Achieving Long-Term Fairness in Performative Reinforcement Learning

Uddalak Mukherjee

Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, Howrah
Pin - 711 202, West Bengal

A thesis submitted to
Ramakrishna Mission Vivekananda Educational and Research Institute
in partial fulfillment of the requirements for the degree of
MSc in Big Data Analytics
2025

Achieving Long-term Fairness in Performative Reinforcement Learning

By
Uddalak Mukherjee

Declaration by student:

"I hereby declare that the present dissertation is the outcome of my project work under the guidance of Dr. Debabrota Basu and I have properly acknowledged the sources of materials used in my project report."

Uddalak Mukherjee

(Uddalak Mukherjee, ID No. B2330042)

A project report in the partial fulfilment of the requirements of the degree of MSc in Big Data Analytics

Examined and approved on

28/05/25

by

Dr. Debabrota Basu (supervisor)
Faculty(ISFP), Scool team
Inria Center at University of Lille, France

Countersigned by

Registrar
Ramakrishna Mission Vivekananda Educational and Research Institute

Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, Howrah 711202, West Bengal, India

May 28, 2025

Acknowledgment

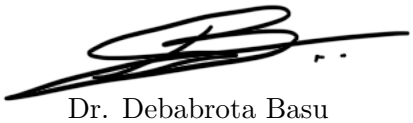
The present project work is submitted in partial fulfillment of the requirements for the degree of Master of Science of Ramakrishna Mission Vivekananda Educational and Research Institute (RKMVERI). I express my deepest gratitude to my supervisor Prof. Dr. Debabrota Basu of Inria Center at University of Lille, France for his inestimable support, encouragement, profound knowledge, largely helpful conversations, and also for providing me with a systematic way for the completion of my project work. His ability to work hard inspired me a lot. Additionally, I want to extend my warmest gratitude to Mr. Udvas Das, PhD Scholar, Inria Center at University of Lille, France for providing me with inputs throughout the course of this project. I am also extremely grateful to the Vice-Chancellor of this University for his encouragement and support throughout the course. This work would not have been possible without the immense support of my parents and last but not the least, my fellow classmates.

Belur
May 28, 2025

Uddalak Mukherjee
Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute

CERTIFICATE FROM THE SUPERVISOR

This is to certify that the thesis entitled *'Achieving Long-term Fairness in Performative Reinforcement Learning'* submitted by *Mr. Uddalak Mukherjee*, who has been registered for the award of MSc in Big Data Analytics degree of Ramakrishna Mission Vivekananda Educational and Research Institute, Belur Math, Howrah, West Bengal is absolutely based upon his own work under the supervision of *Dr. Debabrota Basu*, faculty (ISFP) at the Scool project-team (previously called SequeL) of Inria Center at University of Lille in France, that neither his thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.



Dr. Debabrota Basu
Tenured Faculty (ISFP)
Scool team
Inria Center at University of Lille, France

Abstract

This work explores the study of performative reinforcement learning (RL) through the lens of policy gradient methods, with a view toward supporting long-term fairness objectives. In many real-world settings, deployed policies influence the environment in ways that alter future data distributions, a phenomenon known as performative feedback. Classical policy gradient methods ignore this feedback, often leading to instability or sub-optimality. We propose *PePG*—a performative policy gradient algorithm that operates in two phases: a warm-up phase that mimics standard policy gradient, followed by a performative-aware update phase incorporating estimated models of the environment. While PePG does not explicitly incorporate fairness constraints, it sets the groundwork for future work in this direction, particularly since policy gradient approaches have been shown to naturally align with long-term fairness formulations. Experiments in a stylized grid-world demonstrate that PePG attains performative stability and approximate optimality, significantly outperforming its non-performative counterpart. This work represents a first step toward integrating fairness considerations into performative RL through principled policy optimization.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Fairness in machine learning is an interdisciplinary endeavor that examines how automated decision-making systems can inadvertently perpetuate or exacerbate societal biases, aiming to ensure that model outcomes do not disproportionately harm or advantage individuals based on sensitive attributes such as race, gender, or socioeconomic status (Barocas et al., 2023a). Literature formalizes key concepts like group fairness, where statistical measures (e.g., equalized odds or demographic parity) enforce parity across protected groups, and individual fairness, where a task-specific metric guarantees that similar individuals receive similar treatment, hence providing both the theoretical foundations and practical algorithms to adjust models accordingly (Hardt et al., 2016; Dwork et al., 2012).

Fairness Types. Fairness in traditional machine learning, now often called *static fairness* assumes a fixed data distribution and enforces group or individual level constraints on a trained model. By contrast, *dynamic fairness* addresses sequential decision-making systems whose actions influence future data: it splits into two complementary paradigms. *Instantaneous fairness* (also termed *stepwise* or *per-iteration fairness*) requires that group-fairness constraints hold at each decision step on the samples observed in that round (Deng et al., 2022; Bechavod and Ligett, 2019). *Long-term fairness* instead imposes that these constraints be satisfied cumulatively over time, accounting for feedback loops between decisions and population dynamics as suggested by Xu et al. (2024); Chi et al. (2022). A rich literature now studies algorithms ranging from no-regret online learners (Yin et al., 2024) to reinforcement-learning frameworks with cumulative-constraint guarantees that ensure fairness in an online fashion while balancing utility and long-term equilibrium outcomes (Yin et al., 2023; Chiappa, 2023). Optimization-based approaches (Du et al., 2024) explore online mirror descent algorithms with fairness constraints to attain long-term fairness.

Working Example What motivates our project is the adoption of long-term fairness in reinforcement learning (Xu et al., 2024; Chi et al., 2022; Yin et al., 2023). Many of these works do not consider the interaction of the model with data, which leads to shifts in the data distribution in a performative fashion (Brown et al., 2022; Piliouras and Yu, 2023a; Mendler-Dünner et al., 2020) that can affect our long-term satisfaction of fairness constraints. We look into this problem with a real-world example of credit loaning. We adopt the illustrative example from (Xu et al., 2024), where a bank acts as the decision-making agent with binary actions: to either approve or reject loan applications. In this setup, while the bank enforces *instantaneous fairness* at each decision point, the system nonetheless exhibits *long-term unfairness*.

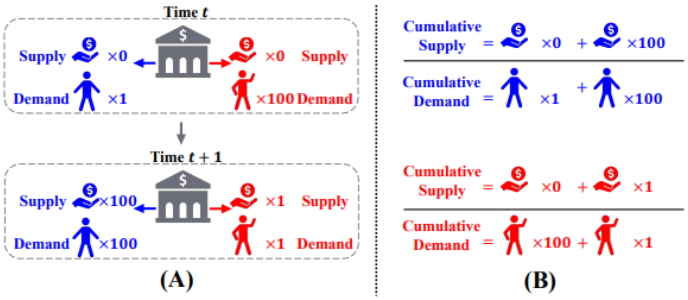


Figure 1.1: **(A)** At time t, the bank approves 0 loans out of 1 blue applicant and 0 out of 100 red applicants. At $t+1$, it approves 100 out of 100 blue applicants and 1 out of 1 red applicant.
(B) The long-term benefit rate bias is computed as $|\frac{100}{101} - \frac{1}{101}|$, revealing disparity from the bank's decisions.

As demonstrated in Figure 1.1, at both time steps t and $t+1$, the disparity in supply-demand ratios across the two demographic groups is zero, suggesting fairness at the respective time instances. However, the cumulative difference in acceptance ratios at time $t+1$ reaches $\frac{99}{101}$, a value significantly close to 1 and far from zero. This observation indicates a considerable violation of long-term fairness.

This problem can be mitigated by employing long-term fairness-aware reinforcement learning algorithms, such as the fairness-constrained policy gradient method proposed in the same work (see Section 3.3 for more details). While such methods successfully enforce fairness over time horizons, they typically neglect the performative effects—that is, how the environment (in this case, the distribution of applicant credit scores) evolves in response to the deployed decision-making policy.

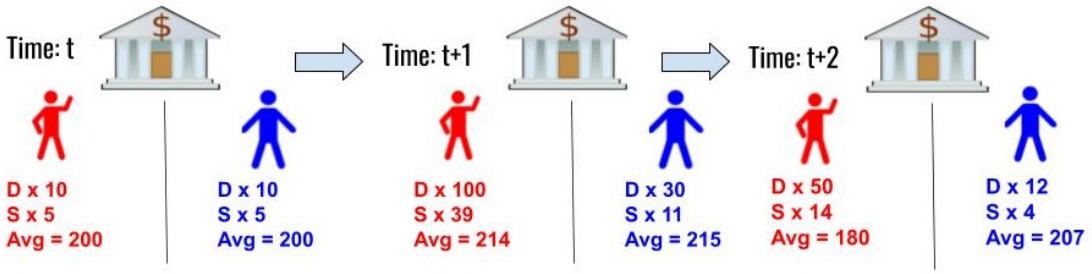


Figure 1.2: Performative effect: Average credit score disparity grows over time

Figure 1.2 further highlights this issue. Although the applied algorithm ensures long-term fairness across all three time steps, we observe a growing disparity in the average population credit scores (denoted using "Avg") between the two groups. While the change between time steps t and $t+1$ is negligible, the divergence becomes pronounced between $t+1$ and $t+2$. This phenomenon stems from the high rejection rate faced by the red group at $t+1$, where 61 applicants were denied credit. Assuming approximately equal population sizes, this indicates that a substantial fraction of the red group applied for loans and faced rejection. Consequently, many individuals—potentially with respectable credit scores, were denied access to financial support, leading to a decline in their scores and, by extension, a decrease in the group's average creditworthiness. In contrast, fewer applicants from the blue group sought loans, resulting in a smaller performative effect on their credit score distribution.

The above example underscores the importance of adopting learning frameworks that explicitly model the feedback loop between policy decisions and the evolving data distribution. In such environments, where the agent's actions influence the future data-generating process, it is inadequate to rely solely on traditional fairness-aware policy gradient algorithms. Instead, it is crucial to adopt *performative* variants of such methods that account for the dynamic and endogenous nature of the environment. This is precisely the focus of the **performative prediction** framework.

Justification for Choosing Policy Gradient Policy gradient methods are particularly well-suited for incorporating long-term fairness constraints, as highlighted in prior work (Xu et al., 2024). This is primarily because fairness considerations can be naturally embedded into the policy optimization framework by modifying the objective to include fairness-aware terms—often achieved by replacing the standard advantage function with a fairness-adjusted variant. Importantly, such modifications do not compromise the convergence guarantees of the underlying algorithm. Furthermore, since the policy directly governs the agent's decision-making process, optimizing it allows for explicit control over behavioral outcomes. This makes policy gradient a principled and practical choice for our performative reinforcement learning framework with fairness considerations.

Our Contribution However, to the best of our knowledge, there exists no well-established **Performative Policy Gradient** algorithm that integrates fairness constraints in the context of performative environments. To address this gap, we propose a novel REINFORCE-style algorithm, which we term *PePG*, designed within the framework of *Performative Reinforcement Learning* (Mandal et al., 2023), where both the reward and transition dynamics are influenced by the agent's current policy.

Chapter 2

Literature Review

2.1 Fairness in Machine Learning

We aim to explore the notion of long-term fairness in a sequential decision making problem. For that it becomes inherent to discuss first the nuances or recent advancements in static fairness for RL problems.

Static Fairness. Static algorithmic fairness refers to fairness criteria evaluated on a fixed snapshot of decisions, typically by checking if an algorithm's outcomes are evenly distributed across groups (for example, requiring a model to approve loans at equal rates for different demographics or to have the same true-positive rate for each group) (Oneto and Chiappa, 2020; Caton and Haas, 2024; Hort et al., 2023; Balseiro et al., 2021; Hsu et al., 2022; Hardt and Recht, 2021). In the last decade attempting to formalize the decision making problem as a constrained optimization, researchers have come up with different metrics of static fairness (such as Equal Opportunity: equality in prevalence rates, Demographic Parity: equality in TPR etc.) (Barocas et al., 2023b; Caton and Haas, 2024; Kilbertus et al., 2017). A recurring bottleneck in these works revolves around a well-known phenomenon called the "*Impossibility Theorem*" (Kleinberg et al., 2016; Fudenberg and Levine, 2012; Saravanakumar, 2021; Bell et al., 2023). It says if the classifier has to ensure calibration within groups, balance for positive and negative groups simultaneously, then the classifier either has to be *perfect* or the data must have equal *base rates* across the groups. An ϵ-approximate version of the theorem was also introduced in the paper. Though Kleinberg et al. (2016) includes a brief discussion on this setting, (Bell et al., 2023) encompasses a more detailed analysis.

Why Long-term? Several works (Ge et al., 2021; Hu and Zhang, 2022; D'Amour et al., 2020; Zhang et al., 2020) show if an algorithm satisfies a static fairness rule, it doesn't nec-

essarily improve the long-term outcomes for the very groups it aims to protect (Liu et al., 2018). So it is myopic to define fairness in machine learning as a certificate of *instantaneous* statistical parity (Caton and Haas, 2024; Barocas et al., 2023b; Oneto and Chiappa, 2020). It becomes imperative to consider the effect of decisions and corresponding feedback on future fairness guarantees of the algorithm, thus analyzing long-term algorithmic fairness.

Long-term Fairness. The literature on Long-term fairness is stratified into two main regimes.

(a) **MDP.** Yin et al. (2023) propose model iterative fair classification as a finite-horizon RL problem, treating the population distribution as the state and classifiers as stochastic actions sampled from the agent's policy. The optimization problem proposed in the paper involves the maximization of the reward that is associated with the loss function under the constraint of bounded fairness violation. Under some theoretical assumptions, the paper proposes a fair version of Least Squares Value Iteration UCB called L-UCBFair which successfully optimizes the concerned objective.

Extending the idea on MDPs (Kokhlikyan et al., 2022; Li et al., 2023, 2025) , (Xu et al., 2024) proposes a policy gradient approach towards mitigating long-term bias. The agent receives immediate demand and supply along with the reward while interacting with the environment. The difference in the group-wise cumulative demand and supply ratio is termed as "ratio-before-aggregation" or bias associated with fairness violation in the problem. This work proposes a policy-gradient algorithm named ELBERT-PO which is necessarily a fair version of the *PPO* algorithm using the aforementioned bias in the constrained optimisation.

For entirety, we should also mention other studies like (Chi et al., 2022; Yu et al., 2022; Xian et al., 2023; Rateike et al., 2024) where they often assume non-interacting separate MDPs for groups of the sensitive feature.

(b) **Optimization.** A fundamental work in this sub-domain was done by (Du et al., 2024). They minimize the total loss incurred till pre-specified horizon T subject to the constraint that the cumulative fairness violation is strictly zero. They leverage Online Mirror Descent(OMD) as the regret minimizing algorithm. Other studies like (Zhao et al., 2021) try to incorporate long-term fairness in meta learning. (Balseiro et al., 2021) mainly focuses on resource allocation ahead of the actual fairness constraints, and propose optimization techniques for fair allocation while works like (Zhan et al., 2024; Vadavathi et al., 2024; Ortmann et al., 2024; Rateike et al., 2024) studies long-term fairness in very specific setups that do not provide unified theoretical analyses on the subject.

2.2 Gaps in Long-term Fairness Literature

A major assumption in long-term fairness aware Reinforcement Learning Literature is the static-ness of the environment. Works like (Yin et al., 2023; Xu et al., 2024; Chi et al., 2022) that are based on the MDP setup assume that the transition dynamics are mostly static implying that the setup has some true transition function and reward distribution,i.e., a true underlying MDP the agent tries to learn overtime. However, this does not have very close link with realism because from a realistic point of view, the decision making process may cause shifts in the data generating distribution as the policy interacts with the environment. For the optimization based approaches, it is almost the same scenario. Works like (Du et al., 2024; Zhan et al., 2024; Zhao et al., 2021) do not really explore the performative side of the algorithm, as a result, the assumptions under which the algorithms attain stable regret are not complete in the sense of realism. This exact issue invokes the need for something more than long-term fairness. A truly good attempt at achieving long-term fairness requires us to take into account the performative nature of the decision-making process.

2.3 Performative Prediction & Reinforcement Learning

Why Performative Prediction? Performative prediction acknowledges that predictions themselves can alter the underlying data distribution by influencing individuals' behaviors and choices. In sequential decision-making, this feedback loop compounds over time, necessitating models that account for the evolving environment. By incorporating the performative effects of past decisions, one can better anticipate and correct potential biases. This approach promotes long-term fairness, ensuring equitable outcomes across repeated interventions. As suggested in (Mishler and Dalmasso, 2022), Machine Learning models which are fair during training can become unfair post deployment as the decisions taken by the model affects the data generating distribution, hence the fairness constraints get violated in long-term. Related works like (Roh et al., 2023; Zezulka and Genin, 2024; Baharlouei et al., 2023) also pose similar questions to us.

Performative Prediction in RL. We can dissect the literature on *Performative Prediction* largely into two strata.
(a) **MDP.** Mandal et al. (2023) introduced performative V-value which is the V-value under the assumption of changing reward and transition probability functions, as a response. It maximizes a discounted state-action occupancy measure derived from the performative V-value under the standard Bellman constraint. Repeated Retraining(RR) on the objective ensures convergence to the performatively stable point. An extension to this work by

Rank et al. (2024), proposes a k-delayed RR framework which converges to performative stability a rate faster than ordinary RR but is more space complex. Additionally, it also introduces a mixed delayed RR that that employs samples from previous iterations as well but putting more weight to the more recent samples, also ensuring convergence to performative stability. Some works further extending on this idea like (Mandal and Radanovic, 2024; Yan and Cao, 2023; Rodemann et al., 2024; Chen et al., 2024a) add some model variations to the existing framework.

