Hyperparameters We use a learning rate that decays as a harmonic series:

$$\eta^t = \eta^1/t; \quad t \in \{1, 2, ...\}. \tag{4.12}$$

All experiments follow the same decay schedule and run for the same number of steps (i.e., 30), but the initial learning rate η^1 is equal to half of the diameter of $\mathcal{A}$. Each experiments run in less than 60 seconds on a typical laptop CPU.

We set initial conditions $\boldsymbol{\ell}^0$, the participation function parameters (b, s) (Section 4.1.3), and the fairness slack parameter ε (Eq. (4.7)) to demonstrate qualitatively diverse simulation outcomes among our included results.

Results In all experiments, CPG achieved a feasible local optimum of the objective, while RGD and MPG did so only rarely. In (Figs. 4.1 to 4.3), we highlight a few examples of our experimental results on the following tasks:

- **Income**: "Income" task of Ding et al. (2021b) with groups redefined to coincide with the binary classification label and restricted to data from Alabama.
- **MovieLens**: From user age, occupation, and gender, predict whether this user exhibits a stronger-than-median preference for "mystery" rather than "adventure" films, using zero-one loss and targeting equal user rates across gender. The data for this task comes from Harper and Konstan (2015).
- **IncomeThree**: "Income" task of Ding et al. (2021b) with groups expanded to three divisions of income (below \$60K, between \$60K and \$120K, above \$120K) and restricted to data from Alabama.

As our algorithms are deterministic, we do not consider multiple runs for the same setting and leave characterizations of the robustness of these algorithms in terms of different hyperparameters to future work.

In Figs. 4.1 to 4.3, the first pane visualizes the set of achievable losses, $\mathcal{A}$, and the samples used to generate it, where the axes correspond to each group's loss. The

second pane visualizes the corresponding set of achievable participation rates $\boldsymbol{\rho}$ with axes corresponding to each group's participation rate. The last pane plots total loss and disparity vs. time-step for all three methods in the given setting. In Figs. 4.1 and 4.2, an additional pane demonstrates the non-convexity of the total loss and disparity surfaces along a curve corresponding to all $\boldsymbol{\ell} \in \mathcal{A}$ which maximize distance from the origin, with angle from the x-axis parameterized by ϕ. In all figures, a distinct marker represents each method (i.e., RRM, MPG, CPG) and their shared initialization across all panes.

4.2 Alignment via Feedback Control

Parts of this section are adapted from previously published work by Raab and Liu (2021).

In this section, we revisit the dynamical environment of Section 3.5.1. In this setting, agents affected by a binary classification policy (Section 2.1.4) used for approving loans must choose whether to become (un)qualified for the loan as time evolves. Modelling this decision according to evolutionary game theory, we assume that group qualification rates obey the replicator equation (Eq. (3.10) and Appendix A).

We recall

$$s_g^t = \frac{s_g^{t-1} Q_g(\theta^t)}{s_g^{t-1} Q_g(\theta^t) + (1 - s_g^{t-1}) N_g(\theta^t, t)}, \tag{4.13}$$

where s_g^t represents the qualification rate of group g at time /(t), $Q_g(\theta^t)$ represents the expected utility to an agent in group g of qualification as a consequence of the policy parameterized by θ^t, and $N_g(\theta^t)$ likewise represents the expected utility of non-qualification.

As shown in Section 3.5, it is possible for present-normative interventions in this setting, such as demographic parity or equalized odds (Section 2.2) to backfire, artificially

causing underlying inequities in the population to persist, requiring interventions with increasing magnitude over time, and increasing loss for the policy-maker.

Laissez-Faire Solutions

Before analyzing how we can utilize the dynamics of the setting to refine our algorithm design, we first perform a dynamical analyses of the laissez-faire setting, in which fairness constraints are absent, when the policy-maker uses repeated risk minimization (RRM, Section 3.3.1). We recall that the loan-approval policy is a threshold policy, accepting applicants with features above a group-specific threshold $X \geq \theta_g$ and rejecting all others. We also recall the definition of $q_y(x) \coloneqq \Pr(x \mid Y{=}y)$ where $y \in \{-1, 1\}$ (Eq. (3.26)), such that $q_1(x)/q_{-1}(x)$ is a monotonically increasing function of x (Asm. 3.5.1).

In this case, the thresholds θ_g^{t+1} may be found in closed form by inverting the equations

$$\forall g, \quad \frac{q_1(\theta_g^{t+1})}{q_{-1}(\theta_g^{t+1})} = \left(\frac{1 - s_g^t}{s_g^t}\right)\left(\frac{C_{-1,-1} - C_{-1,1}}{C_{1,1} - C_{1,-1}}\right), \tag{4.14}$$

where $C_{y,\hat{y}}$ represents the utility to the classifier of an individual outcome with qualification y and decision $\hat{y}$. This solution corresponds to setting the group-thresholds at the feature values for which an individual in each group yields an expected classifier utility that is independent of whether the individual is qualified or unqualified.

Note, as a consequence of Asm. 3.5.1, that

$$\frac{\partial}{\partial \theta_g^{t+1}} s_g^t < 0. \tag{4.15}$$

That is, as the qualification rate s of group g increases, the critical threshold value θ_g with *decrease* in response. Intuitively, when a group is statistically more qualified,

middling credit scores are more likely to be the consequence of noise for truly qualified individuals while only very low credit scores reliably indicate non-qualification for a loan. As a result, it becomes *easier* to get a loan when your peer-group is statistically more qualified.

Equilibrium Given analytic solutions for the RRM policy without interventions, we can solve explicitly for system equilibria. Let us abbreviate $Q_g(\theta_g^{t+1}) \equiv Q_g^{t+1}$ and $N_g(\theta_g^{t+1}) \equiv N_g^{t+1}$. Inspecting Eq. (3.32), we notice that

$$Q_g^{t+1} > N_g^{t+1} \implies \frac{Q_g^{t+1}}{s_g^t Q_g^{t+1} + (1 - s_g^t) N_g^{t+1}} > 1 \implies s_g^{t+1} > s_g^t. \tag{4.16a}$$

$$Q_g^{t+1} = N_g^{t+1} \implies \frac{Q_g^{t+1}}{s_g^t Q_g^{t+1} + (1 - s_g^t) N_g^{t+1}} = 1 \implies s_g^{t+1} = s_g^t. \tag{4.16b}$$

$$Q_g^{t+1} < N_g^{t+1} \implies \frac{Q_g^{t+1}}{s_g^t Q_g^{t+1} + (1 - s_g^t) N_g^{t+1}} < 1 \implies s_g^{t+1} < s_g^t. \tag{4.16c}$$

and, therefore,

$$\operatorname{sgn}\left(s_g^{t+1} - s_g^t\right) = \operatorname{sgn}\left(Q_g(\theta_g^{t+1}) - N_g(\theta_g^{t+1})\right). \tag{4.17}$$

This provides a succinct description of equilibrium: Each group is in equilibrium when qualification and non-qualification yield the same expected utility (discounting the state in which everyone is (non-)qualified):

$$\text{Equilibrium} \iff \Big(Q_g(\theta_g) = N_g(\theta_g) \quad \text{or} \quad s_g \in \{0, 1\}\Big). \tag{4.18}$$

Before proceeding, we recall Asm. 3.5.2, which states that agents care about the decision of the classifier, and qualified agents prefer to get the loan (i.e. $A_{-1,1} \neq A_{-1,-1}$ and $A_{1,1} > A_{1,-1}$).

Theorem 4.2.1 (Strict Quasiconcavity). *$Q_g(\theta_g) - N_g(\theta_g)$ is strictly quasi-concave in θ_g. This guarantees that no more than two zeros of the function $(Q_g(\theta_g) - N_g(\theta_g))$ exist, each corresponding to a potential equilibrium.*

Stability Having described the possible equilibria of the system, we now determine whether these equilibria are stable or unstable.

Let us denote the possible zeros of $(Q_g - N_g)$ as θ_g^- and θ_g^+, such that $\theta_g^- < \theta_g^+$. Note that the superscripts of $(+)$ and $(-)$ do not refer to relative time indices. By Eq. (4.17), only θ_g^- is stable, as depicted in Fig. 4.4.

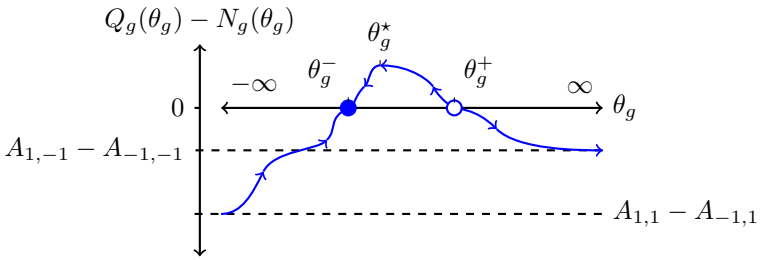

Figure 4.4: $Q_g(\theta_g) - N_g(\theta_g)$ (blue curve) is a strictly quasi-concave function of θ_g. We use $\theta_g^\star$ to denote the unique maximum. The direction of the arrows is a consequence of Eq. (4.17) and Eq. (4.15).

4.2.1 Intervention Geometry

Equipped with an analytical handle on the optimal classifier policy for this setting, we briefly discuss a geometric interpretation of equalized odds in this setting.

Given our assumption that q_y is group-independent, the only way to satisfy equalized odds with threshold policies is to use a group-independent threshold value $\overline{\theta}$ for all g. Under this constraint, we can solve explicitly for the global threshold $\overline{\theta}$ in terms of the

average qualification rate $\overline{s}$ substituted into Eq. (4.14):

$$\frac{q_1(\overline{\theta}^{t+1})}{q_{-1}(\overline{\theta}^{t+1})} = \left(\frac{1-\overline{s}^t}{\overline{s}^t}\right)\left(\frac{C_{-1,-1}-C_{-1,1}}{C_{1,1}-C_{1,-1}}\right), \quad \text{where} \quad \overline{s}^t \coloneqq \mathop{\mathrm{E}}_{G}[s_G^t]. \tag{4.19}$$

Geometrically, all states with a common $\overline{s}$ value form a hyperplane in the vector space of $\mathbf{s}$. Generalizing Thm. 4.2.1, it follows that there are at most two distinct equilibrium *hyperplanes*. This geometry is apparent in the Equalized Odds panes of Fig. 3.4, where the streamlines terminate along a single hyperplane (the line from upper left to lower right).

We reiterate that Equalized Odds does not, in general, lead to qualification rate parity nor group-independent statistics.

4.2.2 Basic Feedback Control

In the setting under consideration, given that laissez-faire policies converge to equitable states, should we dispense with fairness constraints altogether? Perhaps not: Such constraints do at least satisfy intuitive notions of fairness, even if they lead to undesirable long-term results. Is there a way to parameterize the trade-off between the two?

We show that it is possible to interpolate between short-term (present normative) and long-term notions of fairness by perturbing the equalized odds policies in the direction of laisez-faire policies. That is, with parametrically small violations of equalized odds, we can guarantee convergence of the system to the an equitable equilibrium with group-independent qualification rates.

Using $\overline{\theta}$ to denote the group-independent policy given by Eq. (4.19), then we may write an appropriate perturbation for the case of two equal-sized groups as a parametric

example of *feedback control*:

$$\theta_1^{t+1} = \overline{\theta}^{t+1} + \varepsilon \frac{\left(s_1^t - s_2^t\right)}{2s_1^t(1-s_1^t)}. \quad ; \quad \theta_2^{t+1} = \overline{\theta}^{t+1} - \varepsilon \frac{\left(s_1^t - s_2^t\right)}{2s_2^t(1-s_2^t)}. \tag{4.20}$$

We derive the analogous perturbation for N groups of differing size in Raab and Liu (2021). Importantly, this policy does not require intimate knowledge of the utility structure of the classifier nor of the agents, and is driven by measurements of quantities that should be observable in hindsight (true qualification disparities). This intervention also has the desirable property of being *self-abating* (it is only required to *reach* a target equilibrium, rather than maintain it). As the system transitions to the target equilibrium, the intervention vanishes, reverting to laissez-faire, natural equilibrium policies.