(b) **Optimization** A fundamental work in this stratum is (Brown et al., 2022) proposes incorporates performative prediction under adversarial feedback. The models introduced are decision-dependent on data shifts where a decision-maker selects a model at each round, and the environment which adversarially updates the data distribution through a transition map leading to feedback loops. The problem here is referred to as *Repeated Risk Minimization* where the institution minimises it's risk (expected loss for the deployed model in that round w.r.t samples from the new data generating distribution introduced by the environment in that round) at each round depending on the samples generated by the environment as a response to it's previous deployment. This is a special case of Repeated Re-training. Furthermore, it also proposes a k-delayed RRM where k-groups respond slowly following the model published by the institution and the institution minimizes the risk once in k rounds. Again this algorithm ensures faster convergence at the cost of increment in space complexity. Papers extremely similar to this work are (Mendler-Dünner et al., 2020; Kim and Perdomo, 2022; Izzo et al., 2022; Piliouras and Yu, 2023a; Cai et al., 2024) utilize similar underlying setups and propose algorithms which ensure convergence along with providing stable regret.

Other papers like (Miller et al., 2021) provide detailed theoretical analyses. They try to identify the conditions necessary to prove convexity of performative risk and hence proposing simple gradient-free method using random perturbations and a 2-step algorithm using Least Squares while assuming the location family for the data generating distribution. Related works like (Taori and Hashimoto, 2022) manage to showcase how model-generated labels together with human annotations can cause a shift in the data generating distribution hence amplifying the bias. Works like (Chen et al., 2024b) consider multi-armed bandit setup with performative feedback providing similar regret guarantees. Some related papers with similar motivation are (Piliouras and Yu, 2023b; Zheng et al., 2024; Li et al., 2022). (Izzo et al., 2021) provided a performative REINFORCE-style gradient descent algorithm in a non-RL based setup which helped us greatly in designing our own algorithm.

2.4 Policy Gradient Algorithms in Reinforcement Learning

Policy gradient methods estimate the gradient of the expected return with respect to policy parameters. The foundational policy gradient theorem (Sutton et al., 1999) expresses this gradient as an expectation involving the score function and the action-value function. Earlier, Williams (1992a) introduced the REINFORCE algorithm, which provides an unbiased likelihood-ratio estimator for this gradient. Convergence guarantees for gradient-ascent-based policy improvement were also established in these works. Konda and Tsitsiklis (2000) analyzed actor–critic algorithms using two-timescale stochastic approximation. Later, Kakade (2002) proposed the natural policy gradient, which incorporates curvature information via the Fisher information matrix. Recent theoretical analyses, such as (Agarwal et al., 2021), leverage smoothness and Lipschitz conditions to formalize convergence rates and robustness under distribution shifts.

2.5 Overall Summary

To put all the things together, most works in both fairness and performative prediction assume a Markov Decision Process (MDP) framework, while a few adopt alternative sequential decision-making setups. In fairness-aware reinforcement learning, MDP-based approaches typically employ either policy gradient methods or confidence-bound techniques to minimize regret while optimizing cumulative reward or state-value functions under fairness constraints. Alternate formulations often utilize online mirror descent, which provides stable and non-trivial regret and fairness guarantees even though they may not be very good. In performative prediction, irrespective of whether the setup follows an MDP structure, the predominant approach is Repeated Retraining or its variants which all provide good enough (ε, δ) PAC(Provably Approximate Correct) guarantees. Notably, fairness-focused research has largely overlooked the impact of model-data interactions on the broader environment, presenting a significant opportunity for future exploration.

Chapter 3

Preliminaries

3.1 Summary of Notations

Notation	Definition
T	Horizon
$\mathcal{S}, \mathcal{A}$	Set of all states and actions
$\Delta(\mathcal{A})^{\lvert\mathcal{S}\rvert}$	Space/Simplex of all stochastic policies
$s_t, a_t, \pi^{(t)}$	The state, action and policy at round t
$r(x_t, a_t)$	The reward received from state s_t after taking action a_t
$S_g(s_t, a_t), D_g(s_t, a_t)$	Instantaneous Supply and Demand after taking action a_t from s_t
η_g^S, η_g^D	Cumulative Demand and Supply under infinite horizon
θ	Model parameter (also used as policy parameter)
$\mathcal{W}$	Wasserstein-1 distance
θ_{PS}, θ_{PO}	Performatively Stable and Optimal Model
$J(\pi), J(\theta)$	The total Return acting under policy π or policy parameter θ
π^{PS}, π^{PO}	Performatively Stable and Optimal Policy
$d_\pi(s, a)$	Long-term discounted state action occupancy measure for policy π
$d_{s_0}^\pi(s)$	Discounted state occupancy measure for policy π and starting state s_0
ρ, μ	The starting state and intermediate state distributions
$\psi(s, a), \phi(s, a, s')$	Feature maps for (state, action) and (state, action, next state) tuples

Table 3.1: Summary of Notations

3.2 Fairness

The study of fairness in the static decision making setup often employs some metrics to ensure that the concerned algorithm is fair. In this section we recall some of the most commonly used fairness metrics defined primarily using 3 random variables: A: The sensitive Attribute, Y : The label/class and R: The risk function.

Definition 3.2.1 (Independence)**.** *Random variables* (A, R) *satisfy independence if* $A \perp R$*.*

For a binary classifier $\hat{Y}$ we can represent it in terms of R as $\hat{Y} = \mathbb{I}_{\{R>t\}}$. Correspondingly, independence reduces to,

$$\mathbb{P}(\hat{Y} = 1|A = a) = \mathbb{P}(\hat{Y} = 1|A = b)$$

An alternative relaxation is to consider a ratio condition,

$$\frac{\mathbb{P}(\hat{Y} = 1|A = a)}{\mathbb{P}(\hat{Y} = 1|A = b)} \geq 1 - \epsilon$$

Permitting an $\epsilon > 0$ amount of slack.

Definition 3.2.2 (Seperation)**.** *Random variables* (R, A, Y) *satisfy separation if* $R \perp A|Y$*.*

Once again for a binary classifier $\hat{Y}$, the condition reduces to,

$$\mathbb{P}(\hat{Y} = 1|Y = 1, A = a) = \mathbb{P}(\hat{Y} = 1|Y = 1, A = b)$$

$$\mathbb{P}(\hat{Y} = 1|Y = 0, A = a) = \mathbb{P}(\hat{Y} = 1|Y = 0, A = b)$$

Here we are ensuring equality of both true and false positive rates.

Definition 3.2.3 (Sufficiency)**.** *We say the random variables* (R, A, Y) *satisfy sufficiency if* $Y \perp A|R$*.*

Once again in terms of risk score function R, the condition reduces to,

$$\mathbb{P}(Y = 1|R = r, A = a) = \mathbb{P}(Y = 1|R = r, A = b)$$

Definition 3.2.4 (Calibration)**.** *If a score* R *is calibrated with respect to an outcome variable* Y *if for all score values* $r \in [0, 1]$*, then we have*

$$\mathbb{P}(Y = 1|R = r) = r$$

To formalize the connection between sufficiency and calibration we introduce the notion of calibration by group.

Definition 3.2.5. *Calibration by group satisfies,*

$$\mathbb{P}(Y = 1|R = r, A = a) = r \tag{3.1}$$

for all score values r *and groups* a*.*

3.3 Long-Term Fairness

Several authors have proposed various frameworks for incorporating long-term fairness into sequential decision-making. Du et al. (2024) introduced a fair variant of the online mirror descent algorithm, termed Lotfair, to address this challenge (An enhancement of this method—offering tighter regret bounds, improved guarantees on long-term fairness violations, and greater feasibility has been outlined in the supplementary document accompanying this project, however, this improvement has not yet been experimentally validated). Despite its theoretical appeal, the Lotfair methodology does not account for the changes the environment undergoes as a result of the decision-making process. This limitation renders it less reflective of real-world dynamics when compared to more general formulations such as Markov Decision Processes (MDPs).

Another work (Xu et al., 2024) addresses biases in sequential decision-making by introducing a long-term fairness concept named Equal Long-term Benefit Rate (ELBERT). This concept is seamlessly integrated into a Markov Decision Process (MDP) to consider the future effects of actions on long-term fairness, thus providing a unified framework for fair sequential decision-making problems.

In the sequential setting, each time step t corresponds to a static dataset that comes with group supply and group demand. To adapt this to a Markov Decision Process (MDP), we assume that in addition to the immediate reward $R(s_t, a_t)$, the agent receives the immediate group supply $S_g(s_t, a_t)$ and the immediate group demand $D_g(s_t, a_t)$ at every time step t.

We define the Supply-Demand MDP (SD-MDP) as a tuple:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, S_g, D_g, \gamma)$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s_{t+1} \mid s_t, a_t)$ is the transition probability function, $R(s_t, a_t)$ is the immediate reward function, $S_g(s_t, a_t)$ is the immediate group supply function, $D_g(s_t, a_t)$ is the immediate group demand function, $\gamma \in [0, 1]$ is the discount factor.

Definition 3.3.1. *We define the* ***cumulative group supply*** *under policy π as:*

$$\eta_g^S(\pi) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t S_g(s_t, a_t) \right]. \tag{3.2}$$

Similarly, we define the ***cumulative group demand*** *under policy π as:*

$$\eta_g^D(\pi) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t D_g(s_t, a_t) \right]. \tag{3.3}$$

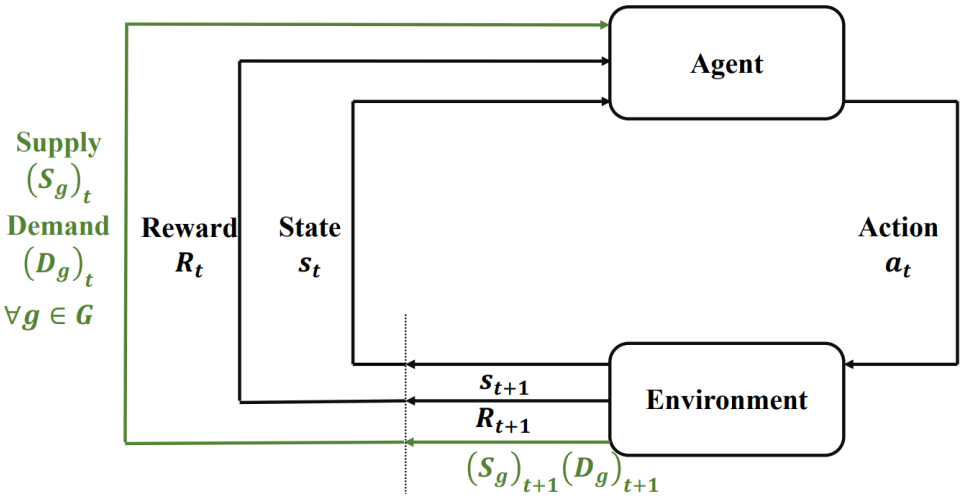


Figure 3.1: The general setup for SD-MDP

These cumulative measures capture the long-term supply and demand in the system, discounted over time with the factor $\gamma \in [0, 1]$.

Definition 3.3.2. *Define the Longterm Benefit Rate of group* g *as* $\frac{\eta_g^S(\pi)}{\eta_g^D(\pi)}$ *Define the bias of a policy as the maximal difference of Long-term Benefit Rate among groups, i.e.,*

$$b(\pi) = \max_{g \in G} \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)} - \min_{g \in G} \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)} \tag{3.4}$$

The Optimization Problem under the framework of ELBERT, the goal of reinforcement learning with fairness constraints is to find a policy to maximize the cumulative reward and keep the bias under a threshold δ. In other words,

$$\begin{aligned} &\text{maximize} \quad \eta(\pi) \\ &\text{subject to} \quad b(\pi) := \max_{g \in G} \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)} - \min_{g \in G} \frac{\eta_g^S(\pi)}{\eta_g^D(\pi)} \leq \delta \end{aligned}$$

This problem can be reformulated as the unconstrained problem for 2 groups as,

$$\max_\pi J(\pi) = \eta(\pi) - \alpha \left(\frac{\eta_1^S(\pi)}{\eta_1^D(\pi)} - \frac{\eta_2^S(\pi)}{\eta_2^D(\pi)} \right)^2 \tag{3.5}$$

Xu et al. (2024) solves this problem using standard policy gradient where it proposes a fair version of the PPO algorithm named ELBERT-PO. (Note:- In order to bridge the gap between standard fairness notions introduced in the last section and the fairness notion introduced in this section with respect to the ELBERT framework, we provide some additional notes in Appendix A).

3.4 Performative Prediction

As suggested before, long-term fairness literature does not take into account how the model itself can affect the data-generating distribution over the course of time. Thus, in this section, we shall explore the fundamental mathematical concepts of this statistical learning phenomenon which is termed as Performative Prediction.
When learning or evaluating a machine learning model, we assess its quality by how well it is able to predict an outcome variable y from features x on a given target population.

Definition 3.4.1. *We denote by $\Delta(\mathcal{X} \times \mathcal{Y})$ the simplex of probability distributions over the domain $\mathcal{X} \times \mathcal{Y}$. For a given distribution $D \in \Delta(\mathcal{X} \times \mathcal{Y})$, the risk of a model θ for a loss function ℓ is given as*

$$\mathit{Risk}(\theta, \mathcal{D}) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(\theta; z)].$$

The core conceptual device in the performative prediction framework is the notion of a *distribution map*

$$\mathcal{D} : \Theta \to \Delta(\mathcal{X} \times \mathcal{Y}),$$

which captures the dependence of the data-generating distribution $\mathcal{D}(\theta) \in \Delta(\mathcal{X} \times \mathcal{Y})$ on the model parameters θ.

Definition 3.4.2 (Sensitivity)**.** *We say that the distribution map $D(\cdot)$ is ϵ-sensitive if for all $\theta, \theta' \in \Theta$, it holds that*

$$\mathcal{W}(D(\theta), D(\theta')) \leq \epsilon \|\theta - \theta'\|_2,$$

where $\mathcal{W}$ denotes the Wasserstein-1 distance.

Sensitivity amounts to a Lipschitz condition and quantifies how far the distribution map is from being constant. The special case where $\epsilon = 0$ implies that $D(\theta) = D(\theta')$ for all $\theta, \theta' \in \Theta$, thereby recovering a static, non-performative setting.
Given the concept of a distribution map, it is natural to assess a model's risk with respect to the distribution that manifests from its deployment. This gives rise to two natural notions of optimality.
We define an equilibrium notion termed *performative stability*. It requires that the model appears optimal with respect to the distribution it induces. More formally,

Definition 3.4.3 (Performative Stability.)**.** *A model θ_{PS} is said to be* performatively stable *if it satisfies the fixed-point condition:*

$$\theta_{PS} \in \arg\min_{\theta \in \Theta} \mathit{Risk}(\theta, D(\theta_{PS})).$$

This means, based on data collected after the deployment of θ_{PS}, there is no incentive to deviate from the model. It is optimal on the static problem defined by $\mathcal{D}(\theta_{PS})$.

Definition 3.4.4 (Performative Optimality.)**.** *We say that a predictive model with parameters θ_{PO} is* performatively optimal *if it satisfies*

$$\theta_{PO} \in \arg\min_{\theta \in \Theta} Risk(\theta, D(\theta)).$$

In contrast to the stability condition, the objective in performative optimality is a moving target. A performatively optimal model must minimize the risk with respect to the distribution shift that arises as a consequence of its own deployment. In general, performatively stable points need not be optimal, and optimal points need not be stable. However, it always holds that

$$\mathrm{PR}(\theta_{\mathrm{PO}}) \leq \mathrm{PR}(\theta_{\mathrm{PS}})$$

for any performative optimum θ_{PO} and stable point θ_{PS}.
There exists a simple empirical check for stability: collect data under current conditions, solve the corresponding risk minimization problem, and verify whether the model is at least as good as the risk minimizer. In contrast, performative optimality admits no such straightforward empirical check. The data available prior to deployment generally do not inform us about the model's post-deployment performance.

Definition 3.4.5. *We define the* performative risk *of a model θ as*

$$PR(\theta) := Risk(\theta, D(\theta)). \tag{4}$$

Performatively optimal models minimize performative risk by definition.

Practitioners often employ a technique called Empirical Repeated Risk Minimization (ERRM) that involves the repeated retraining of the classifier per iteration on the newly collected samples to minimize performative risk. This technique is also backed by theory with results assuring convergence to performatively satble (and sometimes even performatively optimal) points (Brown et al., 2022).

3.5 Policy Gradient

Policy gradient methods aim to directly optimize the expected long-term return of an agent by adjusting the parameters of a stochastic policy in the direction of the performance gradient. For a policy π_θ parameterized by θ, the classic *Policy Gradient Theorem* (Sutton et al., 1999) provides a convenient expression for this gradient:

$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim d^\pi,\, a\sim\pi_\theta} \left[\nabla_\theta \log \pi_\theta(a|s)\, Q^\pi(s,a)\right],$$

where $Q^\pi(s,a)$ is the expected return when starting from state s, taking action a, and thereafter following policy π, and d^π denotes the (discounted) distribution of visited states (more on this later) under π.

The above expression underlies a family of algorithms including REINFORCE (Williams, 1992a) and actor-critic methods (Konda and Tsitsiklis, 2000), where the gradient is estimated from trajectories. These methods enjoy convergence guarantees to locally optimal policies under standard assumptions like smoothness and boundedness of the policy class. In a more recent and refined result, Agarwal et al. (2021) analyze the convergence rate of projected policy gradient ascent under a simple tabular setting with direct parameterization. They show that for an appropriately chosen step size, if π^* is the optimal policy, and $\{\pi^{(t)}\}_{t=1}^T$ be the sequence of policies generated by projected gradient ascent using the aforementioned gradient. Then for any starting state distribution ρ, the minimum sub-optimality over T steps satisfies:

$$\min_{t<T} \left\{ V^{\pi^*}(\rho) - V^{\pi^{(t)}}(\rho) \right\} \leq \varepsilon$$

as long as the number of iterations T is of the order $\mathcal{O}(1/\varepsilon^2)$.

The exact convergence bound depends on properties of the state and action spaces, the discount factor, and the distance between the initial and optimal policies. However, for our purposes, it suffices to remember that policy gradient enjoys a well-established $\mathcal{O}(1/\varepsilon^2)$ convergence rate to ε-optimal policies in this setting. Later in this work, we shall prove that incorporating additional performative terms in the gradient to account for the shifting rewards and transitions, does not affect our convergence and as a matter of fact, the rate also remains the same.

Chapter 4

Problem Formulation: Performative Reinforcement Learning

First introduced by Mandal et al. (2023), this setup is extremely essential for our project as our proposed algorithm (to be discussed in the next section) utilizes this framework as the base.

The notation $M(\pi)$ represents the corresponding Markov Decision Process (MDP), given by $M(\pi) = (S, A, P_\pi, r_\pi, \rho)$. Here, the transition probability function P_π and the reward function r_π vary based on the policy π. When an agent follows policy π while the underlying MDP is $M(\pi') = (S, A, P_{\pi'}, r_{\pi'}, \rho)$, the transition probabilities are determined accordingly. The probability of a trajectory $\tau = (s_k, a_k)_{k=0}^{\infty}$ under policy π in the MDP $M(\pi')$ is given by:

$$P(\tau) = \rho(s_0) \prod_{k=0}^{\infty} P_{\pi'}(s_{k+1} \mid s_k, \pi(s_k)).$$

We denote such a trajectory as $\tau \sim P_\pi^{\pi'}$.

Definition 4.0.1. *Given a policy π and a starting state distribution $\rho \in \Delta(S)$, the performative value function $V_\pi^\pi(\rho)$ is defined as:*

$$V_\pi^\pi(\rho) = \mathbb{E}_{\tau \sim P_\pi^\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_\pi(s_t, a_t) \mid \rho \right].$$

Now we look into the solution concepts of this setup,

Definition 4.0.2. *A policy π is performatively optimal if it maximizes the performative value function:*

$$\pi \in \arg\max_{\pi'} V_{\pi'}^{\pi'}(\rho).$$

We denote the performatively optimal policy as π^P. However, although π^P maximizes the performative value function, it may not be stable, meaning it is not necessarily optimal with respect to the changed environment $M(\pi^P)$.

Definition 4.0.3. *A policy π is performatively stable if it satisfies:*

$$\pi \in \arg\max_{\pi'} V_{\pi'}^{\pi}(\rho).$$

Lastly, we define the discounted state-action occupancy measure as follows:

Definition 4.0.4. *The **long-term discounted state-action occupancy measure** of a policy π in the MDP $M(\pi)$ is defined as:*

$$d_\pi(s,a) = \mathbb{E}_{\tau \sim P_\pi^\pi}\left[\sum_{k=0}^{\infty} \gamma^k 1\{s_k = s, a_k = a\} \mid \rho\right].$$

More details on the algorithms employed by Mandal et al. (2023) to attain performative stability is covered in Appendix B.

Based on the proposed formulation, Mandal et al. (2023) introduced several algorithmic approaches designed to attain performative stability or performative optimality within such environments (more in Appendix B). In contrast to their methodology, we pursue a more classical trajectory grounded in first-order optimization theory. Specifically, we aim to extend the standard policy gradient framework to a performative setting and establish theoretical guarantees for its convergence. In the subsequent chapter, we formally define a modified policy gradient ascent step that explicitly accounts for the performative dependence of both the reward and transition functions on the deployed policy. We then demonstrate that, despite these structural deviations from the classical setting, the proposed algorithm retains convergence guarantees to a performatively optimal policy under suitable assumptions.

Chapter 5

Methodology

5.1 Performative Projected Gradient Ascent

Recall the definition of Performative Value Function and Performatively Optimal Policy from Chapter 5. Our goal in this section is to propose a policy gradient ascent step which ensures convergence to a performatively optimal policy that will take into consideration the performative nature of the rewards and transitions. In order to achieve this, we need to assume a few things and recall some standard results from classic policy gradient. These assumptions are as follows:

Direct Parameterization. We consider the following class of stochastic policies where the policies are directly parameterized as

$$\pi_\theta(a \mid s) = \theta_{s,a}, \tag{5.1}$$

where $\theta \in \Delta(\mathcal{A})^{|\mathcal{S}|}$, i.e., θ satisfies $\theta_{s,a} \geq 0$ and $\sum_{a \in \mathcal{A}} \theta_{s,a} = 1$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.
Recall (4.0.4), in order to introduce our performative policy gradient, it is useful to define the *discounted state visitation distribution (state occupancy measure)* $d^\pi_{s_0}$ of a policy π as:

$$d^\pi_{s_0}(s) := (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \overset{\pi}{\Pr}(s_t = s \mid s_0), \tag{5.2}$$

where $\Pr^\pi(s_t = s \mid s_0)$ denotes the probability that the agent is in state s at time t under policy π, starting from initial state s_0. Overloading notation, we define the state-visitation distribution under an initial state distribution ρ as:

$$d^\pi_\rho(s) := \mathbb{E}_{s_0 \sim \rho}\left[d^\pi_{s_0}(s)\right],$$

where d^π_ρ is the *discounted state visitation distribution* when the starting state s_0 is drawn from distribution ρ, the starting state distribution.

The classic non-performative policy gradient in its standard functional form (see, e.g., (Williams, 1992b; Sutton et al., 1999)) is given by:

$$\nabla_\theta V^{\pi_\theta}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi_\theta},\ a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a \mid s) Q^{\pi_\theta}(s,a)\right]. \tag{5.3}$$

Furthermore, if we are working with a differentiable parameterization of $\pi_\theta(\cdot \mid s)$ that explicitly enforces $\pi_\theta \in \Delta(\mathcal{A})^{|\mathcal{S}|}$ for all θ, then the gradient can also be expressed using the advantage function:

$$\nabla_\theta V^{\pi_\theta}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi_\theta},\ a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a \mid s) A^{\pi_\theta}(s,a)\right]. \tag{5.4}$$

Note that the gradient expression in Equation (5.4) does not hold for the direct parameterization, whereas Equation (5.3) remains valid in general.
It is also important to note that for the direct policy parameterization where $\theta_{s,a} = \pi_\theta(a \mid s)$, the gradient of the value function with respect to the policy is given by:

$$\frac{\partial V^\pi(\mu)}{\partial \pi(a \mid s)} = \frac{1}{1-\gamma} d_\mu^\pi(s) Q^\pi(s,a), \tag{5.5}$$

as a direct consequence of Equation (5.3) (where μ represents the state distribution for all intermediate states). In particular, under this parameterization, we may write $\nabla_\pi V^\pi(\mu)$ instead of $\nabla_\theta V^{\pi_\theta}(\mu)$. This will be a useful tool even when studying the nature of convergence for the performative policy gradient ascent.
With the help of these standard results, we can now formally define the gradient of our performative value function with the help of this following lemma.

Lemma 5.1.1. *Let $V_\pi^\pi(\tau) = \mathbb{E}_{\tau\sim P_\theta}[\sum_{t=0}^{T-1}\gamma^t R_{\pi_\theta}(s_t,a_t)]$ denote the performative value function, with rewards R_{π_θ} and transitions P_{π_θ} depending on policy π_θ and trajectory τ. Then the gradient of the performative value function w.r.t θ is given by,*

$$\begin{aligned}
\nabla_\theta V_{\pi_\theta}^{\pi_\theta}(\tau) = \mathbb{E}_{\tau\sim P_\theta}&\left[\left(\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right)\cdot\left(\sum_{t=0}^{T-1}\gamma^t R_{\pi_\theta}(s_t,a_t)\right)\right] \\
+\mathbb{E}_{\tau\sim P_\theta}&\left[\left(\sum_{t=1}^{T}\nabla_\theta \log P_{\pi_\theta}(s_t|s_{t-1},a_{t-1})\right)\cdot\left(\sum_{t=0}^{T-1}\gamma^t R_{\pi_\theta}(s_t,a_t)\right)\right] \\
+\mathbb{E}_{\tau\sim P_\theta}&\left[\sum_{t=0}^{T-1}\gamma^t \nabla_\theta R_{\pi_\theta}(s_t,a_t)\right]
\end{aligned} \tag{5.6}$$

Proof. We denote,

$$f_\theta(\tau) = \sum_{t=0}^{T-1}\gamma^t R_{\pi_\theta}(s_t,a_t)$$

So, $\nabla_\theta V_{\pi_\theta}^{\pi_\theta}(\tau) = \nabla_\theta \mathbb{E}_{\tau\sim P_\theta}[f_\theta(\tau)]$

$$\begin{aligned}
\nabla_\theta \mathbb{E}_{\tau\sim P_\theta}[f_\theta(\tau)] &= \nabla_\theta \int P_\theta(\tau) f_\theta(\tau) d\tau \\
&= \int \nabla_\theta (P_\theta(\tau) f_\theta(\tau)) d\tau \quad \text{(swap integration with gradient)} \\
&= \int (\nabla_\theta P_\theta(\tau)) f_\theta(\tau) d\tau + \int P_\theta(\tau)(\nabla_\theta f_\theta(\tau)) d\tau \\
&\overset{(a)}{=} \int P_\theta(\tau)(\nabla_\theta \log P_\theta(\tau)) f_\theta(\tau) d\tau + \mathbb{E}_{\tau\sim P_\theta}[\nabla_\theta f_\theta(\tau)] \\
&= \mathbb{E}_{\tau\sim P_\theta}\left[(\nabla_\theta \log P_\theta(\tau)) f_\theta(\tau)\right] + \mathbb{E}_{\tau\sim P_\theta}[\nabla_\theta f_\theta(\tau)]
\end{aligned}$$

where (a) holds since $\nabla \log P_\theta(\tau) = \frac{\nabla P_\theta(\tau)}{P_\theta(\tau)}$

Now we have a sample-based estimator for $\nabla_\theta \mathbb{E}_{\tau\sim P_\theta}[f_\theta(\tau)]$. Let $\tau^{(1)}, \ldots, \tau^{(n)}$ be n empirical samples from P_θ (which are obtained by running the policy π_θ for n times, with T steps for each run). We can estimate the gradient of $\eta(\theta)$ by

$$\mathbb{E}_{\tau\sim P_\theta}\left[(\nabla_\theta \log P_\theta(\tau)) f_\theta(\tau)\right] \approx \frac{1}{n}\sum_{i=1}^{n} (\nabla_\theta \log P_\theta(\tau^{(i)})) f_\theta(\tau^{(i)}) \tag{5.7}$$

Here recall that ρ is used to denote the density of the distribution of s_0. It follows

$$\begin{aligned}
\log P_\theta(\tau) = \log \rho(s_0) + \log \pi_\theta(a_0 \mid s_0) + \log P_{\pi_\theta}(s_1|s_0, a_0) + \log \pi_\theta(a_1 \mid s_1) \\
+ \log P_{\pi_\theta}(s_2|s_1, a_1) + \cdots + \log P_{\pi_\theta}(s_t|s_{T-1}, a_{T-1})
\end{aligned}$$

Taking the gradient with respect to θ, we obtain

$$\begin{aligned}
\nabla_\theta \log P_\theta(\tau) = \nabla_\theta \log \pi_\theta(a_0 \mid s_0) + \nabla_\theta \log \pi_\theta(a_1 \mid s_1) + \cdots + \nabla_\theta \log \pi_\theta(a_{T-1} \mid s_{T-1}) \\
+ \nabla_\theta \log P_{\pi_\theta}(s_1|s_0, a_0) + \nabla_\theta \log P_{\pi_\theta}(s_2|s_1, a_1) \cdots + \nabla_\theta \log P_{\pi_\theta}(s_T|s_{T-1}, a_{T-1})
\end{aligned}$$

Hence, substituting in the main equation we get,

$$\begin{aligned}\nabla_\theta \mathbb{E}_{\tau\sim P_\theta}[f_\theta(\tau)] = \mathbb{E}_{\tau\sim P_\theta}&\left[\left(\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right)\cdot\left(\sum_{t=0}^{T-1}\gamma^t R_\pi(s_t,a_t|\theta)\right)\right]\\ &+\mathbb{E}_{\tau\sim P_\theta}\left[\left(\sum_{t=1}^{T}\nabla_\theta \log P_{\pi_\theta}(s_t|s_{t-1},a_{t-1})\right)\cdot\left(\sum_{t=0}^{T-1}\gamma^t R_\pi(s_t,a_t|\theta)\right)\right]\\ &+\mathbb{E}_{\tau\sim P_\theta}\left[\sum_{t=0}^{T-1}\gamma^t\nabla_\theta R_\pi(s_t,a_t|\theta)\right]\end{aligned}$$

□

We also present an alternate version of this gradient which replaces the trajectory τ explicitly with $d_{s_0}^{\pi_\theta}$,$\pi_\theta(a \mid s)$ and P_{π_θ} and present it in a similar fashion as equation (5.3):

$$\begin{aligned}\nabla_\theta V_\pi^\pi(s) = &\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^\pi}\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[\nabla_\theta \log \pi_\theta(a \mid s)\cdot Q(s,a)\Big]\\ &+\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^\pi}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P_\theta(s'|s,a)}\Big[\nabla_\theta \log P_{\pi_\theta}(s' \mid s,a)\cdot Q(s,a)\Big]\\ &+\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^\pi}\mathbb{E}_{a\sim\pi(\cdot|s)}\Big[\nabla_\theta R_{\pi_\theta}(s,a)\Big]\end{aligned}\tag{5.8}$$

We shall now formally define our projected gradient ascent step in the performative setup as follows:

Definition 5.1.1. *The projected gradient ascent algorithm updates the policy according to*

$$\theta^{(t+1)} = P_{\Delta(\mathcal{A})^{|\mathcal{S}|}}\left(\theta^{(t)} + \eta\nabla_\theta V^{(t)}(\mu)\right),\tag{5.9}$$

where $P_{\Delta(\mathcal{A})^{|\mathcal{S}|}}$ denotes the projection onto the product simplex $\Delta(\mathcal{A})^{|\mathcal{S}|}$ under the Euclidean norm. Here, $\Delta(\mathcal{A})^{|\mathcal{S}|}$ is the space of all stochastic policies with $\mathcal{A}$ denoting the action space and $|\mathcal{S}|$ denoting the cardinality of the state space.

In the aforementioned gradient ascent step, $\nabla_\pi V^{(t)}(\mu)$ denotes the gradient of the performative value function with state distribution μ at iteration t computed using (5.8). In the next section, we will provide mathematical guarantees to ensure asymptotic convergence to performative optimality of this proposed ascent step.

5.2 Theoretical Convergence Analysis

In this section we provide mathematical guarantees for convergence of our proposed policy gradient ascent step to a performatively optimal policy. Here, we use π instead of θ due

to direct parameterization.

To establish the asymptotic convergence of the proposed policy gradient ascent method to a performatively optimal policy, we proceed in three steps. First, we derive a bound on the sub-optimality gap of an arbitrary policy by formulating a performative analogue of the classical gradient domination lemma. This result enables us to relate the value sub-optimality directly to the norm of the policy gradient. Second, we demonstrate that the performative value function exhibits β-smoothness, a property crucial for analyzing the dynamics of gradient-based optimization algorithms. Finally, leveraging a standard result from (Beck, 2017) for smooth non-convex optimization, we obtain a convergence guarantee for the policy gradient method. Specifically, to ensure that the sub-optimality gap falls below a tolerance threshold ε, it suffices to perform $T = \mathcal{O}(1/\varepsilon^2)$ iterations of gradient ascent. This establishes that the algorithm achieves convergence to a stationary point of the performative objective at a rate comparable to that of standard policy gradient methods.

Before doing so, we start with two very standard assumptions.

Assumption 5.2.1. *The rewards are bounded by* -1 *and* 1*, i.e.,* $r_\pi \in [-1, 1]$

Assumption 5.2.2. *The reward and probability transition mappings are* $(\varepsilon_r, \varepsilon_p)$*-sensitive, i.e., the following holds for any two occupancy measures* d *and* d'*:*

$$\|r_\pi - r_{\pi'}\|_2 \leq \varepsilon_r \|\pi - \pi'\|_2, \quad \|P_\pi - P_{\pi'}\|_2 \leq \varepsilon_p \|\pi - \pi'\|_2$$

While the first assumption is extremely common in Reinforcement Learning Literature, the second one was introduced by Mandal et al. (2023) with the only difference in their version being that they used occupancy measure d instead of policy π. Now we present some results that will aid us in proving the main result.

Lemma 5.2.1 (Performative Performance Difference Lemma)**.** *For all policies* π, π_0 *and states* s_0

$$\begin{aligned} V_\pi^\pi(s_0) - V_{\pi'}^{\pi'}(s_0) = \gamma \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \Big[\big[\mathbb{E}_{s_1 \sim P_\pi(s_0,a_0)} V_\pi^\pi(s_1) \big] - \big[\mathbb{E}_{s_1 \sim P_{\pi'}(s_0,a_0)} V_{\pi'}^{\pi'}(s_1) \big] \Big] \\ + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \Big[A^{\pi'}(s_0, a_0) \Big] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \Big[r_\pi(s_0, a_0) - r_{\pi'}(s_0, a_0) \Big] \end{aligned} \tag{5.10}$$

Proof. For completeness, we provide the proof for this lemma in Appendix C. □

In the following results we denote the Value function for performatively optimal policy using V^* and the performative value function as V^π instead of V_π^π for simplicity. Also $\|\cdot\|$ denotes $\ell 2$-norm, $\|V^*(\mu)\|$ denote the $\ell 2$-norm of the optimal value function over the state distribution μ.

Lemma 5.2.2 (Performative Gradient Domination Lemma)**.** *For the direct policy parameterization (as in (5.1)), for all state distributions $\mu, \rho \in \Delta(\mathcal{S})$, we have:*

$$\begin{aligned} V^\star(\rho) - V^\pi(\rho) \leq \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} (\bar{\pi} - \pi)^\top \nabla_\pi V^\pi(\mu) \\ + \frac{M_1}{(1-\gamma)^2} \left\| \frac{d_\rho^{\pi^*}}{d_\mu^\pi} \right\|_\infty \|V^*(\mu)\| \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| \\ + \frac{M_2}{(1-\gamma)^2} \left\| \frac{d_\rho^{\pi^*}}{d_\mu^\pi} \right\|_\infty \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| \\ + \epsilon_r \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| + \gamma \epsilon_p \cdot \|V^*(\mu)\| \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| \end{aligned} \tag{5.11}$$

were M_1 and M_2 are bounds on the gradients of the log transitions and rewards.