4.2.3 Experiments

We now compare the qualitative effects of the feedback policy given in Eq. (4.20) to the set of policies previously considered in Section 3.5.1 in simulation.

Fig. 4.5 provides "phase diagrams" of the dynamics that arise from the mutual recursion between policy-maker and the population under different interventions. It mirrors the presentation of Fig. 3.4, but includes an example of the feedback control policy (Eq. (4.20)) in the third pane. We highlight that the feedback control policy achieves the same, desirable equilibrium that the laissez-faire policies do while minimally violating equalized odds (represented visually by the symmetry of shading, which is perfect in the second pane and only slightly distorted in the third pane).

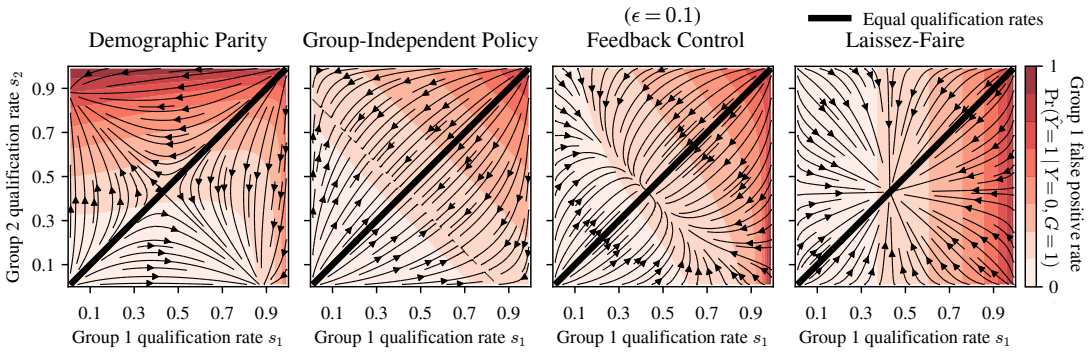

Figure 4.5: Simulated dynamics for two groups of equal size, subject to different global interventions. Streamlines approximate system time evolution. Color represents the false positive classification rate for group 1. q_{-1} and q_1 are Gaussians with unit variance and have means -1 and 1, respectively. For this example, $(A_{-1,-1} = 0.1; A_{-1,1} = 5.5; A_{1,-1} = 0.5; A_{1,1} = 1.0; C_{-1,-1} = 0.5; C_{-1,1} = -0.5; C_{1,-1} = -0.25; C_{1,1} = 1.0)$.

4.3 New Possibilities for Algorithmic Fairness

In Sections 4.1 and 4.2, we introduced new fairness constraints: The first of these was participation rate parity, while the second was qualification rate parity.

Importantly, these definitions do *not* fit into the paradigm of present normative interventions. They are not operative without knowledge of distribution shift, and are not compatible with the assumption that the distribution is static. Only by anticipating policy-induced distribution shift can either of these fairness definitions be systematically improved.

Reintroducing general group-based fairness constraints as formulated in Eq. (2.15), we define the disparity function $\mathcal{V}$ in terms of a group-specific quantity ξ_g that depends

on the policy θ and the distribution $\mathcal{D}$.

$$\begin{aligned}\mathcal{V}(\theta, \mathcal{D}) = \operatorname*{Var}_{g}\left[\xi_g(\theta, \mathcal{D})\right] &\coloneqq \frac{1}{n}\sum_{g=1}^{n}\left\|\xi_g(\theta, \mathcal{D}) - \overline{\xi}(\theta, \mathcal{D})\right\|^2 \\ &= \frac{1}{n}\sum_{g=1}^{n}\left\|\xi_g(\theta, \mathcal{D})\right\|^2 - \left\|\overline{\xi}(\theta, \mathcal{D})\right\|^2,\end{aligned} \tag{4.21}$$

Continuing to use this formulation we define participation parity and qualification rate parity.

Participation Parity is defined by the group-independence of participation rates by combining Eqs. (4.21) and (4.22):

$$\xi_g(\theta, \mathcal{D}) \coloneqq \Pr_{G\sim\mathcal{D}}\Big(G = g\Big). \tag{4.22}$$

Qualification Rate Parity may likewise be defined as the group-independence of qualification rates by combining Eqs. (4.21) and (4.23)

$$\xi_g(\theta, \mathcal{D}) \coloneqq \Pr_{Y,G\sim\mathcal{D}}\Big(Y = 1 \;\Big|\; G = g\Big). \tag{4.23}$$

This measure of disparity is considered by Raab and Liu (2021); Zhang et al. (2020).

Unlike standard measures of unequal treatment subject to ML policy, neither Eq. (4.22) nor Eq. (4.23) depend on the policy (i.e., the variable $\hat{Y}$) at all. Instead, both measure are *inherent to the distribution* and can only be acted on by the policy by inducing changes in the underlying distribution.

By integrating awareness of policy-induced distribution shift into our algorithm development, this chapter has shown that we may treat a strictly wider class of disparity measures than standard, present-normative measures. In the next chapter, we show

that a reinforcement-learning based approach to alignment in dynamical environments can learn to avoid the familiar trap of pursuing short-term fairness metrics only to widen underlying disparities that demand escalating interventions at the expense of utility (Section 4.1.3).

5 Bringing Alignment Online

How can we guarantee alignment for AI systems that interact with complex, multiagent environments that are difficult to model or predict?

In unknown environments, we are forced to *learn* policies (or models of the environment) *online*—i.e., through live, irreversible interactions with the environment. This problem domain is known as "online learning".

The use online learning for socially consequential tasks is fraught with ethical complications: To some degree, online learning amounts to some degree of "live experimentation" on affected populations. A more conservative interpretation of online learning, however, would be "learning from past mistakes". We endorse the latter interpretation and emphasize that the alternative to online learning is *not learning* from one's actions.

In order to consider online learning techniques for realistic use cases, we typically want theoretical *guarantees* about the performance and alignment of the system. While guarantees relating to the primary objective (i.e., $\mathcal{L}$) are standard within this domain, guarantees about *constraints* (i.e., $\mathcal{V}$ in our formulation) remains an active area of research, including "safe reinforcement learning" as discussed in Section 2.4.3.

In this chapter, we formulate the central problem of AI alignment as an online reinforcement learning problem (i.e., a "partially observable Markov decision process" (POMDP)). To address this formulation, we also review an online reinforcement learning technique that provide *simultaneous* guarantees for $\mathcal{L}$ and $\mathcal{V}$ (subject to certain assumptions). We also consider the applicability of more established reinforcement learning (RL) techniques to the problem.

Parts of this chapter are adapted from previously published work by Yin et al. (2023).

5.1 Alignment as an RL Problem

In Section 4.2, we considered a alignment as *control problem*, in which the goal of an algorithmic agent is to drive system dynamics towards a desirable outcome. We saw that such policies could be parameterized to make trade-offs between short-term goals (such as normative present group-fairness targets) and long-term goals (such as reaching inherently equitable system states). Moreover, we saw that the alignment between short-term and long-term could be dependent on the state of the system.

In this section, we reformulate our objectives to account for such trade-offs. In order to limit short-term disparities while also seeking aligned long-term outcomes, we will define *cumulative* notions of loss ($\mathcal{L}$) and disparity ($\mathcal{V}$). In building towards this reformulation of our objectives, we first adapt standard RL concepts to our purpose.

Markovian Dynamics In Section 3.2, we discussed different models of how multiagent systems might dynamically evolve and react to algorithmic policies. One such model, highlighted in Section 3.2.3, is that of Markovian dynamics, in which the state s of the system (from which the distribution $\mathcal{D}$ may be derived) evolves stochastically according to a "transition kernel" $\mathbf{P}$ that depends on the deployed policy.

$$s^t \sim \mathbf{P}(\theta^t, s^{t-1}). \tag{5.1}$$

We stipulate that $\mathbf{P}$ may be initially unknown, but we assume that it does change with time.

5.1.1 States, Policies, and Meta-Policies

In the standard reinforcement learning setting, we consider a policy for stochastically selecting actions in different states. In our current setting, we wish to know the appropriate policy (e.g., for classification or prediction tasks on a population) to deploy given the currently observed state (and distribution). To disambiguate these notions of *policy*, we refer to a "meta-policy" π that stochastically selects the policy θ (h_θ) as a function of the state (and thus distribution) s. This identifies θ with an "action" in the typical RL framing.

$$\theta^t \sim \pi^t(s^t). \tag{5.2}$$

Our goal is to find an appropriate algorithm for refining π over time.

To simplify our setting further, we assume that the state s, evolving according to Eq. (5.1), is fully observable to the policy-maker, and that we identify the state with the joint distribution of variables relevant to the policy h_θ, i.e., $s \equiv \mathcal{D}$. In the notation of Section 3.2.2, where we relate state variables s and the entailed distribution $\mathcal{D}$ according to $\mathcal{D}^t = \mathrm{M}(s^t)$, this assumption is equivalent to setting M equal to the identity, such that

$$s^t \equiv \mathcal{D}^t. \tag{5.3}$$

5.1.2 Value and Quality Functions

The standard objective of reinforcement learning, which we seek to *maximize*, is the expected cumulative value of some *reward* subject to dynamic uncertainties. We will identify reward with negative loss, $-\mathcal{L}$. The difficulty of this problem is that this expectation value must be taken over all possible future *trajectories*, as both θ (governed by π) and s (governed by $\mathbf{P}$) evolve in time.

Given a fixed policy π, this objective is easily expressed using the letter "V", for "value", as in

$$V_{\mathcal{L}}(\pi, s) \coloneqq -\operatorname{E}\left[\sum_{\tau=0}^{H} \gamma^{\tau} \mathcal{L}(\theta^{\tau}, s^{\tau}) \;\middle|\; s^{0} = s\right]. \tag{5.4}$$

The parameter $\gamma \in (0, 1]$ corresponds to an *exponential* discounting, giving multiplicatively less weight to future outcomes compared to proximal outcomes, while the value of H parameterizes a (finite) time-horizon.

The value function V is related to the commonly defined Q function, which evaluates the "quality" of state-action pairs, by the Bellman equation and "dynamic programming":

$$Q_{\mathcal{L}}(\pi, s, \theta) \coloneqq \operatorname*{E}_{s' \sim \mathbf{P}(s,\theta)}\Big[-\mathcal{L}(\theta, s) + \gamma V_{\mathcal{L}}(\pi, s')\Big] \tag{5.5}$$

$$V_{\mathcal{L}}(\pi, s) = \operatorname*{E}_{\theta \sim \pi(s)}\Big[Q_{\mathcal{L}}(\pi, s, \theta)\Big] \tag{5.6}$$

For our purposes, we define Q and V functions for both cumulative loss ($\mathcal{L}$) *and* cumulative disparity ($\mathcal{V}$):

$$Q_{\mathcal{V}}(\pi, s, \theta) \coloneqq \operatorname*{E}_{s' \sim \mathbf{P}(s,\theta)}\Big[-\mathcal{V}(\theta, s) + \gamma V_{\mathcal{V}}(\pi, s')\Big] \tag{5.7}$$

$$V_{\mathcal{V}}(\pi, s) = \operatorname*{E}_{\theta \sim \pi(s)}\Big[Q_{\mathcal{L}}(\pi, s, \theta)\Big] \tag{5.8}$$

5.1.3 The Constrained RL Problem

For a given initial state s^0, we formulate the following RL problem:

$$\operatorname*{maximize}_{\pi} \quad V_{\mathcal{L}}(\pi, s^0) \tag{5.9}$$

$$\text{subject to} \quad V_{\mathcal{V}}(\pi, s^0) \geq \varepsilon. \tag{5.10}$$

Note that the semantics of the optimization problem have changed slightly compared to our earlier optimization problems, because we now seek to *maximize* both V functions (which is the convention in RL), rather minimize them (which is the convention in optimization).