Proof. For completeness, we provide the proof for this lemma in Appendix C. □

Lemma 5.2.3. *Let $\pi_\alpha := \pi_{\theta + \alpha u}$, and let $\tilde{V}(\alpha)$ be the corresponding value at a fixed state s_0, i.e.,*

$$\tilde{V}(\alpha) := V^{\pi_\alpha}(s_0).$$

Assume that

$$\begin{aligned} \left| \sum_{a \in \mathcal{A}} \frac{d\pi_\alpha(a \mid s_0)}{d\alpha} \Big|_{\alpha=0} \right| \leq C_1, \quad \left| \sum_{a \in \mathcal{A}} \frac{d^2\pi_\alpha(a \mid s_0)}{d\alpha^2} \Big|_{\alpha=0} \right| \leq C_2 \\ \left| \sum_{s \in \mathcal{S}} \frac{dP_\alpha(s \mid s_0, a_0)}{d\alpha} \Big|_{\alpha=0} \right| \leq T_1, \quad \left| \sum_{s \in \mathcal{S}} \frac{d^2P_\alpha(s \mid s_0, a_0)}{d\alpha^2} \Big|_{\alpha=0} \right| \leq T_2 \\ \left| \sum_{a \in \mathcal{A}} \frac{dr_\alpha(s_0, a)}{d\alpha} \Big|_{\alpha=0} \right| \leq R_1, \quad \left| \sum_{a \in \mathcal{A}} \frac{d^2r_\alpha(s_0, a)}{d\alpha^2} \Big|_{\alpha=0} \right| \leq R_2 \end{aligned}$$

Then

$$\max_{\|u\|_2 = 1} \left\| \frac{d^2\tilde{V}(\alpha)}{d\alpha^2} \Big|_{\alpha=0} \right\| \leq \frac{C_2}{1-\gamma} + 2C_1\beta_1 + C_2\beta_2$$

where $\beta_1 = \frac{\gamma}{(1-\gamma)^2}(C_1 + T_1) + \frac{R_1}{1-\gamma}$ and $\beta_2 = \frac{2\gamma^2}{(1-\gamma)^3}(C_1 + T_1)^2 + \frac{\gamma}{(1-\gamma)^2}(C_2 + 2C_1T_1 + T_2) + \frac{2\gamma R_1}{(1-\gamma)^2}(C_2 + 2C_1T_1 + T_2) + \frac{R_2}{1-\gamma} + \frac{\gamma C_1 R_1}{(1-\gamma)^2}$

Proof. For completeness, we provide the proof for this lemma in Appendix C. □

Lemma 5.2.4 (Smoothness For Direct Parameterization)**.** *For all starting states s_0,*

$$\left\| \nabla_\pi V^\pi(s_0) - \nabla_\pi V^{\pi'}(s_0) \right\|_2 \leq (2\sqrt{|\mathcal{A}|}\beta_1 + \beta_2) \left\| \pi - \pi' \right\|_2.$$

where $\beta_1 = \frac{\gamma}{(1-\gamma)^2}(\sqrt{\mathcal{A}} + T_1) + \frac{R_1}{1-\gamma}$ *and* $\beta_2 = \frac{2\gamma^2}{(1-\gamma)^3}(\sqrt{|\mathcal{A}|} + T_1)^2 + \frac{\gamma}{(1-\gamma)^2}(2T_1\sqrt{\mathcal{A}} + T_2) + \frac{2\gamma R_1}{(1-\gamma)^2}(2T_1\sqrt{|\mathcal{A}|} + T_2) + \frac{R_2}{1-\gamma} + \frac{\gamma R_1\sqrt{|\mathcal{A}|}}{(1-\gamma)^2}$

Proof. For completeness, we provide the proof for this lemma in Appendix C. □

Assumption 5.2.3. *Let* $f : \mathbb{R}^d \to (-\infty, \infty]$ *be a proper and closed function. Assume that* $\mathrm{dom}(f)$ *is convex and that* f *is* β*-smooth over* $\mathrm{int}(\mathrm{dom}(f))$.

Lemma 5.2.5 (Lemma 3 in (Ghadimi and Lan, 2016))**.** *Suppose that the above-mentioned assumption holds. Let the update be given by*

$$x^+ = x - \eta G_\eta(x),$$

then

$$\nabla f(x^+) \in \mathcal{N}_C(x^+) + \varepsilon(\eta\beta + 1)\mathbb{B}_2,$$

where $\mathbb{B}_2$ *is the unit* ℓ_2*-ball, and* $\mathcal{N}_C(x^+)$ *is the normal cone of the set* C *at* x^+.

Proof. The proof can be found in the cited paper. □

Lemma 5.2.6. *Let* $V^\pi(\mu)$ *be* β*-smooth in* π*. Define the gradient mapping*

$$G_\eta(\pi) = \frac{1}{\eta}\left(P_{\Delta(\mathcal{A})^{|\mathcal{S}|}}(\pi + \eta\nabla_\pi V^\pi(\mu)) - \pi\right),$$

and define the projected gradient update rule as $\pi^+ = \pi + \eta G_\eta(\pi)$*. If* $\|G_\eta(\pi)\|_2 \le \varepsilon$*, then*

$$\max_{\pi+\delta\in\Delta(\mathcal{A})^{|\mathcal{S}|},\, \|\delta\|_2\le 1} \delta^\top \nabla_\pi V^{\pi^+}(\mu) \le \varepsilon(\eta\beta + 1).$$

Proof. For completeness, we provide the proof for this lemma in Appendix C. □

We consider solving the following problem

$$\min_{x\in C} f(x) \tag{5.12}$$

with C being a nonempty closed and convex set.

Lemma 5.2.7 (Theorem 10.15 in (Beck, 2017))**.** *Suppose that Assumption E.1 holds and let* $\{x_k\}_{k\ge 0}$ *be the sequence generated by the gradient descent algorithm for solving the problem* (5.12) *with stepsize* $\eta = \frac{1}{\beta}$*. Then:*

1. *The sequence* $\{f(x_t)\}_{t\ge 0}$ *is non-increasing.*
2. $G_\eta(x_t) \to 0$ *as* $t \to \infty$.

3. $\min_{t=0,1,\ldots,T-1} \|G_\eta(x_t)\| \leq \sqrt{\frac{2\beta(f(x_0)-f(x^*))}{T}}$.

Proof. The proof of this theorem can be found in the cited paper. □

We now state and prove the main result of our project which ensures the convergence of the proposed policy gradient ascent step to performative optimality in asymptotic sense.

Theorem 5.2.1. *The projected gradient ascent algorithm on $V^\pi(\mu)$ with stepsize $\eta = \frac{1}{\beta}$ satisfies the following: For all distributions $\rho \in \Delta(S)$,*

$$\min_{t<T} \left\{ V^*(\rho) - V^{(t)}(\rho) \right\} \leq \varepsilon$$

whenever

$$T \geq \frac{32|S|\beta\kappa_\pi^2(1-\gamma)^3}{\left[\varepsilon(1-\gamma)^3 - \epsilon_1 - \epsilon_2\kappa_\pi\right]^2} = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$$

where $\beta = 2\sqrt{|\mathcal{A}|}\beta_1 + \beta_2$ (β_1 and β_2 are defined as they are in lemma 5.2.4) and $\epsilon_1 = 4\sqrt{|S|}(1-\gamma)^2\left(1-\gamma+\sqrt{|S|}\right)$, $\epsilon_2 = 2M\sqrt{|S|}(1-\gamma+\sqrt{|S|})$, $\kappa_\pi = \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty$, $M = \max\{M_1, M_2\}$

Proof. Recall the definition of the gradient mapping:

$$G_\eta(\pi) = \frac{1}{\eta}\left(P_{\Delta(\mathcal{A})^{|\mathcal{S}|}}(\pi + \eta\nabla_\pi V^{(t)}(\mu)) - \pi\right)$$

From lemma 5.2.4, we know that $V^\pi(s)$ is β-smooth for all states s, and hence $V^\pi(\mu)$ is also β-smooth, with β as defined above. Then from lemma 5.2.7 (standard smooth convex optimization result), using step-size $\eta = \frac{1}{\beta}$, we have:

$$\min_{t=0,\ldots,T-1} \|G_\eta(\pi^{(t)})\|_2 \leq \frac{\sqrt{2\beta(V^*(\mu) - V^{(0)}(\mu))}}{\sqrt{T}}$$

Then, from lemma 5.2.6, we obtain:

$$\min_{t=0,\ldots,T} \pi^{(t)} \max_{\substack{+\delta\in\Delta(\mathcal{A})^{|\mathcal{S}|} \\ \|\delta\|_2\leq 1}} \delta^\top \nabla_\pi V^{\pi^{(t+1)}}(\mu) \leq (1+\eta\beta)\frac{\sqrt{2\beta(V^*(\mu) - V^{(0)}(\mu))}}{\sqrt{T}}$$

Observe that:

$$\max_{\bar{\pi}\in\Delta(\mathcal{A})^{|\mathcal{S}|}} (\bar{\pi} - \pi)^\top \nabla_\pi V^\pi(\mu) = 2\sqrt{|\mathcal{S}|} \max_{\substack{\pi+\delta\in\Delta(\mathcal{A})^{|\mathcal{S}|} \\ \|\delta\|_2\leq 1}} \delta^\top \nabla_\pi V^\pi(\mu)$$

since $\|\bar{\pi} - \pi\|_2 \leq 2\sqrt{|\mathcal{S}|}$.

Using lemma 5.2.4 and $\eta\beta = 1$, we conclude:

$$\begin{aligned}\min_{t=0,\ldots,T} V^*(\rho) - V^{(t)}(\rho) \leq \frac{4\sqrt{|\mathcal{S}|}}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \frac{\sqrt{2\beta(V^*(\mu) - V^{(0)}(\mu))}}{\sqrt{T}} \\ + \frac{M_1}{(1-\gamma)^2} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \|V^*(\mu)\| \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| \\ + \frac{M_2}{(1-\gamma)^2} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| + \epsilon_r \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| \\ + \gamma\epsilon_p \cdot \|V^*(\mu)\| \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\|\end{aligned}$$

We know, $\|V^*(\mu)\| \leq \frac{\sqrt{|S|}}{1-\gamma}$ since the rewards are bounded by $[-1, 1]$. Also because of this, we can take $\epsilon_r = 2, \epsilon_p = 1$ and also since $\|\bar{\pi} - \pi\|_2 \leq 2\sqrt{|\mathcal{S}|}$ and we can set $M = \max(M_1, M_2)$, hence taking all of these into consideration and choosing $\kappa_\pi = \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty$, we arrive at,

$$\begin{aligned}\min_{t=0,\ldots,T} V^*(\rho) - V^{(t)}(\rho) \leq \frac{4\sqrt{|\mathcal{S}|}}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \frac{\sqrt{2\beta(V^*(\mu) - V^{(0)}(\mu))}}{\sqrt{T}} + 2|S| \frac{M}{(1-\gamma)^3} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \\ + 2\sqrt{|S|} \frac{M}{(1-\gamma)^2} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty + 2\sqrt{|S|} \left(2 + \frac{\gamma\sqrt{|S|}}{1-\gamma} \right)\end{aligned}$$

We can get our required bound of ϵ, if we set T such that,

$$\begin{aligned}\frac{4\sqrt{|\mathcal{S}|}}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \frac{\sqrt{2\beta(V^*(\mu) - V^{(0)}(\mu))}}{\sqrt{T}} + 2\sqrt{|\mathcal{S}|} \frac{M\left(\frac{\sqrt{|\mathcal{S}|}}{1-\gamma} + 1\right)}{(1-\gamma)^2} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \\ + 2\sqrt{|\mathcal{S}|} \left(2 + \frac{\gamma\sqrt{|\mathcal{S}|}}{1-\gamma} \right) \leq \varepsilon\end{aligned}$$

or, equivalently,

$$T \geq \frac{32|S|\beta\kappa_\pi^2(1-\gamma)^4 \left(V^*(\mu) - V^{(0)}(\mu)\right)}{\left[\varepsilon(1-\gamma)^3 - 2\sqrt{|S|}(1-\gamma)^2 \left(2(1-\gamma) + 2\sqrt{|S|}\right) - 2\sqrt{|S|}(1-\gamma+\sqrt{|S|})\kappa_\pi M\right]^2}$$

since we know, $V^*(\mu) - V^{(0)}(\mu) \leq \frac{1}{1-\gamma}$ and setting $\epsilon_1 = 4\sqrt{|S|}(1-\gamma)^2 \left(1 - \gamma + \sqrt{|S|}\right)$,

$\epsilon_2 = 2M\sqrt{|S|}(1-\gamma+\sqrt{|S|})$, we arrive at,

$$T \geq \frac{32|S|\beta\kappa_\pi^2(1-\gamma)^3}{\left[\varepsilon(1-\gamma)^3-\epsilon_1-\epsilon_2\kappa_\pi\right]^2} = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$$

where $\beta = 2\sqrt{|\mathcal{A}|}\beta_1 + \beta_2$ (β_1 and β_2 are defined as they are in lemma 5.2.4)

□

5.3 PePG: A Novel Performative Policy Gradient Algorithm

Having established the convergence of our proposed ascent step to the performatively optimal policy, we now proceed to formalize an algorithm that incorporates this update mechanism. To do so, we must first model the performative nature of both the reward and transition functions. For analytical tractability and interpretability, we adopt parametric forms grounded in well-studied statistical families.

Specifically, we model the performative rewards using a **linear location family**, where the observed reward is a linear perturbation of a baseline reward function. For the transition dynamics, we employ a **Gibbs family** (softmax-based parameterization), which models performative shifts via a log-linear modification of a baseline transition kernel. These formulations are chosen for their compatibility with gradient-based optimization and their ability to capture policy-induced distributional shifts in the environment.

Once again, assuming direct parameterization, the mathematical formulations of the performative reward and transition models are presented below.

$$R(s,a) \;=\; R_0(s,a) \;+\; \theta_{s,a}\,\psi(s,a) \tag{5.13}$$

$$T(s,a,s') \;=\; \frac{T_0(s,a,s')\;\exp\big(\theta_{s,a}\,\phi(s,a,s')\big)}{\sum_b T_0(s,a,b)\;\exp\big(\theta_{s,b}\,\phi(s,a,b)\big)} \tag{5.14}$$

Without counting for the performative updates, the simple policy gradient for direct parameterization can be given using (5.5) as

$$\frac{\partial J}{\partial \theta_{s,a}} = d(s)\cdot Q(s,a) \tag{5.15}$$

Where d is the state visitation count (occupancy measure) as defined in (5.2). Here we can omit $\frac{1}{1-\gamma}$ during optimization.

Algorithm 1: PePG: Two-Phase Performative Policy Gradient

Warmup Phase: Initialize $\pi^{(1)}$;
for $k \leftarrow 1$ **to** H **do**
  Collect trajectories with $\pi^{(k)}$;
  Estimate $R^{(k)}$, $T^{(k)}$, $d^{(k)}$, $Q^{(k)}$;
  Append to histories $(R^{(k)}, T^{(k)}, \pi^{(k)})$;
  Compute ordinary gradient $g^{(k)}$ using (5.15);
  Update $\pi^{(k+1)} \leftarrow \pi^{(k)} + \eta\, g^{(k)}$;
Performative Update Phase: Fit (R_0, ψ, T_0, ϕ) using histories through a NN;
for $k \leftarrow H+1$ **to** T **do**
  Collect trajectories with $\pi^{(k)}$;
  Estimate $d^{(k)}$, $Q^{(k)}$;
  Compute performative gradient $g^{(k)}_{\text{perf}}$ using (5.16) ;
  Update $\pi^{(k+1)} \leftarrow \pi^{(k)} + \eta\, g^{(k)}_{\text{perf}}$;

Accounting for the performative updates assumed to be modeled using the aforementioned linear scale families, the gradient using (5.8) becomes,

$$\frac{\partial J}{\partial \theta_{s,a}} = d(s) \left[Q(s,a) + \sum_{b \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \left(\phi(s,b,s') - \mathbb{E}_{s'}[\phi(s,b,s')] \right) Q(s,b) + \sum_{b \in \mathcal{A}} \psi(s,b) \right] \tag{5.16}$$

where $\mathbb{E}_{s'}[\phi(s,b,s')] = \sum_b T(s,a,b)\phi(s,a,b)$

Algorithm Design. We are now prepared to introduce our novel Performative Policy Gradient algorithm *PePG*, which draws conceptual inspiration from the performative gradient descent method, PerfGD (Izzo et al., 2021). The algorithm operates in two distinct phases: a *warmup phase* and a *performative phase*.

During the warmup phase, the agent follows a standard policy gradient procedure, deploying the current policy and collecting trajectories over a fixed number of iterations, denoted by H. At each iteration $t \leq H$, the algorithm records the empirically estimated reward matrix R_t, transition matrix T_t, and the deployed policy π_t, storing the triplet (R_t, T_t, π_t) for subsequent modeling.

Once the warmup phase concludes, the algorithm enters the performative phase. It first fits parametric models to estimate the performative reward and transition dynamics. Specifically, it estimates the baseline reward $\hat{R}_0$, reward shift coefficients $\hat{\psi}$, baseline transition kernel $\hat{T}_0$, and transition shift parameters $\hat{\phi}$, using a supervised learning approach—preferably leveraging deep neural networks to capture complex dependencies.

With these estimates in place, the algorithm proceeds to iteratively apply the perfor-

mative policy gradient updates, now accounting for how the policy-induced shifts in the environment affect both rewards and transitions. The pseudocode for PePG is presented as Algorithm 1.

The convergence of this algorithm need not be worried about as we have shown that the policy update with the gradient as (5.16) ensures convergence to a performatively optimal solution in aymptotic sense.

Chapter 6

Experimental Evaluation

6.1 Experimental Setup

In this section, we empirically assess the effectiveness of various repeated retraining methods and analyze how different parameters influence their convergence behavior. All experiments are performed within a grid-world environment originally introduced by (Triantafyllou et al., 2021). Below, we describe our modifications to this environment for simulating performative reinforcement learning dynamics.