Prob. (5.9) is a *constrained* reinforcement learning problem. In general, such problems are very difficult without restrictive assumptions about the system dynamics and the structure of $\mathcal{L}$ and $\mathcal{V}$. This remains an active area of research.

5.2 Scheduled Lagrangian Regularization

One way to approximately solve Prob. (5.9) is to solve a *Lagrangian relaxation* of the problem. The new, "relaxed" objective may be expressed as

$$\underset{\pi}{\text{maximize}} \quad V_{\mathcal{L}}(\pi, s^0) + \nu V_{\mathcal{V}}(\pi, s^0), \tag{5.11}$$

for some chosen *dual variable* $\nu \geq 0$. In Appendix C, the expression that we seek to extremize in Prob. (5.11) as the Lagrangian (Appendix B).

This new problem is an *unconstrained* optimization problem, which is the bread-and-butter of machine learning: Standard RL techniques will address this problem subject to Markovian dynamics. The only trick is to choose ν appropriately.

5.2.1 Finite Horizon Interpolation

A standard trick for such Lagrangian relaxation is to vary ν according to an increasing "schedule" in time t, effectively requiring increasing strictness with respect to the constraint over time.

We can use this idea by varying the weight assigned to the terms that add to $V_{\mathcal{L}}$ and $V_{\mathcal{V}}$

over time (Eq. (5.4)). For simplicity, we can simply linearly interpolate between $V_{\mathcal{L}}$ and $V_{\mathcal{V}}$ over the finite time horizon H. Let us define the following objective, once more expressed as a target to *minimize*:

$$\underset{\pi}{\text{minimize}} \quad \mathrm{E}\left[\sum_{t=0}^{H}(1-\lambda_t)\mathcal{L}(\theta^t, s^t) + \lambda_t \mathcal{V}(\theta^t, s^t)\right], \tag{5.12}$$

where $\lambda_t = t/H$.

In experiment (Section 5.4), we use an "off-the-shelf" RL algorithm, viz., "Twin-Delayed Deep Deterministic Policy Gradient" (TD3) (Fujimoto et al., 2018) to solve Prob. (5.12) episodically, with an implementation and default parameters provided by the open-source package "Stable Baselines 3" Raffin et al. (2021). We will refer to this implementation as "R-TD3".

5.3 Bounding Regret

The method provided in Section 5.2, while intuitive, does not provide rigorous guarantees regarding the constraint of $V_{\mathcal{V}}$. Part of the problem is that the dual variable ν has a "correct" value in theory, but Section 5.2 does not seek to find it: it merely starts off relatively ignorant of the constraint, then becomes overly conservative with time in the hope that the dynamics have been appropriately "learned" by that point. Can we provide guarantees regarding $V_{\mathcal{V}}$ while learning an appropriate value of ν?

Because we have assumed stochastic dynamics, it is not possible to guarantee constraint satisfaction in Prob. (5.9) over a finite number of online steps. Rather than seeking to solve this problem *exactly*, we can instead develop algorithms that solve this problem approximately—i.e., by finding policies with bounded suboptimally or constraint violation.

Again because the Markovian setting is inherently probabalistic, it is also impossible to guarantee hard limits on suboptimality or constraint violations. Instead, we can only provide probabalistic guarantees for algorithm performance, limiting the degree to which the proposed methods might fail to solve Prob. (5.9).

In this section, we build towards an algorithm that provides probabalistically sublinear bounds on two forms of *regret*, a standard notion in online learning that measures *suboptimiality in hindsight*.

5.3.1 Two Types of Regret

Denote the optimal value of π (defined with respect to $V_{\mathcal{L}}$ alone, ignoring $V_{\mathcal{V}}$ constraints) as $\pi^\star$. We define loss-regret as the suboptimality of $V_{\mathcal{L}}$ with respect to $\pi^\star$ and constraint-regret as the suboptimality of $V_{\mathcal{V}}$ with respect to the maximum allowed value ε:

$$\mathrm{Regret}_{\mathcal{L}}(\pi, s) \coloneqq V_{\mathcal{L}}(\pi^\star, s) - V_{\mathcal{L}}(\pi, s). \tag{5.13}$$

$$\mathrm{Regret}_{\mathcal{V}}(\pi, s) \coloneqq \max\left[0, \varepsilon - V_{\mathcal{V}}(\pi, s)\right]. \tag{5.14}$$

5.3.2 Novel Theoretical Guarantees

With Yin et al. (2023), I helped to establish the first model-free, simulator-free algorithm to provide simultaneous, probabalistic bounds on both $\mathrm{Regret}_{\mathcal{L}}$ and $\mathrm{Regret}_{\mathcal{V}}$ for continuous states (s) and policies (θ). These results rely on several simplifying assumptions (Asms. 5.3.1 to 5.3.3).

Assumption 5.3.1 (Strict Feasibility)**.** Prob. (5.9) is strictly feasible. That is,

$$\exists \pi, \quad \forall s, \quad V_{\mathcal{V}}(\pi, s) > \varepsilon \tag{5.15}$$

Asm. 5.3.1 is a standard assumption necessary that ensures that it is possible to satisfy the constraint of Prob. (5.9) in theory.

Next, generalizing Section 3.2.3, we assume a *known* feature map ϕ for which the state transitions *and* the values of $\mathcal{L}$ and $\mathcal{V}$ are linear.

Assumption 5.3.2 (Linear MDP)**.** Given a feature map ϕ, there exist a family of vectors $\mu[s']$ mapped by s', as well as vectors $\mu[\mathcal{L}]$ and $\mu[\mathcal{V}]$, such that

$$\Pr(s' \mid s, \theta) = \big\langle \phi(s, \theta), \mu[s'] \big\rangle. \tag{5.16}$$

$$\mathcal{L}(s, \theta) = \big\langle \phi(s, \theta), \mu[\mathcal{L}] \big\rangle. \tag{5.17}$$

$$\mathcal{V}(s, \theta) = \big\langle \phi(s, \theta), \mu[\mathcal{V}] \big\rangle. \tag{5.18}$$

The assumption of the *existence* of ϕ is no less justified than Eq. (3.12) when s belongs to the infinite-dimensional Hilbert space of all possible observables of a system. In practical use, s must be finite-dimensional, and we can only hope to approximate the conditions assumed by Asm. 5.3.2. The assumption that ϕ is *known* still qualifies L-UCBFair as an online algorithm, insofar as the μ vectors are not known.

Assumption 5.3.3 (Lipschitz Policies)**.** ϕ is Lipschitz continuous in θ. That is, there exists a value $\alpha > 0$, such that

$$\|\phi(s, \theta) - \phi(s, \theta')\|_2 \leq \alpha \|\theta - \theta'\|_2. \tag{5.19}$$

Asm. 5.3.3 is useful for justifying an approximation that discretizes the space of policies represented by θ.

L-UCB Fair

L-UCBFair, or "LSVI-UCB for Fairness" (Yin et al., 2023) is based on a Least-Squares Value Iteration (LSVI) with an optimistic Upper-Confidence Bound (UCB), as in LSVI-UCB (Jin et al., 2020). This algorithm operates on *episodes* of length H, pre-committing to a policy π_k for the entire episode before updating to policy π_{k+1} for the next episode in the online setting.

During each H-step episode, L-UCBFair maintains constant estimates for $Q_{\mathcal{L}}(\pi^\star, \cdot, \cdot)$ and $Q_{\mathcal{V}}(\pi^\star, \cdot, \cdot)$ (Section 5.1.2). These estimates are achieved, as in Jin et al. (2020), by assuming a linear form analogous to and (justified by) Eq. (5.17), parameterized by a learned weight vector $\mathbf{w}$:

$$Q_{\mathcal{L}}(\pi^\star, s, \theta) \approx \big\langle \phi(s, \theta), \mathbf{w} \big\rangle. \tag{5.20}$$

L-UCBFair also maintains estimates for Lagrangian dual variable ν (Prob. (5.11)) for which the objective values of Prob. (5.9) and Prob. (5.11) achieved by $\pi^\star$ (Section 5.3.1) coincide.

To deal with the fact that θ-space is continuous (and potentially quite large), we rely on the Lipschitz assumption (Asm. 5.3.3) to justify discretizing this space according to a Voronoi scheme with M regions of maximum radius ϵ_I. Policies are selected stochastically, as a function of state, according to a distribution over partitions, then uniformly at random within each partition. [Bounded Regrets (Yin et al., 2023)] With probability p, over K episodes, L-UCBFair achieves

$$\mathrm{Regret}_{\mathcal{L}}, \mathrm{Regret}_{\mathcal{V}} \in \mathcal{O}\bigg(\log\left(\frac{HdK}{1-p}\right) H^2 \sqrt{d^3 K} \bigg). \tag{5.21}$$

We provide additional details and discuss the proof of Section 5.3.2 in Appendix C.

5.4 Experiments

In this section, we evaluate a Lagrangian-relaxation approach to Prob. (5.9) (i.e., R-TD3 Section 5.2) to the Voronoi-LSVI-UCB method (i.e., L-UCBFair) and present-normative baselines. Our central hypothesis is that an RL formulation of alignment should allow a model to learn to sacrifice short-term utility and present-normative constraints in order to drive the system towards more desirable equilibria, *where beneficial*. Moreover, we hope to show that this can be achieve *online*.

5.4.1 Revisting a Familiar Setting

We evaluate our proposed methods in semi-synthetic experiments, which model a multiagent environment using the now-familiar model of Section 3.5: We model evolving group qualification rates according to the replicator equation Eqs. (3.32) to (3.34) and consider a binary classification task (Section 2.1.4) using a threshold classifier on scalar features.

Data and Task We incorporate real-world data into our numerical experiments, provided by a standard "Census Income" or "Adult" data set (for which the canonical task is the prediction of income from 1993 census information (Dua and Graff, 2017)).

In Fig. 5.1, we compare L-UCBFair and R-TD3 to repeated risk minimization (RRM; Section 3.3.1).

Loss For this task, we define loss $\mathcal{L}$ as the opposite of the true-positive classification rate.

$$\mathcal{L}(\theta, s) = -\Pr\Big(\hat{Y}{=}1 \mid Y{=}1\Big). \tag{5.22}$$

Comparing this notion of loss to the model of classifier utility defined in Eq. (3.29), this is equivalent to choosing a classifier utility matrix $C_{y,\hat{y}}$ equal to

$$\begin{bmatrix} C_{-1,-1}=0 & C_{-1,1}=0 \\ C_{1,-1}=0 & C_{1,1}=1 \end{bmatrix}. \tag{5.23}$$

Disparity We define disparity $\mathcal{V}$ as twice the value defined by (Eqs. (2.15) and (2.18)).

$$\mathcal{V}(\theta, s) = \frac{1}{2}\Big\| \Pr_{\theta,s}(\hat{Y}{=}1 \mid G{=}1) - \Pr_{\theta,s}(\hat{Y}{=}1 \mid G{=}2)\Big\|^2 \tag{5.24}$$

$$= 2 \operatorname*{Var}_{g}\Big[\Pr_{\theta,s}(\hat{Y}{=}1 \mid G{=}g)\Big]. \tag{5.25}$$

Preprocessing To map the classification task to one over scalar features while incorporating real-world data in multiple dimensions, we require a preprocessing pipeline. For each simulated state s_g, we perform a preprocessing step re-weights sample-losses on an initial regression task (to mimic counter-factual group qualification rates s_g): The outputs of the trained regressor, corresponding to predicted probabilities of qualification, are used as synthetic feature values to generate semi-synthetic *conditional* distributions $\Pr(X \mid Y, G)$. The full, semi-synthetic distribution reported to the classifier is a product of these conditional distributions and the modelled group qualification rates.

The advantage of using a logistic regression pre-processing step is the ability to satisfy Eq. (3.27) (which posits that X is a "well-behaved" feature on which the likelihood of

qualification is monotonically dependent). This justifies the use of threshold policies.