Environment Design. The cost structure in the grid-world is defined as follows (see Fig. 6.1): when an agent visits either a blank cell or the start cell (denoted by S), a minor penalty of -0.01 is incurred by all agents. Visiting a goal cell (F) imposes a slightly higher cost of -0.02, while stepping into a hazard cell (H) results in a substantial penalty of -0.5. Furthermore, when any follower agent A_j chooses to intervene, it incurs an additional penalty of -0.05.

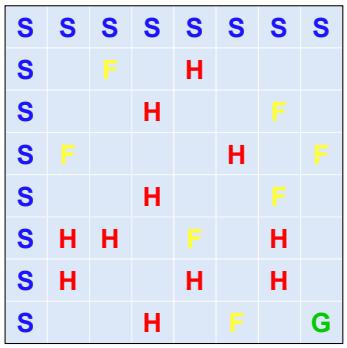


Figure 6.1: Design of the Gridworld Setup

Response Model. We model agent A_1 as a learning agent that undergoes repeated retraining. Initially, A_1 adopts an ε-optimal single-agent policy (with $\varepsilon = 0.1$) under the assumption that no other agent is actively influencing the environment. We generally use the direct parameterization, i.e., **Tabular policy** for A_1. After initialization, A_1 begins a retraining process using a specified learning method, such as gradient ascent.
The remaining agents respond to the policy of A_1 using a structured response mechanism. Given the current policy π_1 of A_1, we compute for each follower agent A_j (for $j = 2, \ldots, n+1$) the optimal Q-function $Q_j^{*|\pi_1}$, which is derived relative to a perturbed version of the original grid-world. These perturbations are introduced randomly and independently for each agent, simulating heterogeneous environmental perceptions or preferences.
We then define an average Q-function over the follower agents:

$$Q^{*|\pi_1}(s,a) = \frac{1}{n} \sum_{j=2}^{n+1} Q_j^{*|\pi_1}(s,a).$$

This averaged Q-function is used to determine the collective response policy π_2 via a Boltzmann softmax operator with temperature parameter β:

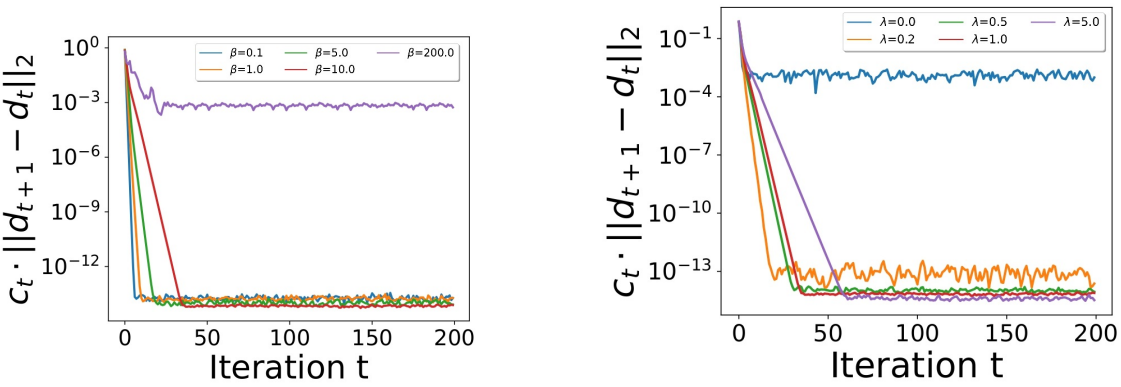
$$\pi_2(a_i \mid s) = \frac{\exp\left(\beta \cdot Q^{*|\pi_1}(s,a_i)\right)}{\sum_{a_j} \exp\left(\beta \cdot Q^{*|\pi_1}(s,a_j)\right)}.$$

The policy π_2 effectively represents the evolving environment faced by A_1, as shaped by the aggregate behavior of the follower agents. The parameter β modulates the responsiveness of this environment—lower values yield smoother, more exploratory behavior, while higher values lead to more deterministic reactions.

6.2 Evaluating Existing Performative Optimization Methods

In this section, we reproduce and evaluate the repeated optimization methods proposed by Mandal et al. (2023) to attain performative stability and subsequently optimality (more details about these algorithms in Appendix Appendix B). Also, the experiments were performed setting the maximum number of iterations to 200 unlike in (Mandal et al., 2023) where it was set to 1000.

Repeated Policy Optimization We begin by analyzing the setting where agent A_1 is granted full access to the updated reward and transition dynamics at each time step t. In this setup, agent A_1 selects a policy π_1^t, following which the follower agents respond using the softmax operator to produce a joint response policy π_2^t. Agent A_1 then observes the



(a) Convergence for varying β ($\lambda = 1$).

(b) Convergence for varying λ ($\beta = 10$).

Figure 6.2: Convergence behavior of Repeated Policy Optimization (RPO) under different parameter regimes.

updated environment, characterized by the new transition matrix P^t and reward function r^t.

With access to (P^t, r^t), agent A_1 proceeds to solve the optimization problem described in Equation (B.2) (From Appendix B), yielding a new occupancy measure d_1^{t+1}. The policy π_1^{t+1} for the next round is then obtained by normalizing d_1^{t+1} according to Equation (B.1) (From Appendix B).

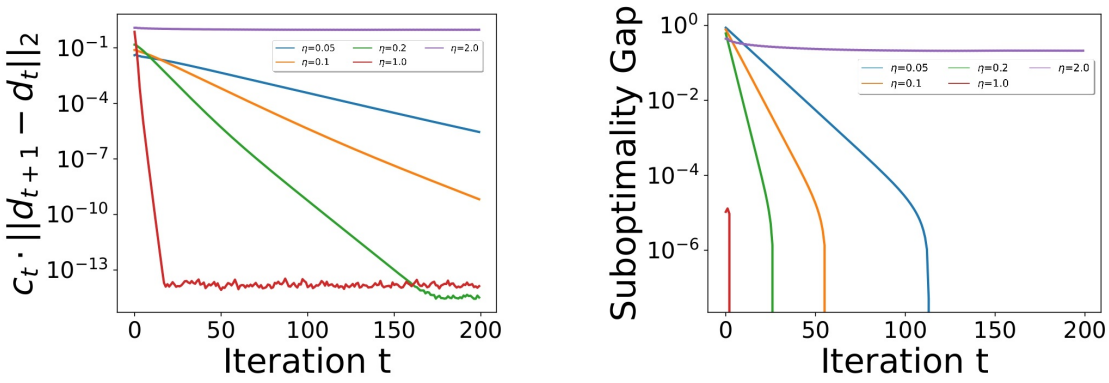
Figure 6.2b illustrates the convergence behavior of repeated policy optimization under different values of the response smoothness parameter β, while keeping the regularization parameter λ fixed at 1. We observe that for large values of β (e.g., $\beta = 200$), the response function becomes too sharp, causing the learning dynamics to destabilize and the algorithm to fail to converge.

In contrast, Figure 6.2a fixes $\beta = 10$ and varies the regularization strength λ. The results indicate that the algorithm converges only for sufficiently large values of λ, and the convergence speed improves as λ increases.

Repeated Gradient Ascent We now turn to a more incremental variant of policy optimization, where agent A_1 adopts a repeated gradient ascent approach instead of solving the full optimization problem in each iteration. As before, A_1 observes the response policy π_2^t at time t, and thereby reconstructs the updated transition dynamics P^t and reward function r^t.

Instead of full optimization, A_1 takes a projected gradient ascent step (B.3) (from Appendix B) to compute the updated occupancy measure d_1^{t+1}.

Figure 6.3a presents the convergence behavior of this approach for varying values of the step-size η. When η is small (e.g., $\eta \leq \mathcal{O}(1/\lambda)$), the method converges reliably to a stable point. However, larger step-sizes cause divergence. Notably, convergence is faster when η



(a) Convergence for varying step-size η.

(b) Suboptimality gap for varying η.

Figure 6.3: Performance of Repeated Gradient Ascent (RGA) for different step-sizes.

is chosen close to $1/\lambda$.

Since this method does not fully optimize the objective, we also track the *suboptimality gap*, defined as the difference between the objective value of the best feasible solution (with respect to M^t) and the current solution d_1^t, normalized by the former. The trend in suboptimality gap, as a function of η, mirrors that of convergence speed in Figure 6.3a, reaffirming the effectiveness of moderate step sizes.

6.3 Evaluating Vanilla PG and PePG

The figures in this section are plotted with respect to different values of parameter ν which will be discussed below.

Vanilla Policy Gradient The proposed implementation methodology formalizes a principled, optimization-based policy update framework. The procedure begins with the initialization of the environment and the involved agents, including a primary agent whose policy is updated iteratively, and a fixed or reference agent that influences the environment. Given access to the initial state distribution ρ and discount factor γ, the algorithm proceeds by estimating the discounted state-action occupancy measure $d^\pi(s, a)$, as well as computing the environment-specific Q-function $Q(s, a)$ and expected utility matrix.

The policy gradient is then computed using (5.15) as

$$DU(s, a) = d(s) \cdot Q(s, a),$$

which represents the direction of steepest ascent in expected return. The updated policy is obtained by solving a regularized proximal objective that penalizes deviations from a

target policy π_{target}, defined as:

$$\pi_{\text{target}} = \pi_{\text{last}} - \eta \cdot DU - \nu(1 + \log \pi_{\text{last}}),$$

where η is the learning rate and ν is a regularization constant. The additive log-barrier term $-\nu \log \pi_{\text{last}}$ serves a dual purpose: it enforces strict positivity of the policy variables, and it acts as an implicit barrier preventing updates from approaching the boundary of the probability simplex. Near the simplex edges, where some $\pi(s, a)$ approach zero, the term $\log \pi(s, a)$ diverges to $-\infty$, thereby inducing an infinite cost in the optimization objective and effectively repelling solutions away from the boundary. This ensures numerical stability and preserves the feasibility of the updated policy.

The full convex optimization problem is given by:

$$\min_{\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}} \|\pi - (\pi_{\text{last}} - \eta \cdot DU - \nu(1 + \log \pi_{\text{last}}))\|_2^2$$

$$\text{subject to} \quad \sum_a \pi(s, a) = 1 \quad \forall s, \quad \pi(s, a) \geq 0.$$

This is solved using the SCS convex solver with high numerical precision. The raw output is post-processed to ensure strictly positive values and proper normalization using a small smoothing constant $\delta = 10^{-7}$, as follows:

$$\hat{\pi}(s, a) = \frac{\pi(s, a) + \delta}{\sum_{a'} \pi(s, a') + |\mathcal{A}|\delta}.$$

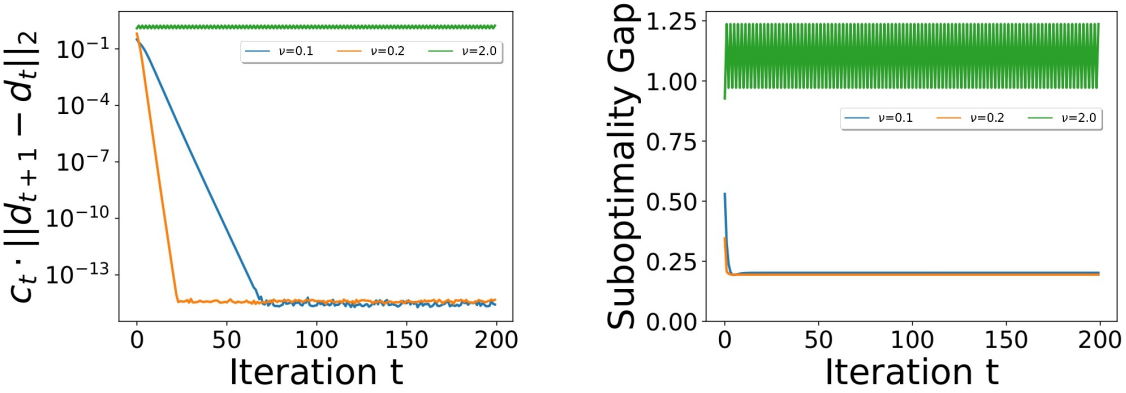
The updated policy is then assigned to the agent, and the new occupancy measure is recomputed and compared to the previous one using the relative ℓ_2-norm. To evaluate performance, the suboptimality gap is computed via a constrained maximization over valid occupancy measures d, subject to discounted flow conservation constraints. This gap is quantified as:

$$\text{gap} = \max\left(\frac{V^* - V^\pi}{|V^*|}, 0\right),$$

where V^* denotes the optimal value under a performative-aware occupancy measure, and V^π is the expected return of the current policy, thereby providing a normalized measure of convergence and solution quality.

Although a formal theoretical guarantee of stability for the policy gradient method remains elusive, empirical evidence as in Figure 6.4a suggests that the algorithm consistently converges to a stable solution. However, this stability does not imply optimality. Since the method does not explicitly account for performative feedback in the policy space, it converges to a suboptimal solution. This is evidenced by the suboptimality gap, which stabilizes around a value of 0.2022 (as suggested by Figure 6.4b), converging at a notably

rapid rate.



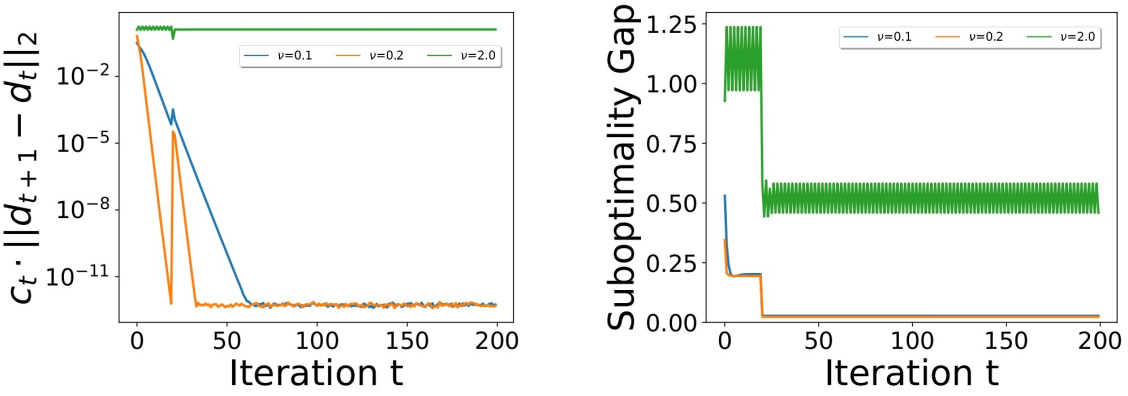
(a) Stability of the policy gradient method.

(b) Convergence of the suboptimality gap.

Figure 6.4: Empirical behavior of the policy gradient method: While it consistently attains stability (left), the method converges to a suboptimal policy due to a lack of performative-aware updates (right).

PePG: Performative Policy Gradient The Performative Policy Gradient (PePG) algorithm closely resembles the standard policy gradient method, with the primary distinction lying in its incorporation of the performative framework introduced in Section 4.3. As with the vanilla approach, PePG employs projection and includes a log-barrier term to enforce simplex constraints and encourage interior solutions. The algorithm proceeds in two distinct phases. During the *warmup phase* (lasting approximately 20 iterations), the agent updates its policy using standard policy gradient steps while simultaneously recording the reward matrices, transition dynamics, and policy snapshots into historical buffers. At the conclusion of this phase, we compute point estimates of $\widehat{R}_0$, $\widehat{\psi}$, and $\widehat{\phi}$.
Subsequently, the *performative phase* begins. Here, policy updates incorporate the performative gradient terms, and a performative adjustment is applied to the occupancy measure d using the current policy and the previously estimated quantities. This enforced performative update encourages exploratory behavior aligned with the performative environment's evolution and is designed to bridge the gap between theoretical assumptions and practical constraints—particularly relevant given the limitations of the grid-world setup, as discussed in Section 5.3.
Figure 6.5a illustrates that the stability observed in vanilla policy gradient is largely preserved. Although a transient instability is observed at the transition between phases, attributable to the sudden change in gradient scale, the algorithm quickly regains stability and eventually satisfies performative stability. More notably, the suboptimality gap exhibits significant improvement. While it mirrors the vanilla behavior during the warmup phase, a sharp decline is observed once performative updates are introduced. The gap

converges to a small, non-zero value (approximately 2.7×10^{-3}), which we attribute to violations of ideal theoretical assumptions in the experimental setup. Nevertheless, this outcome can be interpreted as *approximate optimality* in practice.



(a) Policy stability over iterations.

(b) Suboptimality gap for PePG.

Figure 6.5: Performance of the PePG algorithm: The left plot demonstrates that performative stability is preserved, while the right plot shows that the suboptimality gap drops significantly after the performative phase begins.

We summarize the results reported in this section through the following table. Here "Perf. Stab." is supposed to indicate whether performative stability was attained while "Perf. Opt." is supposed to indicate whether performative optimality was attained (and the type, i.e., absolute or approximate). "# iters" in both cases is supposed to state the number of iterations in which the stability or optimality (or convergence to sub-optimality) was attained and "param" in both cases is supposed to state the best parameter for which the values are reported in the table. "sub-gap" is supposed to denote the sub-optimality gap finally attained by the algorithm.

Table 6.1: Performance Comparison of Algorithms on Stability and Optimality

Algorithm	Perf. Stab.	# Iters	Param	Perf. Opt.	Sub-Gap	# Iters	Param
RPO (β)	Attained	10	$\beta = 0.1$	–	–	–	–
RPO (λ)	Attained	40	$\lambda = 1.0$	–	–	–	–
RGA	Attained	20	$\eta = 1.0$	Absolute	0	5	$\eta = 1.0$
Vanilla PG	Attained	25	$\nu = 0.2$	No	0.2022	6	$\nu = 0.2$
PePG	Attained	40	$\nu = 0.2$	Approx.	2.7×10^{-3}	10	$\nu = 0.2$

6.4 Challenges

While the proposed Performative Policy Gradient (PePG) algorithm demonstrates strong empirical performance and near-performative optimality, there are two significant challenges that hinder convergence to true optimality.

Lack of Trajectory-Based Sampling. One of the primary limitations arises from the structural design of the environment, which does not support direct sampling of trajectories. Instead, the learning process is driven by the repeated update of pre-initialized matrices representing the reward function, transition probabilities, occupancy measures, policy distributions, and Q-value estimates. While this approach facilitates modular experimentation and faster iteration, it introduces bias in the gradient estimates as it lacks the stochasticity and variance inherent in real trajectory-based rollouts. This structural mismatch compromises the fidelity of the policy gradient updates, especially in non-stationary performative settings, thereby contributing to the deviation from full performative optimality.

Violation of Smoothness Assumptions. A second fundamental challenge is the violation of theoretical regularity assumptions required for convergence guarantees. In particular, the reward distribution in the grid-world environment is defined over a finite and discrete set of values, violating the assumed $\mathcal{C}^2$-smoothness of the reward mapping. Although the state-action space itself is discrete, many theoretical results underpinning performative optimality and stability assume differentiability of the reward function with respect to the agent's policy parameters. This discontinuity in the reward structure results in non-smooth objective surfaces, which in turn hampers the efficacy of gradient-based optimization techniques and further contributes to the observed sub-optimality gap.
Taken together, these two challenges—structural limitations on sampling and violation of smoothness assumptions—highlight important practical obstacles that must be addressed in future work in order to bridge the gap between empirical performance and theoretical guarantees of performative optimality.

Chapter 7

Discussions and Future Works

7.1 Main Contributions

We have two major contributions towards the study of long-term fairness in performative reinforcement learning:

1. **A step forward to achieve long-term fairness.** It emphasizes the critical role of *performative prediction* in achieving long-term fairness in reinforcement learning. While most existing approaches assume a static environment, this work brings to light how the deployed policy can influence future data distributions—a hallmark of performativity. By explicitly modeling and accounting for this feedback loop, we underscore that fairness guarantees derived in static settings may fail under dynamic conditions. This serves as a foundational motivation for integrating performative modeling assumptions in algorithm design, especially when fairness is a long-term concern.

2. **Novel algorithm design.** It introduces a novel **Performative Policy Gradient (PePG)** algorithm tailored for the performative reinforcement learning setup. The proposed method is shown to attain *performative stability*, albeit at a slower rate than the classic policy gradient due to a transient instability caused by the shift from the warm-up phase to the performative update phase, which introduces a sharp change in the gradient scale. Although this stability property cannot yet be established theoretically, it is consistently observed in empirical evaluations. More importantly, the PePG algorithm shows *approximate performative optimality* even in a discrete, non-smooth grid-world environment that is not well-suited for standard policy gradient techniques—wherein the vanilla policy gradient method fails to achieve any form of optimality and instead converges to a non-trivial suboptimality gap.

7.2 Future Works

To conclude, this work presents an initial step toward designing reinforcement learning algorithms capable of addressing long-term fairness constraints through a performative lens. We introduce *PePG*, a performative policy gradient algorithm motivated by recent literature (Xu et al., 2024), which highlights the compatibility between policy gradient methods and long-term fairness objectives. Empirical results demonstrate that PePG outperforms the vanilla policy gradient in a performative grid-world setting. However, the development process revealed key implementation challenges, notably the lack of trajectory-based sampling and violations of theoretical smoothness assumptions in environment proposed by Mandal et al. (2023). These limitations motivate several promising directions for future research. First, there is a need to construct more realistic benchmark environments for evaluating performative RL algorithms, moving beyond synthetic settings such as grid-worlds. Leveraging real-world datasets—such as the *GiveMeSomeCredit* dataset from Kaggle, which has been previously used for performative prediction tasks—offers a natural path forward. Additionally, we aim to extend PePG to handle parameterized policies beyond direct tabular representations, including softmax policies. Future work also includes developing a natural policy gradient variant with variance reduction techniques, followed by an actor-critic extension of PePG. That said, incorporating Long-Term fairness in *PePG* remains our "long-term" goal. Lastly, we plan to investigate the performance of PePG relative to methods in the Posterior Sampling for Reinforcement Learning (PSRL) framework within performative settings.

Appendix A

Bridging the Gap between Classic Fairness and SD-MDP

A.1 Fairness notions with the supply and demand formulation

Demographic Parity (DP). The well-being of a group g in DP is defined as

$$\Pr[\hat{Y} = 1 \mid G = g] \;=\; \frac{\Pr[\hat{Y} = 1,\, G = g]}{\Pr[G = g]},$$

and DP requires these probabilities to be equal across groups. In practice, given a dataset, one computes for each group g a ratio $\frac{S_g}{D_g}$, where S_g is the number of samples with $\{\hat{Y} = 1, G = g\}$ (e.g., accepted individuals in group g) and D_g is the total number of samples with $\{G = g\}$.

Equal Opportunity (EO). EO requires

$$\Pr[\hat{Y} = 1 \mid G = g,\, Y = 1] \;=\; \frac{\Pr[\hat{Y} = 1,\, Y = 1,\, G = g]}{\Pr[Y = 1,\, G = g]},$$

to be equal across groups. In practice, the well-being of group g is $\frac{S_g}{D_g}$, where S_g counts $\{\hat{Y} = 1,\, Y = 1,\, G = g\}$ (qualified and accepted) and D_g counts $\{Y = 1,\, G = g\}$ (qualified individuals in group g).

Equality of Discovery Probability. This notion is a special case of EO, often used in settings like predictive policing. It requires that, given a person actually committed a crime ($Y = 1$), the probability of being apprehended ($\hat{Y} = 1$) should be independent of group identity.

Equalized Odds. Equalized Odds requires both the True Positive Rate (TPR) and False Positive Rate (FPR) to be equal across groups. That is, TPR $= \Pr[\hat{Y} = 1 \mid Y = 1, G = g]$ and FPR $= \Pr[\hat{Y} = 1 \mid Y = 0, G = g]$ should not vary with g. Concretely, TPR for group g is given by $\frac{S_g^T}{D_g^T}$ where $S_g^T = \{\hat{Y} = 1, Y = 1, G = g\}$ and $D_g^T = \{Y = 1, G = g\}$; FPR for group g is $\frac{S_g^F}{D_g^F}$ where $S_g^F = \{\hat{Y} = 1, Y = 0, G = g\}$ and $D_g^F = \{Y = 0, G = g\}$.

Extending Equalized Odds to Sequential Settings. In a sequential scenario, define for each group g two supply-demand pairs: (D_g^T, S_g^T) and (D_g^F, S_g^F). Their long-term cumulative quantities under a policy π are

$$\eta_{D,g}^T(\pi) = \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t \, D_g^T(s_t, a_t)\Big], \quad \eta_{S,g}^T(\pi) = \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t \, S_g^T(s_t, a_t)\Big],$$

and similarly $\eta_{D,g}^F(\pi)$, $\eta_{S,g}^F(\pi)$ for FPR. Define the TPR bias

$$b^T(\pi) \;=\; \max_{g \in G}\!\left(\frac{\eta_{S,g}^T(\pi)}{\eta_{D,g}^T(\pi)}\right) \;-\; \min_{g \in G}\!\left(\frac{\eta_{S,g}^T(\pi)}{\eta_{D,g}^T(\pi)}\right),$$

and the FPR bias $b^F(\pi)$ analogously. Enforcing $b^T(\pi) \le \epsilon$ and $b^F(\pi) \le \epsilon$ yields the constrained objective

$$\max_{\pi} \; \eta(\pi) \quad \text{subject to} \quad b^T(\pi) \le \epsilon, \; b^F(\pi) \le \epsilon.$$

In practice, one often relaxes these hard constraints into a regularized objective:

$$J(\pi) \;=\; \eta(\pi) - \alpha \, b^T(\pi)^2 - \alpha \, b^F(\pi)^2,$$

where $\alpha > 0$ balances return and fairness. The gradients of these ratios remain tractable and can be computed with policy gradient methods.

Accuracy Parity. Accuracy Parity defines the well-being of group g as $\Pr[\hat{Y} = Y \mid G = g] = \frac{\Pr[\hat{Y}=Y, G=g]}{\Pr[G=g]}$, which is the fraction of correctly predicted samples in group g. In practice, this is $\frac{S_g}{D_g}$ where $S_g = \{\hat{Y} = Y, G = g\}$ and $D_g = \{G = g\}$.

Appendix B

Details on Repeated Optimization Methods in Performative RL

B.1 Repeated Policy Optimization

For a given occupancy measure d, the corresponding policy π_d with occupancy measure d is defined as:

$$\pi_d(a \mid s) = \begin{cases} \frac{d(s,a)}{\sum_b d(s,b)}, & \text{if } \sum_a d(s,a) > 0, \\ \frac{1}{A}, & \text{otherwise.} \end{cases} \tag{B.1}$$

Using this definition, according to (Mandal et al., 2023) a **performatively stable occupancy measure** is the optimal solution to the following problem:

$$\begin{aligned} &\max_{d \geq 0} \sum_{s,a} d(s,a) r_d(s,a) \\ &s.t., \sum_a d(s,a) = \rho(s) + \gamma \sum_{s',a} d(s',a) P_d(s',a,s), \quad \forall s. \end{aligned}$$

Even though this optimization problem is guaranteed to have a stable solution, it is not clear that repeatedly optimizing this objective converges to such a point.

The corresponding regularized version of this objective is,

$$\max_{d \geq 0} \sum_{s,a} d(s,a) r_d(s,a) - \frac{\lambda}{2} \|d\|_2^2$$
$$s.t., \sum_a d(s,a) = \rho(s) + \gamma \sum_{s',a} d(s',a) P_d(s',a,s), \quad \forall s. \tag{B.2}$$

They show that a stable solution guarantees *approximate stability* with respect to the original unregularized objective.
Here, $\lambda > 0$ is a constant that determines the **strong concavity** of the objective.
They show that repeatedly solving the problem converges to a stable point.

Assumption B.1.1. *The reward and transition probability mappings are $(\varepsilon_r, \varepsilon_p)$-sensitive, i.e., the following holds for any two occupancy measures d and d':*

$$\|r_d - r_{d'}\|_2 \leq \varepsilon_r \|d - d'\|_2, \quad \|P_d - P_{d'}\|_2 \leq \varepsilon_p \|d - d'\|_2$$

Theorem B.1.1 (Theorem 1 from Mandal et al. (2023))**.** *Suppose Assumption 4 holds with*

$$\lambda > \frac{12 S^{3/2}(2\varepsilon_r + 5S\varepsilon_p)}{(1-\gamma)^4}. \quad \textit{And Let} \quad \mu = \frac{12 S^{3/2}(2\varepsilon_r + 5S\varepsilon_p)}{\lambda(1-\gamma)^4}.$$

Then for any $\delta > 0$, we have

$$\|d_t - d^*\|_2 \leq \delta \quad \forall t \geq \frac{2}{1-\mu} \ln\left(\frac{2}{\delta(1-\gamma)}\right).$$

B.2 Repeated Gradient Ascent

Now we extend the results to the case where the learner does not fully solve the optimization problem (B.2) at each iteration. Instead, the learner performs a single gradient step in response to the updated environment at every iteration.
Let $\mathcal{C}_t$ denote the set of occupancy measures that are consistent with the current transition dynamics P_t. Specifically, define

$$\mathcal{C}_t = \left\{ d \ : \ \sum_a d(s,a) = \rho(s) + \gamma \sum_{s',a} d(s',a) P_t(s',a,s) \ \ \forall s, \quad \text{and} \quad d(s,a) \geq 0 \ \ \forall s,a \right\}$$

In this setting, the gradient ascent algorithm proceeds by first taking a gradient step on the regularized objective function:

$$r_t^\top d - \frac{\lambda}{2} \|d\|_2^2,$$

followed by a projection of the updated point onto the set $\mathcal{C}_t$. This gives the update rule:

$$d_{t+1} = \text{Proj}_{\mathcal{C}_t}\left(d_t + \eta(r_t - \lambda d_t)\right) = \text{Proj}_{\mathcal{C}_t}\left((1-\eta\lambda)d_t + \eta r_t\right) \tag{B.3}$$

Here, $\text{Proj}_{\mathcal{C}}(v)$ denotes the Euclidean projection of vector v onto the set $\mathcal{C}$, i.e., the point in $\mathcal{C}$ closest to v in ℓ_2-norm.
We now show that applying projected gradient ascent with a suitable step size η leads to convergence to a stable point.

Theorem B.2.1 (Theorem 2 from Mandal et al. (2023))**.** *Choose step-size* $\eta = \frac{1}{\lambda}$ *and let,*

$$\lambda \geq \max\left\{4\varepsilon_r,\ 2S,\ \frac{20\gamma^2 S^{3/2}(\varepsilon_r + \varepsilon_p)}{(1-\gamma)^2}\right\} \quad \textit{And} \quad \mu = \sqrt{\frac{64\gamma^2\varepsilon_p^2}{(1-\gamma)^4}\left(1 + \frac{30\gamma^4 S^2}{(1-\gamma)^4}\right)}.$$

Suppose Assumption 4 holds and that $\varepsilon_p < \min\left\{\frac{\gamma}{S^3},\ \frac{(1-\gamma)^4}{100\gamma^3 S}\right\}$*. Then, for any* $\delta > 0$*, we have*

$$\|d_t - d^*\|_2 \leq \delta \quad \forall t \geq \frac{1}{1-\mu}\ln\left(\frac{2}{\delta(1-\gamma)}\right).$$

Appendix C

Missing Proofs from Section 5.2

Lemma 5.2.1 (Performative Performance Difference Lemma)**.** *For all policies π, π_0 and states s_0*

$$V_\pi^\pi(s_0) - V_{\pi'}^{\pi'}(s_0) = \gamma \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \Big[\big[\mathbb{E}_{s_1 \sim P_\pi(s_0,a_0)} V_\pi^\pi(s_1)\big] - \big[\mathbb{E}_{s_1 \sim P_{\pi'}(s_0,a_0)} V_{\pi'}^{\pi'}(s_1)\big] \Big] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \Big[A^{\pi'}(s_0, a_0) \Big] + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \Big[r_\pi(s_0, a_0) - r_{\pi'}(s_0, a_0) \Big]$$

Proof.

$$\begin{aligned}
V_\pi^\pi(s_0) - V_{\pi'}^{\pi'}(s_0) &= V_\pi^\pi(s_0) - \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \big[r_{\pi'}(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P_{\pi'}(s_0,a_0)} V_{\pi'}^{\pi'}(s') \big] \\
&\quad + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \big[r_{\pi'}(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P_{\pi'}(s_0,a_0)} V_{\pi'}^{\pi'}(s') \big] - V_{\pi'}^{\pi'}(s_0) \\
&= \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \big[r_\pi(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P_\pi(s_0,a_0)} V_\pi^\pi(s') \big] \\
&\quad - \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \big[r_{\pi'}(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P_{\pi'}(s_0,a_0)} V_{\pi'}^{\pi'}(s') \big] \\
&\quad + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \big[r_{\pi'}(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P_{\pi'}(s_0,a_0)} V_{\pi'}^{\pi'}(s') \big] - V_{\pi'}^{\pi'}(s_0) \\
&= \gamma \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \big[\big[\mathbb{E}_{s_1 \sim P_\pi(s_0,a_0)} V_\pi^\pi(s_1)\big] - \big[\mathbb{E}_{s_1 \sim P_{\pi'}(s_0,a_0)} V_{\pi'}^{\pi'}(s_1)\big] \big] \\
&\quad + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \big[r_{\pi'}(s_0, a_0) + \gamma \mathbb{E}_{s' \sim P_{\pi'}(s_0,a_0)} V_{\pi'}^{\pi'}(s') \big] - V_{\pi'}^{\pi'}(s_0) \\
&\quad + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \big[r_\pi(s_0, a_0) - r_{\pi'}(s_0, a_0) \big] \\
&= \gamma \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \big[\big[\mathbb{E}_{s_1 \sim P_\pi(s_0,a_0)} V_\pi^\pi(s_1)\big] - \big[\mathbb{E}_{s_1 \sim P_{\pi'}(s_0,a_0)} V_{\pi'}^{\pi'}(s_1)\big] \big] \\
&\quad + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \big[Q_{\pi'}^{\pi'}(s_0, a_0) - V_{\pi'}^{\pi'}(s_0) \big] \\
&\quad + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \big[r_\pi(s_0, a_0) - r_{\pi'}(s_0, a_0) \big] \\
&= \gamma \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \big[\big[\mathbb{E}_{s_1 \sim P_\pi(s_0,a_0)} V_\pi^\pi(s_1)\big] - \big[\mathbb{E}_{s_1 \sim P_{\pi'}(s_0,a_0)} V_{\pi'}^{\pi'}(s_1)\big] \big] \\
&\quad + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \big[A^{\pi'}(s_0, a_0) \big] \quad + \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)} \big[r_\pi(s_0, a_0) - r_{\pi'}(s_0, a_0) \big]
\end{aligned}$$

□

Lemma 5.2.2. *For the direct policy parameterization (as in (5.1)), for all state distributions $\mu, \rho \in \Delta(\mathcal{S})$, we have:*

$$
\begin{aligned}
V^*(\rho) - V^\pi(\rho) \leq & \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} (\bar{\pi} - \pi)^\top \nabla_\pi V^\pi(\mu) \\
& + \frac{M_1}{(1-\gamma)^2} \left\| \frac{d_\rho^{\pi^*}}{d_\mu^\pi} \right\|_\infty \|V^*(\mu)\| \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| \qquad \text{(C.1)} \\
& + \frac{M_2}{(1-\gamma)^2} \left\| \frac{d_\rho^{\pi^*}}{d_\mu^\pi} \right\|_\infty \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| \\
& + \epsilon_r \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| + \gamma \epsilon_p \cdot \|V^*(\mu)\| \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| \qquad \text{(C.2)}
\end{aligned}
$$

were M_1 and M_2 are bounds on the gradients of the log transitions and rewards.