An additional pre-processing step that we use, in order to approximately satisfy Asm. 5.3.2, is to train a neural-network to approximate the feature-map ϕ *offline*.

RRM Baseline: "Myopic-Fair" We define the "Myopic-fair" baseline as the policy that optimizes, as a function of state s the regularized objective:

$$\underset{\theta}{\text{minimize}} \quad \mathcal{L}(\theta, s) + \mathcal{V}(\theta, s). \tag{5.26}$$

5.4.2 Results

In Fig. 5.1, we compare the myopic-fair baseline to L-UCBFair and R-TD3. In this example setting, the myopic baseline exhibits qualitatively similar behavior to the baseline in Sections 3.5.2 and 4.2.3 subject to demographic parity: the qualification rates of both groups diverge as a result of the present normative intervention.

In contrast to the myopic baseline, both L-UCBFair and R-TD3 are able do drive the system towards more desirable states characterized by higher group qualification rates that converge closer to each other. Importantly, this result is *aligned* with the *cumulative* definition of loss: lower mean cumulative loss, as depicted in the lower half of the figure, is achieved by L-UCBFair and R-TD3 than the baseline policy.

This experiment confirms that that an RL formulation of alignment can allow a model to learn to sacrifice short-term utility and present-normative constraints (which are greedily pursued by the baseline policy) in order to drive the system towards more desirable equilibria as defined by cumulative loss.

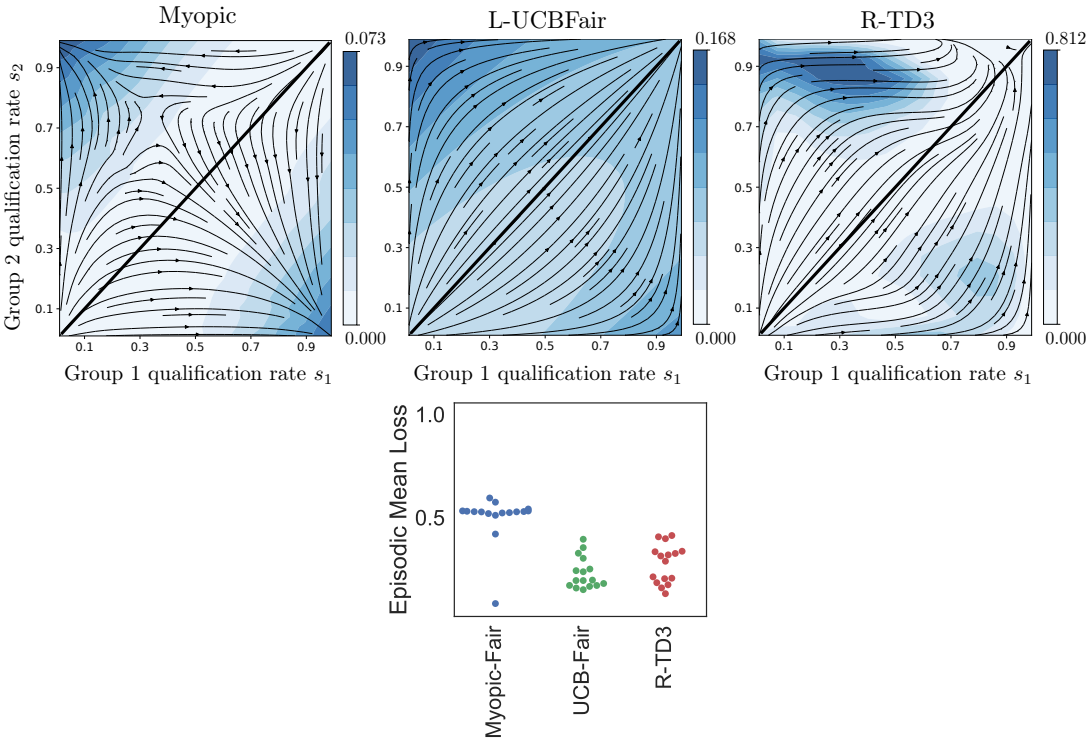

Figure 5.1: (Above) Phase diagrams (introduced in Section 3.5.2) for three algorithms in the same setting: The "Myopic-fair" baseline, L-UCBFair, and R-TD3. (Below) A comparison of cumulative losses between the algorithms achieved from a uniform distribution over initial states. Note that the color scales on the upper diagram are not equal and represent violations of demographic parity.

6 Conclusion

In this dissertation, we have shown that multiagent dynamics, especially in the form of policy-induced distribution shift, must be accounted for if we wish to intervene on AI systems and align their behaviors with human goals, values, and ethical principles. In particular, we have repeatedly highlighted the potential failure modes of "present normative interventions": Solving the alignment will not be accomplished by finding the right optimization problem to solve as an ML task as long as dynamical realities are ignored.

To deal with realities of multiagent dynamics that surround AI systems in socially consequential positions, we have explored problem formulations that rely on constrained optimization, dynamical stability and control, and online reinforcement learning. With each problem formulation, we have explored increasingly more complex models of the dynamics of the multiagent system until arriving at, and incrementally extending, the limit of what current theory can prove as far as safety guarantees are concerned.

6.1 Closing Remarks

In this dissertation, we have considered only a tiny subset of the possible multiagent dynamics that ML systems may interact with, and we have remained largely agnostic to the plethora of mathematical definitions that have been proposed for "alignment". Our contribution might be described an initial "recipe book" of methods and strategies for dealing with potential "alignment problems" on a case-by-case basis, depending on the problem and the anticipated dynamical context.

To the extent that the alignment problem, writ large, is self-consistent or well-defined, it may be the case that it is not really an ML problem: In truth, it is not clear that human institutions have "solved" the alignment problem for *humans* (or for our institutions).

Society continues to grapple with what "human goals, values, and ethical principles" really are, and the historical record indicates that such ideals are culturally constructed, adapted to circumstance, changing over time. To the extent that we agree on common values, codified in law or reinforced as norms, their enforcement mechanisms do not depend on mathematical expression, but on power and politics.

Even if we knew *how* to regulate and control AI, its regulation *in practice* will depend on political force and foresight. My hope is that this dissertation at least frames the problem in terms of dynamical systems and control, multiagent systems, and game-theoretic incentives. As we have shown, when we *do* account for multiagent system dynamics and correctly aligned incentives, the powerful tools we have built for numerical optimization have the potential to realize the long-term goals we give them.

A Replicator Dynamics

"Replicator dynamics" refers to a model for the population dynamics of competing and *replicating* (i.e., reproducing, multiplying) *types* (e.g. species, genes, or phenotypes). Despite originating as a model for biological systems, replicator dynamics also provides a foundational model for evolutionary game theory, where types can imply competing strategies, hypotheses, or norms.

In this appendix, we discuss the replicator dynamics as described by the replicator equation and briefly discuss an original result connecting the replicator equation to the mathematics of gradient descent.

Parts of this appendix are adapted from previously published work by Raab et al. (2022).

A.1 The Replicator Equation

The replicator equation has been applied to game theory (Hofbauer et al., 1998; Sandholm, 2010; Cressman and Tao, 2014; Friedman and Sinervo, 2016), economics (Friedman, 1991), and machine learning (Hennes et al., 2019).

In the replicator equation, the absolute *fitness* (in this paper, the negative *loss* $\mathscr{L}$) of hypotheses $h \in \mathcal{H}$ is identified with its rate of *replication*: exponential growth (or decline) in a population where different hypotheses compete for relative frequency $\rho(h) \in [0, 1]$. For probability distributions over hypothesis space $\mathcal{H}$, this equation induces *replicator dynamics*, selecting hypotheses with lower than average loss.

In continuous time, the replicator equation is

$$\dot{\rho}(h) = \rho(h)\Big[\overline{\mathscr{L}}_\rho - \mathscr{L}(h)\Big], \quad \text{where} \quad \overline{\mathscr{L}}_\rho \coloneqq \sum_h \rho(h)\mathscr{L}(h), \quad \sum_h \rho(h) = 1. \tag{A.1}$$

Although we allow ourselves to omit time-indexing for ρ and $\overline{\mathscr{L}}$ in Eq. (A.1), these quantities are time-varying.

In discrete time, the replicator equation is

$$\rho(h)^{t+1} \coloneqq \rho^t(h)\frac{W^t(h)}{\overline{W}^t_\rho} \quad \text{where} \quad \overline{W}^t \coloneqq \sum_h \rho^t(h)W^t(h), \tag{A.2}$$

where these two equations are related by

$$\log\left[W(h)\right] = -\mathscr{L}(h).$$

A.2 Conjugate Natural Selection

In Raab et al. (2022), I prove that the replicator equation, is *optimally approximated* by Fisher-Rao natural gradient descent (FR-NGD).

FR-NGD is an update given in continuous time by

$$\dot{\theta} = -\nabla_\theta \mathscr{L}(h_\theta)^\top F(\theta)^\dagger, \tag{A.3}$$

where $\nabla\mathscr{L}(h_\theta)$ is the normal loss gradient with respect to θ and F is the "Fisher"—the Hessian of a Kullback-Leibler divergence in ρ-space taken with respect to θ. That is,

$$F[x] = \frac{\partial^2}{\partial x^2} D_{\mathrm{KL}}\Big(\rho_x \parallel \rho^t\Big). \tag{A.4}$$

When x is a vector with components (e.g. x^i, x^j), then F is matrix with entries given by

$$F_{ij}[x] = \operatorname*{Cov}_{H\sim\rho_x}\left[\frac{\partial}{\partial x^i}\log\rho_x(H), \frac{\partial}{\partial x^j}\log\rho_x(H)\right]. \tag{A.5}$$

The dagger (†) in Eq. (A.3) indicates a Moore-Penrose pseudoinverse.

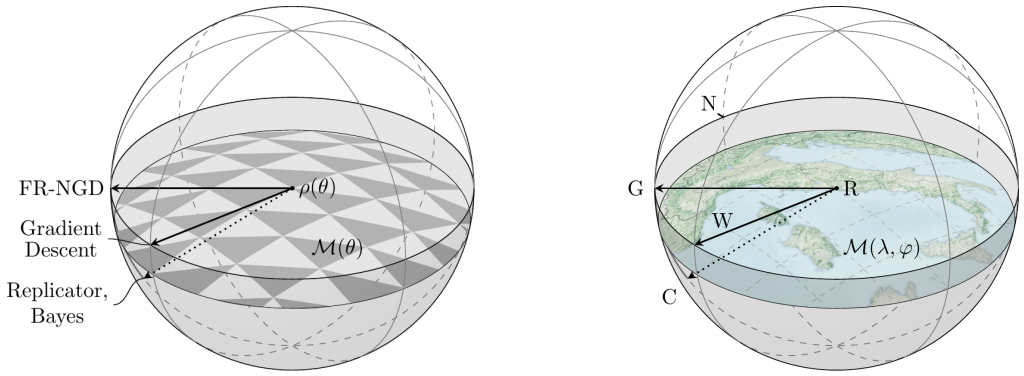

Figure A.1: We visualize an analogy for how the use of the metric $F^\dagger$ in Eq. (A.3) can cause FR-NGD to differ from regular gradient descent by modifying local relative distances: The direction of travel from Rome (point R) that most rapidly decreases one's distance from Chicago, when measured in the Euclidean space of latitude-longitude pairs (λ, φ), is nearly due west (vector W), because Rome and Chicago have nearly the same latitude λ. Performing gradient descent with an implicit Euclidean metric for parameter space is similarly naive. Vector C is tangent to the true shortest path in physical space: north-west at an angle of nearly 35 degrees downwards. Like the update given by the replicator equation, this direction may not be tangent to the manifold $\mathcal{M}$. Constrained to $\mathcal{M}$, the optimal approximation of the direction of C is its projection, G: north-west, tangent to the surface, and tangent to the geodesic from Rome to Chicago on the surface of the sphere. Map Data Credit: NASA Visible Earth.