Proof. We get using lemma 4.2.1,

$$
\begin{aligned}
V^*(s_0) - V^\pi(s_0) &= \gamma \mathbb{E}_{a_0 \sim \pi^*(\cdot|s_0)} \left[\left[\mathbb{E}_{s_1 \sim P_{\pi^*}(s_0, a_0)} V^*(s_1) \right] - \left[\mathbb{E}_{s_1 \sim P_\pi(s_0, a_0)} V^\pi(s_1) \right] \right] \\
&\quad + \mathbb{E}_{a_0 \sim \pi^*(\cdot|s_0)} \left[A^\pi(s_0, a_0) \right] + \mathbb{E}_{a_0 \sim \pi^*(\cdot|s_0)} \left[r_{\pi^*}(s_0, a_0) - r_\pi(s_0, a_0) \right] \\
&\leq \gamma \mathbb{E}_{a_0 \sim \pi^*(\cdot|s_0)} \mathbb{E}_{s_1 \sim P_{\pi^*}(s_0, a_0)} \left[V^*(s_1) - V^\pi(s_1) \right] + \mathbb{E}_{a_0 \sim \pi^*(\cdot|s_0)} \left[A^\pi(s_0, a_0) \right] \\
&\quad + \gamma \mathbb{E}_{a_0 \sim \pi^*(\cdot|s_0)} \mathbb{E}_{s_1 \sim P_{\pi^*} - P_\pi} \left[V^\pi(s_1) \right] + \mathbb{E}_{a_0 \sim \pi^*(\cdot|s_0)} \left[r_{\pi^*}(s_0, a_0) - r_\pi(s_0, a_0) \right] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi^*}} \mathbb{E}_{a \sim \pi^*(\cdot|s)} \left[A^\pi(s, a) \right] + \mathbb{E}_{a_0 \sim \pi^*(\cdot|s_0)} \left[r_{\pi^*}(s_0, a_0) - r_\pi(s_0, a_0) \right] \\
&\quad + \gamma \mathbb{E}_{a_0 \sim \pi^*(\cdot|s_0)} \mathbb{E}_{s_1 \sim P_{\pi^*} - P_\pi} \left[V^\pi(s_1) \right]
\end{aligned}
$$

Hence, by direct parameterization, ϵ_r-sensitivity of rewards and ϵ_p-sensitivity of transitions, we arrive at,

$$
\begin{aligned}
& V^*(\rho) - V^\pi(\rho) \\
\leq & \frac{1}{1-\gamma} \sum_{s,a} d_\rho^{\pi^*}(s) \pi^*(a \mid s) A^\pi(s, a) + \epsilon_r \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| + \gamma \epsilon_p \cdot \|V^*(\mu)\| \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| \\
\leq & \frac{1}{1-\gamma} \sum_{s,a} d_\rho^{\pi^*}(s) \max_{\bar{a}} A^\pi(s, \bar{a}) + \epsilon_r \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| + \gamma \epsilon_p \cdot \|V^*(\mu)\| \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| \\
= & \frac{1}{1-\gamma} \sum_{s} \frac{d_\rho^{\pi^*}(s)}{d_\mu^\pi(s)} \cdot d_\mu^\pi(s) \cdot \max_{\bar{a}} A^\pi(s, \bar{a}) + \epsilon_r \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| + \gamma \epsilon_p \cdot \|V^*(\mu)\| \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| \\
\overset{(a)}{\leq} & \frac{1}{1-\gamma} \left(\max_s \frac{d_\rho^{\pi^*}(s)}{d_\mu^\pi(s)} \right) \sum_s d_\mu^\pi(s) \cdot \max_{\bar{a}} A^\pi(s, \bar{a}) + (\epsilon_r + \gamma \epsilon_p \cdot \|V^*(\mu)\|) \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\|
\end{aligned}
$$

We note that (a) holds since $\max_{\bar{a}} A^{\pi}(s, \bar{a}) \geq 0$ for all states s and policies π. We now aim to upper bound the expression:

$$
\begin{aligned}
&\sum_{s} d_{\mu}^{\pi}(s) \cdot \frac{1}{1-\gamma} \max_{\bar{a}} A^{\pi}(s, \bar{a}) \\
&= \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \sum_{s,a} \frac{d_{\mu}^{\pi}(s)}{1-\gamma} \bar{\pi}(a \mid s) A^{\pi}(s, a) \\
&= \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \sum_{s,a} \frac{d_{\mu}^{\pi}(s)}{1-\gamma} (\bar{\pi}(a \mid s) - \pi(a \mid s)) A^{\pi}(s, a) \\
&= \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \sum_{s,a} \frac{d_{\mu}^{\pi}(s)}{1-\gamma} (\bar{\pi}(a \mid s) - \pi(a \mid s)) Q^{\pi}(s, a) \\
&= \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} (\bar{\pi} - \pi)^{\top} \Big[\nabla_{\pi} V^{\pi}(\mu) - \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi}} \mathbb{E}_{a \sim \pi(\cdot|s)} \mathbb{E}_{s' \sim P(s'|s,a)} \Big[\nabla_{\theta} \log P_{\pi_{\theta}}(s' \mid s, a) \cdot Q(s, a) \Big] \\
&- \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi}} \mathbb{E}_{a \sim \pi(\cdot|s)} \Big[\nabla_{\theta} R_{\pi}(s, a \mid \theta) \Big] \Big] \\
&\leq \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} (\bar{\pi} - \pi)^{\top} \left[\nabla_{\pi} V^{\pi}(\mu) + \frac{1}{1-\gamma} \left| \mathbb{E} \Big[\nabla_{\pi} \log P_{\pi} \cdot Q_{\pi}^{\pi} \Big] \right| + \frac{1}{1-\gamma} \left| \mathbb{E} \Big[\nabla_{\theta} R_{\pi}(s, a \mid \theta) \Big] \right| \right] \\
&\overset{(a)}{\leq} \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} (\bar{\pi} - \pi)^{\top} \left[\nabla_{\pi} V^{\pi}(\mu) + \frac{1}{1-\gamma} \mathbb{E} \left[\left| \nabla_{\pi} \log P_{\pi} \cdot Q_{\pi}^{\pi} \right| \right] + \frac{1}{1-\gamma} \mathbb{E} \left[\left| \nabla_{\theta} R_{\pi}(s, a \mid \theta) \right| \right] \right] \\
&\overset{(b)}{\leq} \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} (\bar{\pi} - \pi)^{\top} \nabla_{\pi} V^{\pi}(\mu) + \frac{M_1}{1-\gamma} \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} (\bar{\pi} - \pi)^{\top} \mathbf{1} \|V^{*}(\mu)\| + \frac{M_2}{1-\gamma} \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} (\bar{\pi} - \pi)^{\top} \mathbf{1}
\end{aligned}
$$

where (a) holds using Jensen's inequality and (b) follows from Cauchy-Schwarz inequality (Also we consider $Q^{*}(\mu) = \max_{a \in \mathcal{A}} Q(\mu, a)$). Hence, we arrive at,

$$
\begin{aligned}
V^{*}(\rho) - V^{\pi}(\rho) \leq \frac{1}{1-\gamma} &\left\| \frac{d_{\rho}^{\pi^{*}}}{\mu} \right\|_{\infty} \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} (\bar{\pi} - \pi)^{\top} \nabla_{\pi} V^{\pi}(\mu) \\
&+ \frac{M_1}{(1-\gamma)^2} \left\| \frac{d_{\rho}^{\pi^{*}}}{\mu} \right\|_{\infty} \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} (\bar{\pi} - \pi)^{\top} \mathbf{1} \|V^{*}(\mu)\| \\
&+ \frac{M_2}{(1-\gamma)^2} \left\| \frac{d_{\rho}^{\pi^{*}}}{\mu} \right\|_{\infty} \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} (\bar{\pi} - \pi)^{\top} \mathbf{1} + \epsilon_r \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| \\
&+ \gamma \epsilon_p \cdot \|V^{*}(\mu)\| \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\|
\end{aligned}
$$

Using Cauchy-Schwarz inequality again and bounding the value function norm with the norm of its optimal value, we get,

$$
\begin{aligned}
V^*(\rho) - V^\pi(\rho) \leq \frac{1}{1-\gamma} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty & \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} (\bar{\pi} - \pi)^\top \nabla_\pi V^\pi(\mu) \\
& + \frac{M_1}{(1-\gamma)^2} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \|V^*(\mu)\| \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| \\
& + \frac{M_2}{(1-\gamma)^2} \left\| \frac{d_\rho^{\pi^*}}{\mu} \right\|_\infty \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| + \epsilon_r \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\| \\
& + \gamma \epsilon_p \cdot \|V^*(\mu)\| \cdot \max_{\bar{\pi} \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \|\bar{\pi} - \pi\|
\end{aligned}
$$

where M_1 and M_2 are bounds on the gradients of the log transitions and rewards.

□

Lemma 5.2.3. *Let $\pi_\alpha := \pi_{\theta + \alpha u}$, and let $\tilde{V}(\alpha)$ be the corresponding value at a fixed state s_0, i.e.,*

$$
\tilde{V}(\alpha) := V^{\pi_\alpha}(s_0).
$$

Assume that

$$
\begin{aligned}
&\left| \sum_{a \in \mathcal{A}} \frac{d\pi_\alpha(a \mid s_0)}{d\alpha} \bigg|_{\alpha=0} \right| \leq C_1, \quad \left| \sum_{a \in \mathcal{A}} \frac{d^2\pi_\alpha(a \mid s_0)}{d\alpha^2} \bigg|_{\alpha=0} \right| \leq C_2 \\
&\left| \sum_{s \in \mathcal{S}} \frac{dP_\alpha(s \mid s_0, a_0)}{d\alpha} \bigg|_{\alpha=0} \right| \leq T_1, \quad \left| \sum_{s \in \mathcal{S}} \frac{d^2P_\alpha(s \mid s_0, a_0)}{d\alpha^2} \bigg|_{\alpha=0} \right| \leq T_2 \\
&\left| \sum_{a \in \mathcal{A}} \frac{dr_\alpha(s_0, a)}{d\alpha} \bigg|_{\alpha=0} \right| \leq R_1, \quad \left| \sum_{a \in \mathcal{A}} \frac{d^2r_\alpha(s_0, a)}{d\alpha^2} \bigg|_{\alpha=0} \right| \leq R_2
\end{aligned}
$$

Then

$$
\max_{\|u\|_2 = 1} \left\| \frac{d^2\tilde{V}(\alpha)}{d\alpha^2} \bigg|_{\alpha=0} \right\| \leq \frac{C_2}{1-\gamma} + 2C_1\beta_1 + C_2\beta_2
$$

where $\beta_1 = \frac{\gamma}{(1-\gamma)^2}(C_1 + T_1) + \frac{R_1}{1-\gamma}$ and $\beta_2 = \frac{2\gamma^2}{(1-\gamma)^3}(C_1 + T_1)^2 + \frac{\gamma}{(1-\gamma)^2}(C_2 + 2C_1T_1 + T_2) + \frac{2\gamma R_1}{(1-\gamma)^2}(C_2 + 2C_1T_1 + T_2) + \frac{R_2}{1-\gamma} + \frac{\gamma C_1 R_1}{(1-\gamma)^2}$

Proof. Consider a unit vector u, and let P_α be the state-action transition matrix under policy π_α, defined as:

$$
[P(\alpha)]_{(s,a) \to (s',a')} = \pi_\alpha(a' \mid s') P_\alpha(s' \mid s, a)
$$

We can differentiate P_α with respect to α to obtain:

$$\left.\frac{dP_\alpha}{d\alpha}\right|_{\alpha=0}(s,a)\to(s',a') = \left.\frac{d\pi_\alpha(a'\mid s')}{d\alpha}\right|_{\alpha=0} P(s'\mid s,a) + \left.\frac{dP_\alpha(s'\mid s,a)}{d\alpha}\right|_{\alpha=0}\pi_\alpha(a'\mid s')$$

Now, for an arbitrary vector x, we have:

$$\left[\left.\frac{dP_\alpha}{d\alpha}\right|_{\alpha=0} x\right]_{(s,a)} = \sum_{s',a'}\left.\frac{d\pi_\alpha(a'\mid s')}{d\alpha}\right|_{\alpha=0} P_\alpha(s'\mid s,a)x_{s',a'} + \sum_{s',a'}\left.\frac{dP_\alpha(s'\mid s,a)}{d\alpha}\right|_{\alpha=0}\pi_\alpha(a'\mid s')x_{s',a'}$$

Taking the maximum over unit vectors u in ℓ_2-norm:

$$\begin{aligned}
\max_{\|u\|_2=1}\left\|\left.\frac{dP_\alpha}{d\alpha}\right|_{\alpha=0} x\right\|_\infty &\le \max_{\|u\|_2=1}\left|\sum_{s',a'}\left.\frac{d\pi_\alpha(a'\mid s')}{d\alpha}\right|_{\alpha=0} P_\alpha(s'\mid s,a)x_{s',a'}\right| \\
&\quad + \max_{\|u\|_2=1}\left|\sum_{s',a'}\left.\frac{dP_\alpha(s'\mid s,a)}{d\alpha}\right|_{\alpha=0}\pi_\alpha(a'\mid s')x_{s',a'}\right| \\
&\le \max_{s,a}\sum_{s'}P(s'\mid s,a)\sum_{a'}\left|\left.\frac{d\pi_\alpha(a'\mid s')}{d\alpha}\right|_{\alpha=0}\right|\cdot\|x\|_\infty \\
&\quad + \max_{s,a}\sum_{a'}\pi_\alpha(a'\mid s')\sum_{s'}\left|\left.\frac{dP(s'\mid s,a)}{d\alpha}\right|_{\alpha=0}\right|\cdot\|x\|_\infty \\
&\le \max_{s,a}\sum_{s'}P(s'\mid s,a)\|x\|_\infty C_1 + \max_{s,a}\sum_{a'}\pi(a'\mid s')\|x\|_\infty T_1 \\
&\le C_1\|x\|_\infty + T_1\|x\|_\infty = (C_1+T_1)\|x\|_\infty
\end{aligned}$$

By the definition of the ℓ_∞-norm, we conclude:

$$\max_{\|u\|_2=1}\left\|\left.\frac{dP_\alpha}{d\alpha}\right|_{\alpha=0} x\right\|_\infty \le (C_1+T_1)\|x\|_\infty$$

Similarly, differentiating $\tilde{P}(\alpha)$ twice w.r.t. α, we get

$$\begin{aligned}
\left[\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}\Big|_{\alpha=0}\right]_{(s,a)\to(s',a')} &= \left.\frac{d^2\pi_\alpha(a'\mid s')}{(d\alpha)^2}\right|_{\alpha=0} P(s'\mid s,a) + \left.\frac{d^2P_\alpha(s'\mid s,a)}{d\alpha^2}\right|_{\alpha=0}\pi_\alpha(a'\mid s') \\
&\quad + 2\left.\frac{d\pi_\alpha(a'\mid s')}{d\alpha}\right|_{\alpha=0}\left.\frac{dP_\alpha(s'\mid s,a)}{d\alpha}\right|_{\alpha=0}
\end{aligned}$$

Now consider the norm bound:

$$\max_{\|u\|_2=1}\left\|\left.\frac{d^2\tilde{P}(\alpha)}{(d\alpha)^2}\right|_{\alpha=0} x\right\|_\infty \le C_2\|x\|_\infty + 2C_1T_1\|x\|_\infty + T_2\|x\|_\infty = (C_2+2C_1T_1+T_2)\,\|x\|_\infty$$

Let $Q_\alpha(s_0, a_0)$ be the Q-function corresponding to the policy π_α at state s_0 and action a_0. Observe that $Q_\alpha(s_0, a_0)$ can be written as:

$$Q_\alpha(s_0, a_0) = e_{(s_0,a_0)}^\top (I - \gamma P_e(\alpha))^{-1} r = e_{(s_0,a_0)}^\top M(\alpha) r_\alpha$$

where $M(\alpha) := (I - \gamma P_e(\alpha))^{-1}$

Differentiating $M(\alpha)$ twice with respect to α gives:

$$\frac{dQ^\alpha(s_0, a_0)}{d\alpha} = \gamma e_{(s_0,a_0)}^\top M(\alpha) \frac{d\tilde{P}(\alpha)}{d\alpha} M(\alpha) r_\alpha + e_{(s_0,a_0)}^\top M(\alpha) \frac{dr_\alpha}{d\alpha}$$

And correspondingly,

$$\begin{aligned}
\frac{d^2Q^\alpha(s_0, a_0)}{d\alpha^2} &= 2\gamma^2 e_{(s_0,a_0)}^\top M(\alpha) \frac{d\tilde{P}(\alpha)}{d\alpha} M(\alpha) \frac{d\tilde{P}(\alpha)}{d\alpha} M(\alpha) r_\alpha + \gamma e_{(s_0,a_0)}^\top M(\alpha) \frac{d^2\tilde{P}(\alpha)}{d\alpha^2} M(\alpha) r_\alpha \\
&+ \gamma e_{(s_0,a_0)}^\top M(\alpha) \frac{d\tilde{P}(\alpha)}{d\alpha} M(\alpha) \frac{dr_\alpha}{d\alpha} + e_{(s_0,a_0)}^\top M(\alpha) \frac{d^2 r_\alpha}{d\alpha^2} \\
&+ \gamma e_{(s_0,a_0)}^\top M(\alpha) \frac{d\tilde{P}(\alpha)}{d\alpha} M(\alpha) \frac{dr_\alpha}{d\alpha}
\end{aligned}$$

By using the power series expansion of the matrix inverse, we can write $M(\alpha)$ as:

$$M(\alpha) = (I - \gamma P_e(\alpha))^{-1} = \sum_{n=0}^{\infty} \gamma^n P_e(\alpha)^n$$

which implies that $M(\alpha) \geq 0$ (componentwise), and

$$M(\alpha)\mathbf{1} = \frac{1}{1-\gamma}\mathbf{1},$$

i.e., each row of $M(\alpha)$ is positive and sums to $\frac{1}{1-\gamma}$.