In Fig. A.1, we provide an analogy for how $F^\dagger$ affects parameter updates in Eq. (A.3), in contrast to the identity matrix (standard gradient descent): Both a flat map and a globe "warp" our perception of local distances and angles, distorting our perception of the shortest route (closeness to the replicator equation) between two points on Earth (the θ-manifold $\mathcal{M}$ of distributions ρ). Intuitively, $F^\dagger$ warps the space to equate distance with the information about ρ implied by changes in θ.

A.2.1 Optimal Approximation

Theorem A.2.1 (Conjugate Natural Selection)**.** *For any twice-differentiable, parametric probability distribution ρ_θ over hypotheses $h \in \mathcal{H}$, FR-NGD minimizes the* Fisher

Divergence *from the replicator equation, denoted $\mathcal{E}$:*

$$\mathcal{E} \coloneqq \frac{1}{2}\int_{\mathcal{H}}\int_{\mathcal{H}}\Big(\dot{\rho}_\theta(h)-\dot{\rho}^\star(h)\Big)F^\dagger[\rho](h,k)\Big(\dot{\rho}_\theta(k)-\dot{\rho}^\star(k)\Big)\,\mathrm{d}h\,\mathrm{d}k\,. \tag{A.6}$$

$\mathcal{E}$ measures how different, on average, $\dot{\rho_\theta}(h)$ (i.e., the update to the distribution prescribed by FR-NGD) is from $\dot{\rho}^\star(h)$, the update prescribed by the replicator equation. This result, which I term "conjugate natural selection" is proven in Raab et al. (2022) along with the correspondence of FR-NGD to continuous Bayesian inference.

A.2.2 Use for Learning

We may use conjugate natural selection to approximate evolution of *distributions over solutions* to highly-convex problems. Such an idea is represented by Fig. A.2: We use FR-NGD to update the parameters of a Gaussian distribution over candidate minimizers of the Rastrigin function,

$$\mathscr{L}(h_x,h_y)=20+h_x^2+h_y^2-10\cos(2\pi h_x)-10\cos(2\pi h_y), \tag{A.7}$$

The surface of this function is depicted in the upper-right pane of Fig. A.2, clearly showing its lack of convexity.

At each time step, $N{=}40$ hypotheses h are sampled from ρ^t and the loss for each h is calculated, yielding a Monte Carlo estimate of the loss gradient.

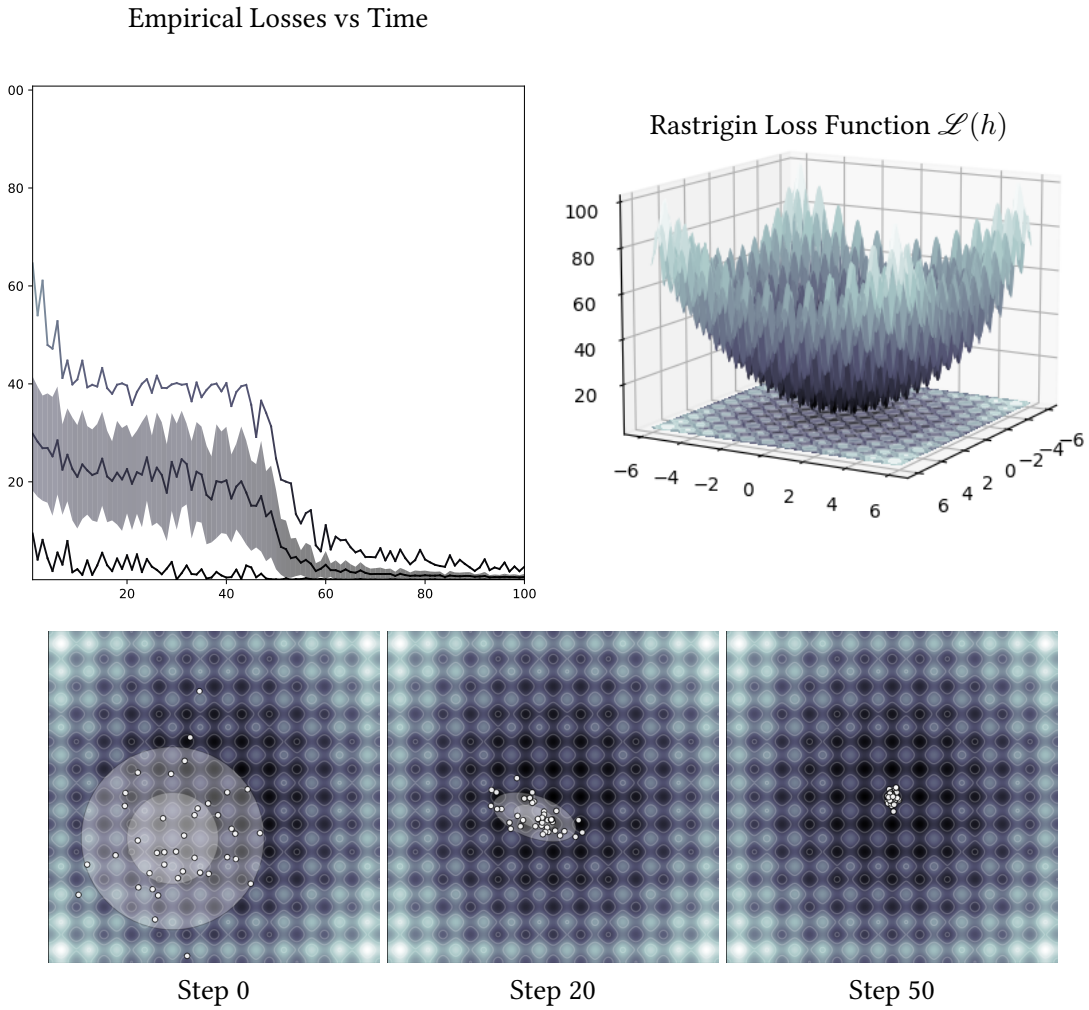

Figure A.2: In the upper left, we plot the mean, standard deviation, and extremal empirical losses for the learned distribution over 100 time steps. In the upper right, the loss function is visualized as a surface over the domain $[-6, 6] \times [-6, 6]$. In the lower three panes, we represent time steps 0, 20, and 50 of the evolution: The Rastrigin function is visualized with shading and highlighted level sets; the sampled hypothesis are represented by white dots; and the 1- and 2-σ ellipses for the evolving Gaussian distribution ρ are shaded white with partial transparency. The distribution is initialized with mean at $[-1.5, -1.5]$ and identity covariance, and we use a constant learning rate of $1e^{-3}$ for the Euler update.

B Constrained Optimization

Throughout this dissertation, we consider constrained optimization problems of the form

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \mathcal{L}(\theta) \\ \text{subject to} \quad & \mathcal{V}(\theta) \leq 0. \end{aligned} \tag{B.1}$$

B.1 The Lagrangian

The theory for solving this sort of problem is well-articulated by introducing the concept of the *Lagrangian*, Λ, which is a function of the decision variable θ (also referred to as the "primal" variable) and a new variable ν that is referred to as a "multiplier" or "dual-variable".

$$\Lambda(\theta, \nu) \coloneqq \mathcal{L}(\theta) + \nu \mathcal{V}(\theta). \tag{B.2}$$

Prob. (B.1) may be re-expressed as

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \left[\max_{\nu} \left[\Lambda(\theta, \nu) \right] \right] \\ \text{subject to} \quad & \nu > 0. \end{aligned} \tag{B.3}$$

Intuitively, ν may be chosen *adversarially* (i.e., after θ is fixed). The problem, therefore, requires the choice of some θ such that the constraint on $\mathcal{V}(\theta)$ is satisfied: If it is not, the objective can be made unboundedly large by some value of ν). Within the "feasible set" of θ values that satisfy $\mathcal{V}(\theta) \leq 0$, we seek to select the one for which $\mathcal{L}(\theta)$ is minimized.

B.2 Primal-Dual Methods

The standard approach to numerically solving problems like Eq. (B.3) in practice is to iteratively refine estimates for the optimal values of θ and ν: a generalization of the idea introduced in Section 2.1.3.

A basic algorithm in the class of such "primal-dual" algorithms is one that relies on gradient descent to update both θ and ν. Generalizing Eq. (2.5) with a Euclidean metrics, at each time-step, we solve

$$\theta^{t+1} = \arg\min_{\theta} \left[\left\langle \nabla_\theta \Lambda(\theta^t, \nu^t), \theta - \theta^t \right\rangle + \frac{1}{2\eta} \|\theta - \theta^t\|^2 \right], \tag{B.4}$$

$$\nu^{t+1} = \arg\max_{\nu \geq 0} \left[(\nu - \nu^t) \frac{\partial}{\partial \nu} \Lambda(\theta^{t+1}, \nu^t) + \frac{1}{2\sigma} \|\nu - \nu^t\|^2 \right], \tag{B.5}$$

for step-sizes η and σ.

Relabeling time-indices to consider the update to ν update *before* θ, this results in the update rules

$$\nu^{t+1} = \max \left[0, \nu^t + \sigma \mathcal{V}(\theta^t) \right]. \tag{B.6}$$

$$\theta^{t+1} = \theta^t - \eta \left(\nabla_\theta \mathcal{L}(\theta^t) + \nu^{t+1} \nabla_\theta \mathcal{V}(\theta^t) \right). \tag{B.7}$$

B.3 Deriving Fletcher's Method

Parts of this section are adapted from previously published work by Raab et al. (2024).

Fletcher's penalty method (Fletcher, 1973) is outlined by Conn et al. (2000, Sec. 14.6) as a surrogate objective or "merit" function Φ that we may seek to minimize in order to

solve (Prob. (B.1)):

$$\Phi(\theta) = \mathcal{L}(\theta) + \max\left(0, \frac{\sigma\mathcal{V}(\theta) - \left\langle \nabla\mathcal{L}(\theta), \nabla\mathcal{V}(\theta)\right\rangle}{\left\langle \nabla\mathcal{V}(\theta), \nabla\mathcal{V}(\theta)\right\rangle}\right)\mathcal{V}(\theta). \tag{B.8}$$

The standard approach is to minimize $\Phi(\theta)$, but the potential non-differentiability of the coefficient of $\mathcal{V}(\theta)$ can be problematic (Conn et al., 2000).

From this point on, we allow the shorthands

$$\mathcal{L}^t \equiv \mathcal{L}(\theta^t), \quad \nabla\mathcal{L}^t \equiv \nabla_\theta\mathcal{L}(\theta^t). \tag{B.9}$$

$$\mathcal{V}^t \equiv \mathcal{V}(\theta^t), \quad \nabla\mathcal{V}^t \equiv \nabla_\theta\mathcal{V}(\theta^t). \tag{B.10}$$

To avoid this potential non-differentiability, let us interpret the coefficient of $\mathcal{V}(\theta)$ in Eq. (B.8) as an estimate for the optimal dual variable ν resulting from a primal-dual method with a non-standard regularization term. That is, replace the Euclidean update penalty for $\|\nu - \nu^t\|^2$ in Eq. (B.5) with the update penalty for θ, expressed as in Eq. (B.7): $\|\theta - \theta^t\|^2 = \eta^2\|\nabla\mathcal{L} + \nu\nabla\mathcal{V}\|^2$. Additionally, approximate

$$\frac{\partial}{\partial\nu}\Lambda(\theta^{t+1}, \nu^t) \approx \mathcal{V}^t + \left\langle \nabla\mathcal{V}^t, \theta^{t+1} - \theta^t\right\rangle. \tag{B.11}$$

Then

$$\nu^{t+1} = \arg\max_{\nu\geq 0}\left[\nu\left(\mathcal{V}^t + \left\langle \nabla\mathcal{V}^t, \theta^{t-1} - \theta^t\right\rangle\right) + \frac{\eta^2}{2\sigma}\|\nabla\mathcal{L}^t + \nu\nabla\mathcal{V}^t\|^2\right] \tag{B.12}$$

$$= \max\left[0, \frac{\frac{\sigma}{\eta^2}\left(\mathcal{V}^t + \left\langle \theta - \theta^t, \nabla\mathcal{V}^t\right\rangle\right) - \left\langle \nabla\mathcal{L}^t, \nabla\mathcal{V}^t\right\rangle}{\left\langle \nabla\mathcal{V}^t, \nabla\mathcal{V}^t\right\rangle}\right]. \tag{B.13}$$

This has the same form as the coefficient in the merit function Eq. (B.8). To use Fletcher's penalty method implicitly, we can use Eq. (B.12) in place of Eq. (B.6) and

iterate Eq. (B.7). Notably, because Eq. (B.12) does not depend on ν^t, we do not require a separate variable to track as algorithm state.

B.4 Generalizing Fletcher's Method

Combining Eq. (B.12) and Eq. (B.7) as prescribed above we find that

$$\left\langle \theta - \theta^t, -\nabla \mathcal{V}^t \right\rangle = \max \left[\eta \left\langle \nabla \mathcal{L}^t, \nabla \mathcal{V}^t \right\rangle, \frac{\sigma}{\eta + \sigma} \mathcal{V}^t \right]. \tag{B.14}$$

It follows that

$$\left\langle \theta - \theta^t, -\nabla \mathcal{V}^t \right\rangle \geq \alpha \mathcal{V}^t. \tag{B.15}$$

for $\alpha = \sigma/(\eta + \sigma)$.

This is precisely the constraint for the CPG method proposed in Section 4.1. As shown in Raab et al. (2024) and stated in Thm. 4.1.1, this constraint is sufficient for solving Prob. (B.1) while allowing a formulation that admits additional convex constraints of the form $\theta \in \mathcal{A}$.

$$\begin{aligned} \theta^{t+1} = \underset{\theta \in \mathcal{A}}{\arg\min} \quad & \left\langle \theta, \nabla \mathcal{L}^t \right\rangle + \frac{1}{2\eta} (\theta - \theta^t)^2. \\ \text{subject to} \quad & \left\langle \theta - \theta^t, -\nabla \mathcal{V}^t \right\rangle \geq \alpha \mathcal{V}^t. \end{aligned} \tag{CPG}$$

C Supplementary Proofs

In this appendix, we provide proofs for theorems stated as such in the main text of this dissertation. These proofs are adapted from my original work as published in Raab and Liu (2021); Chen et al. (2022); Yin et al. (2023); Raab et al. (2024).

Proof of Thm. 3.6.1

Theorem 3.6.1 (Lipschitz Upper Bound for a Curve)**.** *Given an element-wise Lipschitz bound for* $\mathbf{F}$ *along any curve* C *with endpoints* 0 *and* $\mathbf{B}$*, i.e. when there exists some finite* $\mathbf{L} \succeq 0$ *such that*

$$\forall \mathbf{b} \in C, \quad \left|\mathbf{F}(\mathbf{b})\right| \preceq \mathbf{L}, \tag{3.44}$$

we may conclude

$$\mathcal{V}^{\star}_{\mathbf{B}}(\theta, \mathcal{S}) = \mathcal{V}(\theta, \mathcal{S}) - \int_C \left\langle \mathbf{F}(\mathbf{b}), \mathrm{d}\mathbf{b} \right\rangle \tag{3.45a}$$

$$\leq \mathcal{V}(\theta, \mathcal{S}) + \left\langle \mathbf{L}, \mathbf{B} \right\rangle \tag{3.45b}$$

Proof. The equality

$$\mathcal{V}^{\star}_{\mathbf{B}}(\theta, \mathcal{S}) = \mathcal{V}(\theta, \mathcal{S}) - \int_C \left\langle \mathbf{F}(\mathbf{b}), \mathrm{d}\mathbf{b} \right\rangle$$

is a given by Eq. (3.43) as consequence of the definition of the conservative vector field (Eq. (3.41)):

$$\mathbf{F}(\mathbf{b}) \coloneqq -\nabla_{\mathbf{b}} \mathcal{V}^{\star}_{\mathbf{b}}.$$

The desired bound follows by the linearity of integration and the Lipschitz bound $\mathbf{F}(\mathbf{b}) \preceq \mathbf{L}$:

$$|f(x)| \leq y \implies \int_C f(x)\,\mathrm{d}x \leq \int_C y\,\mathrm{d}x\,.$$

Proof of Lem. 3.6.2

Lemma 3.6.2 (Disparity Rates). *The supremal disparity $\mathcal{V}^\star_{\mathbf{B}}$ is subject to a Lipschitz condition given by*

$$\forall g, \quad \frac{\partial}{\partial b_g}\mathcal{V}^\star_{\mathbf{b}}(\theta, \mathcal{S}) \leq L_g = (|\mathcal{G}| - 1)\Big|\beta_g^+ - \beta_g^-\Big|. \tag{3.54}$$

Proof. This is a direct result of differentiation and the label-shift assumption. First, as explained in Section 3.6.2, note that

$$\Pr_{\mathcal{T}}(\hat{Y}{=}1 \mid G{=}g) = \Pr_{\mathcal{T}}(Y{=}1 \mid G{=}g)\beta_g^+ + \Pr_{\mathcal{T}}(Y{=}{-}1 \mid G{=}g)\beta_g^- \tag{C.1a}$$

$$= s_g\beta_g^+ + (1 - s_g)\beta_g^-. \tag{C.1b}$$

where we recall (Eqs. (3.47), (3.49), and (3.50))

$$\beta_g^+(\theta, \mathcal{T}) \coloneqq \Pr_{\theta,\mathcal{T}}(\hat{Y}{=}1 \mid Y{=}1, G{=}g);$$

$$\beta_g^-(\theta, \mathcal{T}) \coloneqq \Pr_{\theta,\mathcal{T}}(\hat{Y}{=}1 \mid Y{=}0, G{=}g);$$

$$s_g \coloneqq \Pr(Y{=}1 \mid G{=}g).$$

Importantly, β_g^+ and β_g^- are constant subject to label shift, which confines distribution shift to the conditional distribution $\Pr(Y{=}1 \mid G)$ (Section 3.1.1). It follows that

$$\frac{\partial}{\partial s_g}\Pr_{\mathcal{T}}(\hat{Y}{=}1 \mid G{=}g) = \big(\beta_g^+ - \beta_g^-\big). \tag{C.2}$$

Next, having defined (Eq. (3.48))

$$D_g(\mathcal{S} \parallel \mathcal{T}) \coloneqq \big|s_g(\mathcal{S}) - s_g(\mathcal{T})\big|,$$

we consider that $\frac{\partial}{\partial b_g}\mathcal{V}^\star_{\mathbf{b}}$, where (Def. 3.6.3 and Eq. (3.53))

$$\forall \mathbf{b} \succeq 0, \quad \mathcal{V}^\star_{\mathbf{b}}(\theta, \mathcal{S}) \coloneqq \sup_{\mathbf{D}(\mathcal{T}\|\mathcal{S}) \preceq \mathbf{b}} \mathcal{V}(\theta, \mathcal{T}),$$
$$\mathcal{V}(\theta, \mathcal{T}) \coloneqq \sum_{g,h \in \mathcal{G}} \Big| \Pr_{\mathcal{T}}(\hat{Y}{=}1 \mid G{=}g) - \Pr_{\mathcal{T}}(\hat{Y}{=}1 \mid G{=}h) \Big|,$$

is bounded by the following:

$$\frac{\partial}{\partial b_g}\mathcal{V}^\star_{\mathbf{b}} = \frac{\partial}{\partial s_g} \sum_{g,h \in \mathcal{G}} \Big| \Pr_{\mathcal{T}}(\hat{Y}{=}1 \mid G{=}g) - \Pr_{\mathcal{T}}(\hat{Y}{=}1 \mid G{=}h) \Big|. \tag{C.3}$$
$$\leq (|\mathcal{G}| - 1) \Big| \frac{\partial}{\partial s_g} \Pr_{\mathcal{T}}(\hat{Y}{=}1 \mid G{=}g) \Big|. \tag{C.4}$$
$$\leq (|\mathcal{G}| - 1) \big| \beta_g^+ - \beta_g^- \big|, \tag{C.5}$$

where we have used Eq. (C.2) to infer Eq. (C.5), thus deriving claimed vector of Lipschitz constants $\mathbf{L}$:

$$\forall g, \quad \frac{\partial}{\partial b_g}\mathcal{V}^\star_{\mathbf{b}}(\theta, \mathcal{S}) \leq L_g = (|\mathcal{G}| - 1) \Big| \beta_g^+ - \beta_g^- \Big|.$$

Proof of Thm. 3.6.3

Theorem 3.6.3 (Demographic Parity under Label Shift)**.** *For demographic parity (Eq.* (3.53) *subject to label-shift (Eq.* (3.51)*),*

$$\mathcal{V}(\theta, \mathcal{T}) - \mathcal{V}(\theta, \mathcal{S}) \leq (|\mathcal{G}| - 1) \sum_g \Big| s_g(\mathcal{S}) - s_g(\mathcal{T}) \Big| \Big| \beta_g^+ - \beta_g^- \Big|. \tag{3.55}$$

Proof. This is a direct result of combining Thm. 3.6.1 and Lem. 3.6.2. The former

gives us

$$\left(\left|\nabla_{\mathbf{b}}\mathcal{V}^{\star}_{\mathbf{b}}\right| \preceq \mathbf{L}\right) \quad \Longrightarrow \quad \left(\mathcal{V}^{\star}_{\mathbf{B}}(\theta,\mathcal{S}) \leq \mathcal{V}(\theta,\mathcal{S}) + \left\langle \mathbf{L},\mathbf{B}\right\rangle\right), \tag{C.6}$$

while the latter implies

$$\left|\frac{\partial}{\partial b_g}\mathcal{V}^{\star}_{\mathbf{b}}\right| \preceq (|\mathcal{G}|-1)\left|\beta_g^+ - \beta_g^-\right| = \mathbf{L}. \tag{C.7}$$

while we define B_g as $\left|s_g(\mathcal{S}) - s_g(\mathcal{T})\right|$(Eq. (3.48)).

By modus ponens,

$$\mathcal{V}(\theta,\mathcal{T}) \leq \mathcal{V}(\theta,\mathcal{S}) + (|\mathcal{G}|-1)\sum_g \left|s_g(\mathcal{S}) - s_g(\mathcal{T})\right|\left|\beta_g^+ - \beta_g^-\right|.$$

Proof of Thm. 4.1.1

To prove Thm. 4.1.1, we first recall our assumptions and then establish three intermediate lemmas (Lems. C.1 to C.3). This proof reiterates the proof provided in Raab et al. (2024), up to a relabelling of variables.

Assumption 4.1.1 (Gradients of Deployed Policy)**.** At each time t, the policy-maker is able to observe $\nabla\mathcal{L}^t$ and $\nabla\mathcal{V}^t$, i.e., the policy-maker has knowledge of the first-order dependence of $\mathcal{L}$ and $\mathcal{V}$ on θ at the currently deployed policy θ^t.

Assumption 4.1.3 (Properties of Disparity)**.** $\mathcal{V}$ is an invex function; that is, every critical point of $\mathcal{V}$ is a global minimum.

Assumption 4.1.2 (Feasibility)**.** The fairness constraint is feasible. That is, $\exists\theta^{\star} \in \mathcal{A}$ such that $\mathcal{V}(\theta^{\star}) \leq 0$. Furthermore, the subproblem in Eq. (CPG) is feasible at each time step t.

In addition to recalling assumptions Asms. 4.1.1 to 4.1.3, we introduce an additional assumption, (Asm. C.1)

Assumption C.1 (Convex Achievable Losses)**.** Independent of the distribution of participating agents, at each time t, the firm is able to select from a constant, convex set of losses $\mathcal{A}$.

Intuitively, the first of these assumptions allows us to invoke gradient methods, while the last three amount to guarantees that:

- Gradient flow of $\mathcal{V}$ constrained to $\mathcal{A}$ converges to feasible $H \leq 0$, and
- Gradient flow of $\mathcal{L}$ constrained to $\mathcal{A}$ and constrained to $H \leq 0$ converges to a local minimum of $\mathcal{L}$.

By assuming sufficiently small step size η, our algorithm effectively performs the necessary gradient flow.

Next, we recall the proposed method given by Eq. (CPG), for parameters $\eta, \sigma > 0$ and $\alpha = \eta\sigma$. Note that relationship between α, η, and σ differs from the relationship we use to explain a relationship to Fletcher's method in Appendix B.

$$\begin{aligned} \theta^{t+1} = \underset{\theta \in \mathcal{A}}{\arg\min} \quad & \left\langle \theta, \nabla \mathcal{L}^t \right\rangle + \frac{1}{2\eta}(\theta - \theta^t)^2. \\ \text{subject to} \quad & \left\langle \theta - \theta^t, -\nabla \mathcal{V}^t \right\rangle \geq \alpha \mathcal{V}^t. \end{aligned} \tag{CPG}$$

In addition to the assumptions above, we acknowledge the implicit assumption that $\theta^t \in \mathcal{A}$ and that $\nabla\mathcal{L}$ and $\nabla\mathcal{V}$ are finite.

Remark C.1. *The objective of the CPG subproblem may be re-expressed as*

$$\left\langle \theta, \nabla \mathcal{L}^t \right\rangle + \frac{1}{2\eta}(\theta - \theta^t)^2 = \frac{1}{2\eta}\Big(\theta - (\theta^t - \eta \nabla \mathcal{L}^t)\Big)^2 + \underbrace{\frac{1}{2\eta}(\theta^t)^2 - \frac{1}{2\eta}(\theta^t - \eta \nabla \mathcal{L}^t)^2}_{\textit{independent of } \theta}.$$

That is, CPG minimizes the distance between θ^{t+1} and $(\theta^t - \eta \nabla \mathcal{L})$, subject to the constraints imposed by $\mathcal{A}$ and $\nabla \mathcal{V}$.

We next establish the following lemmas about the method before restating Thm. 4.1.1 and providing our complete proof.

Lemma C.1 (Satisfiability implies η-bounded updates)**.** *The constraints for CPG are* not *satisfiable, as $\eta \to 0$, only if*

$$\theta^t = \underset{\theta \in \mathcal{A}}{\arg\min} \quad \mathcal{Z} \coloneqq -\left\langle \theta, -\nabla \mathcal{V}^t \right\rangle \quad \textit{and} \quad \mathcal{V}^t > 0. \tag{C.8}$$

That is, a prerequisite for the infeasibility of CPG is that we are outside the feasible set ($\mathcal{V}^t > 0$) and cannot further decrease $-\langle \theta, -\nabla \mathcal{V}^t \rangle$ (i.e., decrease $\mathcal{V}$, linearly approximated) subject to $\theta \in \mathcal{A}$.

In all other cases (which we are guaranteed by Asm. 4.1.2), the magnitude $|\theta^{t+1} - \theta^t|$ is bounded by a quantity that is linear in η.

Proof Sketch. We show that, if Eq. (C.8) does *not* hold, then the constraints of CPG must be satisfiable for some $\eta > 0$, and $|\theta^{t+1} - \theta^t|$ is bounded by a quantity linear in η.

> *Proof.* There are two cases in which Eq. (C.8) fails to hold:
>
> First, we consider the case where $\mathcal{V}^t \leq 0$. In this case, the constraints of CPG are satisfiable by $\theta^t \in \mathcal{A}$, for any $\eta > 0$, by the null update $\theta^{t+1} = \theta^t$. By Rem. C.1, all points with lower objective value than the null update lie within a ball of radius

$2\eta|\nabla\mathcal{L}^t|$ centered at θ^t (specifically, the ball of radius $\eta|\nabla\mathcal{L}^t|$ centered at $\theta^t - \eta\nabla\mathcal{L}^t$, that is, $|\theta^{t+1} - (\theta^t - \eta\nabla\mathcal{L}^t)| < \eta|\nabla\mathcal{L}^t|$). This yields the bound $|\theta^{t+1} - \theta^t| < 2\eta|\nabla\mathcal{L}^t|$, which is an upper bound for the update magnitude that is linear in η, as desired. Second, suppose we are outside the feasible set (that is, $\mathcal{V}^t > 0$), but Eq. (C.8) still does not hold. It follows that there must exist some point $\theta' \in \mathcal{A}$ that improves upon the objective $\mathcal{Z}$ by some nonzero amount. That is, there exists some $\upsilon > 0$ for which

$$\left\langle \theta', -\nabla\mathcal{V}^t \right\rangle - \left\langle \theta^t, -\nabla\mathcal{V}^t \right\rangle > \upsilon. \tag{C.9}$$

By the convexity of $\mathcal{A}$ (Asm. C.1) and the requirement $\theta^t \in \mathcal{A}$, it is clear that all points along the linear path

$$\theta(\omega) \coloneqq \omega\theta' + (1-\omega)\theta^t \colon \omega \in (0,1) \tag{C.10}$$

are also in $\mathcal{A}$, while, by the linearity of $\mathcal{Z}$,

$$\left\langle \theta(\omega), -\nabla\mathcal{V}^t \right\rangle - \left\langle \theta^t, -\nabla\mathcal{V}^t \right\rangle > \upsilon\omega > 0. \tag{C.11}$$

For any sufficiently small step size η such that $0 < \eta < \frac{\upsilon}{\sigma\mathcal{V}^t}$, the point given by $\theta(\omega)$, for $\omega = \eta\frac{\sigma\mathcal{V}^t}{\upsilon}$ (which entails $0 < \omega < 1$), satisfies the constraints of CPG. That is,

$$\begin{aligned} &\left\langle \theta(\omega), -\nabla\mathcal{V}^t \right\rangle - \left\langle \theta^t, -\nabla\mathcal{V}^t \right\rangle = \left\langle \theta(\omega) - \theta^t, -\nabla\mathcal{V}^t \right\rangle > \upsilon\omega = \eta\sigma\mathcal{V}^t = \alpha\mathcal{V}^t. \\ &\theta(\omega) \in \mathcal{A}. \end{aligned} \tag{C.12}$$

By a similar argument as the first case, having established that a feasible point $\theta(\omega)$, for $\omega = \eta\frac{\sigma\mathcal{V}^t}{\upsilon}$, exists and, by (Rem. C.1), noting that the update θ^{t+1} cannot therefore be further than $\theta(\omega)$ away from $\theta^t - \eta\nabla\mathcal{L}^t$, Eq. (C.12) entails, by successive use of

the triangle inequality (see Fig. C.1), the bound

$$\begin{aligned} |\theta^{t+1} - \theta^t| &\leq \underbrace{2|\theta^t - (\theta^t - \eta\nabla\mathcal{L}^t)|}_{2\eta|\nabla\mathcal{L}^t|} + \underbrace{|\theta^t - \theta(\omega)|}_{\omega|\theta^t - \theta'|} \\ &\leq \eta\left(2|\nabla\mathcal{L}^t| + \frac{\sigma\mathcal{V}^t}{\upsilon}|\theta^t - \theta'|\right). \end{aligned} \tag{C.13}$$

We have again bounded the update magnitude linearly in η, as claimed.

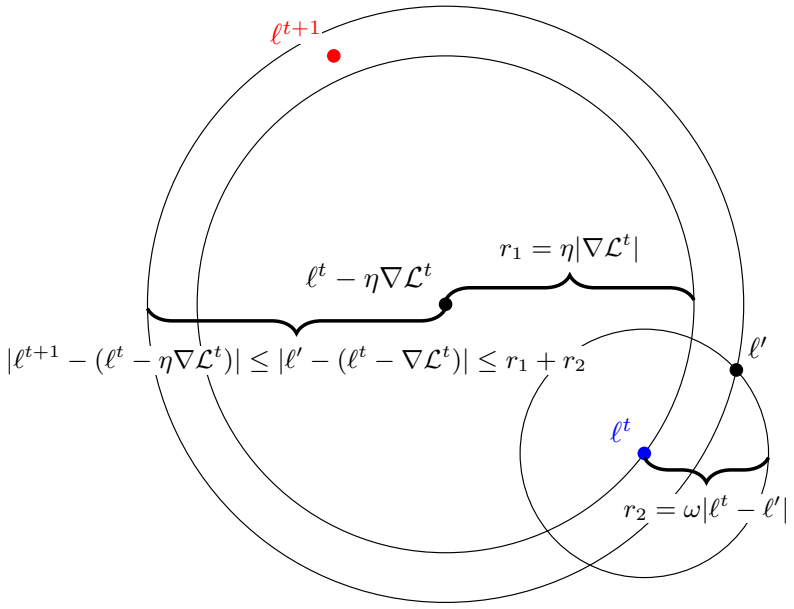

Figure C.1: Successive applications of the triangle inequality yield Eq. (C.13).

Lemma C.2. *Iteration of CPG achieves feasibility (fairness) $\mathcal{V}^t \leq 0$ as ($t \to \infty$) when the step size η is sufficiently small.*

Proof. Note, if $\mathcal{V}^t \leq 0$, we have already achieved fairness. We therefore restrict our attention to the case $\mathcal{V}^t > 0$. In addition, Asm. 4.1.2 guarantees feasibility. It follows that

$$\left\langle \theta^{t+1} - \theta^t, -\nabla\mathcal{V}^t \right\rangle > \alpha\mathcal{V}^t > 0.$$

A sufficiently small step size $\eta > 0$, coupled with the results of Lem. C.1, which bound

the update magnitude linearly in η, ensure that we may rely on linear approximations of $\mathcal{V}$ to adequately describe local behavior. Therefore,

$$\left\langle \theta^{t+1} - \theta^t, -\nabla \mathcal{V}^t \right\rangle > 0 \implies \mathcal{V}^{t+1} < \mathcal{V}^t.$$

It follows that, outside of the feasible set, for sufficiently small step sizes, $\mathcal{V}(\theta)$ is a Lyapunov function for dynamics given by sequential iterations of CPG. This function is monotonically decreasing except at critical points, i.e. where $\nabla \mathcal{V} = 0$. However, by Asm. 4.1.3, any such critical point must a global minimum, which by Asm. 4.1.2 entails that it is feasible. Iterations of CPG therefore achieve feasibility for sufficiently small step sizes.

Lemma C.3. *Given $\mathcal{V}^t \leq 0$, CPG achieves a local optimum (subject to the constraints) of the objective function $\mathcal{L}$ as ($t \to \infty$), when the step size η is sufficiently small.*

Proof. The objective of CPG is convex in θ. Having assumed feasibility for each iterate of CPG (Asm. 4.1.2), this implies that any null update (i.e., $\theta^{t+1} = \theta^t$) occurs only when θ^t is already a feasible local minimum of $\mathcal{L}$. In all other cases, let us establish that $\mathcal{V}^t \leq 0 \implies \left\langle \theta^{t+1} - \theta^t, -\nabla \mathcal{L}^t \right\rangle > 0$.

Lem. C.1 guarantees solutions within an update magnitude of θ^t that is bounded linearly in η. When θ^t is not already a local minimum of $\mathcal{L}$, minimization of the objective will naturally ensure $\left\langle \theta^{t+1} - \theta^t, \nabla \mathcal{L} \right\rangle < 0$. Therefore, for sufficiently small step size $\eta > 0$, we may rely on linear approximations of $\mathcal{L}$ to adequately describe local behavior, and

$$\left\langle \theta^{t+1} - \theta^t, -\nabla \mathcal{L}^t \right\rangle > 0 \implies \mathcal{L}^{t+1} < \mathcal{L}^t.$$

With this observation, it follows that $\mathcal{L}$ is a Lyapunov function inside the feasible set,

which will converge to a local minimum.

Finally, we restate our main result, which we establish by combining the above lemmas.

Theorem 4.1.1 (Asymptotic Convergence)**.** *Subject to Asms. 4.1.1 to 4.1.3, as* $(t \to \infty)$*, constrained projected gradient (Eq.* (CPG)*) converges to a feasible local optimum of the objective when the step size* η *is sufficiently small.*

Proof. This is a direct result of Lems. C.2 and C.3.

Proof of Thm. 4.2.1

Theorem 4.2.1 (Strict Quasiconcavity)**.** $Q_g(\theta_g) - N_g(\theta_g)$ *is strictly quasi-concave in* θ_g*. This guarantees that no more than two zeros of the function* $(Q_g(\theta_g) - N_g(\theta_g))$ *exist, each corresponding to a potential equilibrium.*

Proof of Thm. 4.2.1. We proceed by characterizing the function $B(\theta_g) \coloneqq Q_g(\theta_g) - N_g(\theta_g)$, starting with its zeros. by Eqs. (3.33) and (3.34), the marginal relative increase in the utility of qualification for group g, as the classifier's threshold feature value θ_g is varied, is

$$\frac{\mathrm{d}B}{\mathrm{d}\theta} = q_1(\theta)(A_{1,-1} - A_{1,1}) - q_{-1}(\theta)(A_{-1,-1} - A_{-1,1}) \tag{C.14a}$$

$$= \left(\frac{q_1(\theta)}{q_{-1}(\theta)} - \frac{A_{-1,-1} - A_{-1,1}}{A_{1,-1} - A_{1,1}}\right)\Big(q_{-1}(\theta)(A_{1,-1} - A_{1,1})\Big). \tag{C.14b}$$

Recall that $A_{y,-1} \neq A_{y,1}$ and $A_{1,-1} < A_{1,1}$ (Asm. 3.5.2). By the strict (increasing) monotonicity of $\frac{q_1(\theta)}{q_{-1}(\theta)}$ in θ and strict positivity of $q_{-1}(\theta)$, both guaranteed by Asm. 3.5.1, the sign of this expression can change at most once as θ is varied from $-\infty$ to ∞. We denote the value of θ at which the sign of this first derivative changes as $\theta^\star$:

$$\frac{q_1(\theta^\star)}{q_{-1}(\theta^\star)} = \frac{A_{-1,-1} - A_{-1,1}}{A_{1,-1} - A_{1,1}}. \tag{C.15}$$

Moreover, it follows that

$$\theta < \theta^\star \implies \frac{\mathrm{d}}{\mathrm{d}\theta} B > 0. \tag{C.16a}$$

$$\theta > \theta^\star \implies \frac{\mathrm{d}}{\mathrm{d}\theta} B < 0. \tag{C.16b}$$

$B(\theta_g) \coloneqq Q_g(\theta_g) - N_g(\theta_g)$ is therefore strictly quasi-concave, from which it follows that only two zeros of the function can exist (*By contradiction, more than two zeros would require the function, which has no discontinuities, to invert its slope more than once.*)

For completeness, we may also take a second derivative of B with respect to θ:

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\Big(B(\theta)\Big) = \frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{q_1(\theta)}{q_{-1}(\theta)}\right)\Big(q_{-1}(\theta)(A_{1,-1} - A_{1,1})\Big) \\ + \left(\frac{q_1(\theta)}{q_{-1}(\theta)} - \frac{A_{-1,-1} - A_{-1,1}}{A_{1,-1} - A_{1,1}}\right)\left(\frac{\mathrm{d}}{\mathrm{d}\theta} q_{-1}(\theta)\right) \end{aligned} \tag{C.17}$$

Doing so, we observe that B may have any number of inflection points, but $\theta^\star$ cannot be one of them. We see this because the second term of the expression above evaluated at $\theta^\star$ must be zero, but the first term must be non-zero by Asm. 3.5.1 and Asm. 3.5.2. It follows that $\theta^\star$ is the unique occurrence of a local extremum and therefore a global extremum of $B(\theta)$.

Bibliography

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Alexander Amini, Tsun-Hsuan Wang, Igor Gilitschenski, Wilko Schwarting, Zhijian Liu, Song Han, Sertac Karaman, and Daniela Rus. Vista 2.0: An open, data-driven simulator for multimodal sensing and policy learning for autonomous vehicles. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2419–2426. IEEE, 2022.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias risk assessments in criminal sentencing. *ProPublica*, 2020.

Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3682–3689, 2022.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

Jonas Björnerstedt and Jörgen W Weibull. Nash equilibrium and evolution by imitation. Technical report, IUI Working Paper, 1994.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Kianté Brantley, Miro Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *Advances in Neural Information Processing Systems*, 33:16315–16326, 2020.

Steven L Brunton, Marko Budišić, Eurika Kaiser, and J Nathan Kutz. Modern koopman theory for dynamical systems. *arXiv preprint arXiv:2102.12086*, 2021.

Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 224–232. ACM, 2018.

Yatong Chen, Reilly Raab, Jialu Wang, and Yang Liu. Fairness transferability subject to bounded distribution shift. In *Advances in Neural Information Processing Systems*,

volume 35, pages 11266–11278, 2022.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Brian Christian. *The alignment problem: Machine learning and human values*. WW Norton & Company, 2020.

Stephen Coate and Glenn C Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pages 1220–1240, 1993.

Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust region methods*. SIAM, 2000.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 91–98, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314236. URL `https://doi.org/10.1145/3306618.3314236`.

Kate Crawford and Ryan Calo. There is a blind spot in AI research. *Nature News*, 538 (7625):311, 2016.

Ross Cressman and Yi Tao. The replicator equation and other game dynamics. *Proceedings of the National Academy of Sciences*, 111(supplement_3):10810–10817, 2014.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. 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, 2020.

Jeffrey Dastin. Amazon ditched ai recruiting tool that favored men for technical jobs. *Reuters*, 2018.

Sarah Dean, Mihaela Curmei, Lillian J Ratliff, Jamie Morgenstern, and Maryam Fazel. Multi-learner risk reduction under endogenous participation dynamics. *arXiv preprint arXiv:2206.02667*, 2022.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances*

in Neural Information Processing Systems, 33:8378–8390, 2020.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR, 2021a.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490. Curran Associates, Inc., 2021b. URL `https://proceedings.neurips.cc/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html`.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference of Fairness, Accountability, and Transparency*, 2018.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Roger Fletcher. An exact penalty function for nonlinear programming with inequalities. *Mathematical Programming*, 5:129–150, 1973.

Daniel Friedman. Evolutionary games in economics. *Econometrica: journal of the econometric society*, pages 637–666, 1991.

Daniel Friedman and Barry Sinervo. *Evolutionary Games in Natural, Social, and Virtual Worlds*. Oxford University Press, 2016.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? The effects of machine learning on credit markets. *The Effects of Machine Learning on Credit Markets*, 2018.

Nico Grant. Google chatbot's a.i. images put people of color in nazi-era uniforms. *The New York Times*, 2024.

Nico Grant and Kashmir Hill. Google's photo app still can't find gorillas. and neither can apple's. *The New York Times*, 2023.

Karen Hao. The coming war on the hidden algorithms that trap people in poverty. *MIT Technology Review*, 2020.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, page 111–122, New York, NY, USA, November 2016a. Association for Computing Machinery. URL `http://arxiv.org/abs/1506.06980`. Number: arXiv:1506.06980 arXiv:1506.06980 [cs].

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016b.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Hoda Heidari, Vedant Nanda, and Krishna P. Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. *the International Conference on Machine Learning (ICML)*, 2019.

Daniel Hennes, Dustin Morrill, Shayegan Omidshafiei, Remi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duenez-Guzman, et al. Neural replicator dynamics. *arXiv preprint arXiv:1906.00190*, 2019.

Alex Hern. Google's solution to accidental algorithmic racism: ban gorillas. *The Guardian*, 2018.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

Joe Hernandez. A military drone with a mind of its own was used in combat, U.N. says. *National Public Radio*, 2021.

Josef Hofbauer, Karl Sigmund, et al. *Evolutionary games and population dynamics*. Cambridge university press, 1998.

Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1389–1398. International World Wide Web Conferences Steering Committee, 2018.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.

Yaowei Hu and Lu Zhang. Achieving long-term fairness in sequential decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9549–9557, 2022.

Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26):eaau5872, 2019.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Joseph R. Biden Jr. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 2023. URL `https://www.whitehouse.gov`.

Krishna C Kalagarla, Rahul Jain, and Pierluigi Nuzzo. A sample-efficient algorithm for episodic finite-horizon mdp with constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8030–8037, 2021.

Mintong Kang, Linyi Li, Maurice Weber, Yang Liu, Ce Zhang, and Bo Li. Certifying some distributional fairness with subpopulation decomposition. *arXiv preprint arXiv:2205.15494*, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 576–586, 2021.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

Wei Li, CW Pan, Rong Zhang, JP Ren, YX Ma, Jin Fang, FL Yan, QC Geng, XY Huang, HJ Gong, et al. Aads: Augmented autonomous driving simulation using data-driven algorithms. *Science robotics*, 4(28):eaaw0863, 2019.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018. tex.organization: PMLR.

Lydia T Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 381–391, 2020.

Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021.

Cade Metz and Adam Satariano. An algorithm that grants freedom, or takes it away. *The New York Times*, 2020.

Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 429–438, 2020.

Hussein Mouzannar, Mesrob I Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 359–368. ACM, 2019.

Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J Ratliff. Multiplayer performative prediction: Learning in decision-dependent games. *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (arXiv:2201.03398)*, 2022.

Derek Newton. From admissions to teaching to grading, AI is infiltrating higher education. *The Hechinger Report*, 2021.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative

prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, February 2020. URL http://arxiv.org/abs/2002.06673. arXiv: 2002.06673.

Pope Francis. Messaggio 57 Giornata Mondiale della Pace 2024, 2023. URL https://www.vatican.va.

Reilly Raab and Yang Liu. Unintended selection: Persistent qualification rate disparities and interventions. *Advances in Neural Information Processing Systems*, 34:26053–26065, 2021.

Reilly Raab, Luca de Alfaro, and Yang Liu. Conjugate natural selection. *arXiv preprint arXiv:2208.13898*, 2022.

Reilly Raab, Ross Boczar, Maryam Fazel, and Yang Liu. Fair participation via sequential policies. In *AAAI*, volume 38, 2024.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. Robust fairness under covariate shift. In *AAAI*, 2021.

William H Sandholm. *Population games and evolutionary dynamics*. MIT press, 2010.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H. Chi. Transfer of machine learning fairness across domains, 2019.

Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, New York, NY, USA, 2021. Association for Computing Machinery.

Rahul Singh, Abhishek Gupta, and Ness B Shroff. Learning in markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.

Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382. PMLR, 2019a.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Min Wen, Osbert Bastani, and Ufuk Topcu. Fairness with dynamics. *arXiv preprint arXiv:1901.08568*, 2019.

Joshua Williams and J Zico Kolter. Dynamic modeling and equilibria in fair decision making. *arXiv preprint arXiv:1911.06837*, 2019.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491. PMLR, 2021.

Tongxin Yin, Reilly Raab, Mingyan Liu, and Yang Liu. Long-term fairness with unknown dynamics. *Advances in Neural Information Processing Systems*, 36, 2023.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 228–238. Curran Associates Inc., 2017b. ISBN 9781510860964.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

Xueru Zhang, Mohammad Mahdi Khalili, Cem Tekin, and Mingyan Liu. Group retention when using machine learning in sequential decision making: The interplay between user dynamics and fairness. In *Advances in Neural Information Processing Systems*, pages 15243–15252, 2019.

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellström, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? *arXiv preprint arXiv:2010.11300*, 33:18457–18469, 2020.

Liyuan Zheng and Lillian Ratliff. Constrained upper confidence reinforcement learning. In *Learning for Dynamics and Control*, pages 620–629. PMLR, 2020.