This implies:

$$\max_{\|u\|_2=1} \|M(\alpha)x\|_\infty \leq \frac{1}{1-\gamma}\|x\|_\infty.$$

This gives, using the expressions for $\frac{d^2Q^\alpha(s_0,a_0)}{d\alpha^2}$ and $\frac{dQ^\alpha(s_0,a)}{d\alpha}$, an upper bound on their magnitudes based on $\|x\|_\infty$ and constants arising from bounds on the derivatives of $\tilde{P}(\alpha)$ and r_α.

$$\begin{aligned}
&\max_{\|u\|_2=1} \left\| \frac{d^2 Q^\alpha(s_0,a_0)}{d\alpha^2} \right\|_\infty \\
\leq & 2\gamma^2 \left\| M(\alpha) \frac{d\tilde{P}(\alpha)}{d\alpha} M(\alpha) \frac{d\tilde{P}(\alpha)}{d\alpha} M(\alpha) r_\alpha \right\|_\infty + \gamma \left\| M(\alpha) \frac{d^2\tilde{P}(\alpha)}{d\alpha^2} M(\alpha) r_\alpha \right\|_\infty \\
& + \gamma \left\| M(\alpha) \frac{d^2\tilde{P}(\alpha)}{d\alpha^2} M(\alpha) \frac{dr_\alpha}{d\alpha} \right\|_\infty + \left\| M(\alpha) \frac{d^2 r_\alpha}{d\alpha^2} \right\|_\infty + 2\gamma \left\| M(\alpha) \frac{d\tilde{P}(\alpha)}{d\alpha} M(\alpha) \frac{dr_\alpha}{d\alpha} \right\|_\infty
\end{aligned}$$

Bounding using known bounds on transitions and rewards:

$$\begin{aligned}
\max_{\|u\|_2=1} \left\| \frac{dQ^\alpha(s_0,a_0)}{d\alpha} \right\|_\infty \leq & \frac{2\gamma^2}{(1-\gamma)^3}(C_1+T_1)^2 + \frac{\gamma}{(1-\gamma)^2}(C_2+2C_1T_1+T_2) \\
& + \frac{2\gamma R_1}{(1-\gamma)^2}(C_2+2C_1T_1+T_2) + \frac{R_2}{1-\gamma} + \frac{\gamma C_1 R_1}{(1-\gamma)^2} = \beta_2
\end{aligned}$$

Corresponding bound on the first derivative is:

$$\begin{aligned}
\max_{\|u\|_2=1} \left\| \frac{dQ^\alpha(s_0,a_0)}{d\alpha} \right\|_\infty &\leq \gamma \left\| M(\alpha) \frac{d\tilde{P}(\alpha)}{d\alpha} M(\alpha) \frac{dr_\alpha}{d\alpha} \right\|_\infty + \left\| M(\alpha) \frac{dr_\alpha}{d\alpha} \right\|_\infty \\
&\leq \frac{\gamma}{(1-\gamma)^2}(C_1+T_1) + \frac{R_1}{1-\gamma} = \beta_1
\end{aligned}$$

Consider the expected value under policy π_α:

$$V^\pi(\alpha) = \sum_a \pi_\alpha(a \mid s_0) Q^\alpha(s_0,a)$$

Differentiating twice with respect to α, we obtain:

$$\frac{d^2\tilde{V}(\alpha)}{d\alpha^2} = \sum_a \frac{d^2\pi_\alpha(a \mid s_0)}{d\alpha^2} Q^\alpha(s_0,a) + 2\sum_a \frac{d\pi_\alpha(a \mid s_0)}{d\alpha} \frac{dQ^\alpha(s_0,a)}{d\alpha} + \sum_a \pi_\alpha(a \mid s_0) \frac{d^2 Q^\alpha(s_0,a)}{d\alpha^2}$$

Finally,

$$\max_{\|\mathbf{u}\|_2=1} \left\| \frac{d^2\tilde{V}(\alpha)}{d\alpha^2} \right\|_\infty \leq \frac{C_2}{1-\gamma} + 2C_1\beta_1 + \beta_2$$

□

Lemma 5.2.4. *For all starting states s_0,*

$$\left\|\nabla_\pi V^\pi(s_0) - \nabla_\pi V^{\pi'}(s_0)\right\|_2 \leq (2\sqrt{|\mathcal{A}|}\beta_1 + \beta_2)\left\|\pi - \pi'\right\|_2.$$

where $\beta_1 = \frac{\gamma}{(1-\gamma)^2}(\sqrt{\mathcal{A}} + T_1) + \frac{R_1}{1-\gamma}$ *and* $\beta_2 = \frac{2\gamma^2}{(1-\gamma)^3}(\sqrt{|\mathcal{A}|} + T_1)^2 + \frac{\gamma}{(1-\gamma)^2}(2T_1\sqrt{\mathcal{A}} + T_2) + \frac{2\gamma R_1}{(1-\gamma)^2}(2T_1\sqrt{|\mathcal{A}|} + T_2) + \frac{R_2}{1-\gamma} + \frac{\gamma R_1\sqrt{|\mathcal{A}|}}{(1-\gamma)^2}$

Proof. By differentiating π_α with respect to α, we get

$$\sum_{a\in\mathcal{A}}\left|\frac{d\pi_\alpha(a \mid s_0)}{d\alpha}\right| \leq \sum_{a\in\mathcal{A}}|u_{a,s}| \leq \sqrt{|\mathcal{A}|}.$$

Differentiating again with respect to α, we obtain

$$\sum_{a\in\mathcal{A}}\left|\frac{d^2\pi_\alpha(a \mid s_0)}{d\alpha^2}\right| = 0.$$

Using this with Lemma 4.2.3 with $C_1 = \sqrt{|\mathcal{A}|}$ and $C_2 = 0$, we get

$$\max_{\|u\|_2=1}\left\|\left.\frac{d^2\tilde{V}(\alpha)}{d\alpha^2}\right|_{\alpha=0}\right\|_\infty \leq \frac{C_2}{1-\gamma} + 2C_1\beta_1 + \beta_2 = 2\sqrt{|\mathcal{A}|}\beta_1 + \beta_2,$$

where $\beta_1 = \frac{\gamma}{(1-\gamma)^2}(\sqrt{|\mathcal{A}|} + T_1) + \frac{R_1}{1-\gamma}$ and $\beta_2 = \frac{2\gamma^2}{(1-\gamma)^3}(\sqrt{|\mathcal{A}|} + T_1)^2 + \frac{\gamma}{(1-\gamma)^2}(2\sqrt{|\mathcal{A}|}T_1 + T_2) + \frac{2\gamma R_1}{(1-\gamma)^2}(2\sqrt{|\mathcal{A}|}T_1 + T_2) + \frac{R_2}{1-\gamma} + \frac{\gamma\sqrt{|\mathcal{A}|}R_1}{(1-\gamma)^2}$ this completes the Proof. □

Lemma 5.2.6. *Let $V^\pi(\mu)$ be β-smooth in π. Define the gradient mapping*

$$G_\eta(\pi) = \frac{1}{\eta}\left(P_{\Delta(\mathcal{A})^{|\mathcal{S}|}}(\pi + \eta\nabla_\pi V^\pi(\mu)) - \pi\right),$$

and define the projected gradient update rule as $\pi^+ = \pi + \eta G_\eta(\pi)$. If $\|G_\eta(\pi)\|_2 \leq \varepsilon$, then

$$\max_{\pi+\delta\in\Delta(\mathcal{A})^{|\mathcal{S}|},\,\|\delta\|_2\leq 1}\delta^\top\nabla_\pi V^{\pi^+}(\mu) \leq \varepsilon(\eta\beta + 1).$$

Proof. By Lemma 4.2.5, we have

$$\nabla_\pi V^{\pi^+}(\mu) \in \mathcal{N}_{\Delta(\mathcal{A})^{|\mathcal{S}|}}(\pi^+) + \varepsilon(\eta\beta + 1)\mathbb{B}_2,$$

where $\mathbb{B}_2$ is the unit ℓ_2-ball and $\mathcal{N}_{\Delta(\mathcal{A})^{|\mathcal{S}|}}(\pi^+)$ is the normal cone to the product simplex at π^+.

Since $\nabla_\pi V^{\pi^+}(\mu)$ lies within $\varepsilon(\eta\beta+1)$-distance of the normal cone, and any $\delta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ with $\|\delta\|_2 \leq 1$ in the tangent cone satisfies

$$\delta^\top \nabla_\pi V^{\pi^+}(\mu) \leq \varepsilon(\eta\beta+1),$$

□

Bibliography

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76. 9, 16

Baharlouei, S., Patel, S., and Razaviyayn, M. (2023). f-ferm: A scalable framework for robust fair empirical risk minimization. *arXiv preprint arXiv:2312.03259*. 7

Balseiro, S., Lu, H., and Mirrokni, V. (2021). Regularized online allocation problems: Fairness and beyond. 5, 6

Barocas, S., Hardt, M., and Narayanan, A. (2023a). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. 1

Barocas, S., Hardt, M., and Narayanan, A. (2023b). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. 5, 6

Bechavod, Y. and Ligett, K. (2019). Fairness in dynamic settings through online learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 1

Beck, A. (2017). *First-order methods in optimization*. SIAM. 23, 25

Bell, A., Bynum, L., Drushchak, N., Herasymova, T., Rosenblatt, L., and Stoyanovich, J. (2023). The possibility of fairness: Revisiting the impossibility theorem in practice. 5

Brown, G., Hod, S., and Kalemaj, I. (2022). Performative prediction in a stateful world. 2, 8, 15

Cai, S., Han, F., and Cao, X. (2024). Performative control for linear dynamical systems. *arXiv preprint arXiv:2410.23251*. 8

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38. 5, 6

Chen, Q., Chen, Y., and Li, B. (2024a). Practical performative policy learning with strategic agents. *arXiv preprint arXiv:2412.01344*. 8

Chen, Y., Tang, W., Ho, C.-J., and Liu, Y. (2024b). Performative prediction with bandit feedback: Learning through reparameterization. 8

Chi, J., Shen, J., Dai, X., Zhang, W., Tian, Y., and Zhao, H. (2022). Towards return parity in markov decision processes. 1, 2, 6, 7

Chiappa, S. (2023). Designing long-term group fair policies in dynamical systems. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 1

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., and Halpern, Y. (2020). Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 525–534, New York, NY, USA. Association for Computing Machinery. 5

Deng, Z., Sun, H., Wu, Z. S., Zhang, L., and Parkes, D. C. (2022). Reinforcement learning with stepwise fairness constraints. *arXiv preprint arXiv:2211.03994*. 1

Du, R., Muthirayan, D., Khargonekar, P. P., and Shen, Y. (2024). Long-term fairness for real-time decision making: A constrained online optimization approach. 1, 6, 7, 12

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the Innovations in Theoretical Computer Science Conference*. 1

Fudenberg, D. and Levine, D. K. (2012). Fairness, risk preferences and independence: Impossibility theorems. *Journal of Economic Behavior & Organization*, 81(2):606–612. 5

Ge, Y., Liu, S., Gao, R., Xian, Y., Li, Y., Zhao, X., Pei, C., Sun, F., Ge, J., Ou, W., and Zhang, Y. (2021). Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 445–453. ACM. 5

Ghadimi, S. and Lan, G. (2016). Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99. 25

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*. 1

Hardt, M. and Recht, B. (2021). *Patterns, Predictions, and Actions: A Story About Machine Learning*. Cambridge University Press. 5

Hort, M., Chen, Z., Zhang, J. M., Harman, M., and Sarro, F. (2023). Bias mitigation for machine learning classifiers: A comprehensive survey. 5

Hsu, B., Mazumder, R., Nandy, P., and Basu, K. (2022). Pushing the limits of fairness impossibility: Who's the fairest of them all? 5

Hu, Y. and Zhang, L. (2022). Achieving long-term fairness in sequential decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9549–9557. 5

Izzo, Z., Ying, L., and Zou, J. (2021). How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning*, pages 4641–4650. PMLR. 8, 29

Izzo, Z., Zou, J., and Ying, L. (2022). How to learn when data gradually reacts to your model. In *International Conference on Artificial Intelligence and Statistics*, pages 3998–4035. PMLR. 8

Kakade, S. M. (2002). A natural policy gradient. In *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, pages 1531–1538. 9

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*. 5

Kim, M. P. and Perdomo, J. C. (2022). Making decisions under outcome performativity. *arXiv preprint arXiv:2210.01745*. 8

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. 5

Kokhlikyan, N., Alsallakh, B., Wang, F., Miglani, V., Yang, O. A., and Adkins, D. (2022). Bias mitigation framework for intersectional subgroups in neural networks. *arXiv preprint arXiv:2212.13014*. 6

Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in Neural Information Processing Systems 13 (NIPS 2000)*, pages 1008–1014. 9, 16

Li, J., Khayatkhoei, M., Zhu, J., Xie, H., Hussein, M. E., and AbdAlmageed, W. (2023). Information-theoretic bounds on the removal of attribute-specific bias from neural networks. *arXiv preprint arXiv:2310.04955*. 6

Li, J., Khayatkhoei, M., Zhu, J., Xie, H., Hussein, M. E., and AbdAlmageed, W. (2025). A critical review of predominant bias in neural networks. *arXiv preprint arXiv:2502.11031*. 6

Li, Q., Yau, C.-Y., and Wai, H.-T. (2022). Multi-agent performative prediction with greedy deployment and consensus seeking agents. *Advances in Neural Information Processing Systems*, 35:38449–38460. 8

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. (2018). Delayed impact of fair machine learning. 6

Mandal, D. and Radanovic, G. (2024). Performative reinforcement learning with linear markov decision process. *arXiv preprint arXiv:2411.05234*. 8

Mandal, D., Triantafyllou, S., and Radanovic, G. (2023). Performative reinforcement learning. 4, 7, 17, 18, 23, 32, 40, 43, 44, 45

Mendler-Dünner, C., Perdomo, J., Zrnic, T., and Hardt, M. (2020). Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939. 2, 8

Miller, J., Perdomo, J. C., and Zrnic, T. (2021). Outside the echo chamber: Optimizing the performative risk. 8

Mishler, A. and Dalmasso, N. (2022). Fair when trained, unfair when deployed: Observable fairness measures are unstable in performative prediction settings. 7

Oneto, L. and Chiappa, S. (2020). *Fairness in Machine Learning*, page 155–196. Springer International Publishing. 5, 6

Ortmann, L., Böhm, F., Klein-Helmkamp, F., Ulbig, A., Bolognani, S., and Dörfler, F. (2024). Tuning and testing an online feedback optimization controller to provide curative distribution grid flexibility. *Electric Power Systems Research*, 234:110660. 6

Piliouras, G. and Yu, F.-Y. (2023a). Multi-agent performative prediction: From global stability and optimality to chaos. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 1047–1074. 2, 8

Piliouras, G. and Yu, F.-Y. (2023b). Multi-agent performative prediction: From global stability and optimality to chaos. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 1047–1074. 8

Rank, B., Triantafyllou, S., Mandal, D., and Radanovic, G. (2024). Performative reinforcement learning in gradually shifting environments. 8

Rateike, M., Valera, I., and Forré, P. (2024). Designing long-term group fair policies in dynamical systems. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 20–50. 6

Rodemann, J., Jansen, C., and Schollmeyer, G. (2024). Reciprocal learning. *Advances in Neural Information Processing Systems*, 37:1686–1724. 8

Roh, Y., Lee, K., Whang, S. E., and Suh, C. (2023). Improving fair training under correlation shifts. In *International Conference on Machine Learning*, pages 29179–29209. PMLR. 7

Saravanakumar, K. K. (2021). The impossibility theorem of machine fairness – a causal perspective. 5

Sutton, R. S., McAllester, D. A., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12 (NIPS 1999)*, pages 1057–1063. The MIT Press. 9, 15, 20

Taori, R. and Hashimoto, T. B. (2022). Data feedback loops: Model-driven amplification of dataset biases. 8

Triantafyllou, S., Singla, A., and Radanovic, G. (2021). On blame attribution for accountable multi-agent sequential decision making. *Advances in Neural Information Processing Systems*, 34:15774–15786. 31

Vadavathi, A. R., Hoogsteen, G., and Hurink, J. (2024). Fair and efficient congestion management for low voltage distribution networks. In *2024 IEEE 8th Energy Conference (ENERGYCON)*, pages 1–6. IEEE. 6

Williams, R. J. (1992a). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256. 9, 16

Williams, R. J. (1992b). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256. 20

Xian, R., Yin, L., and Zhao, H. (2023). Fair and optimal classification via post-processing. In *International conference on machine learning*, pages 37977–38012. PMLR. 6

Xu, Y., Deng, C., Sun, Y., Zheng, R., Wang, X., Zhao, J., and Huang, F. (2024). Adapting static fairness to sequential decision-making: Bias mitigation strategies towards equal long-term benefit rate. 1, 2, 3, 6, 7, 12, 13, 40

Yan, W. and Cao, X. (2023). Zero-regret performative prediction under inequality constraints. *Advances in Neural Information Processing Systems*, 36:1298–1308. 8

Yin, T., Raab, R., Liu, M., and Liu, Y. (2023). Long-term fairness with unknown dynamics. 1, 2, 6, 7

Yin, X., Chouldechova, A., and Roth, A. (2024). Long-term fairness inquiries and pursuits in machine learning. *arXiv preprint arXiv:2406.06736*. 1

Yu, E. Y., Qin, Z., Lee, M. K., and Gao, S. (2022). Policy optimization with advantage regularization for long-term fairness in decision systems. *arXiv preprint arXiv:2210.12546*. 6

Zezulka, S. and Genin, K. (2024). From the fair distribution of predictions to the fair distribution of social goods: Evaluating the impact of fair machine learning on long-term unemployment. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1984–2006. 7

Zhan, S., Morren, J., van den Akker, W., van der Molen, A., Paterakis, N. G., and Slootweg, J. G. (2024). Fairness-incorporated online feedback optimization for real-time distribution grid management. *IEEE Transactions on Smart Grid*, 15(2):1792–1806. 6, 7

Zhang, X., Khalili, M. M., and Liu, M. (2020). Long-term impacts of fair machine learning. *Ergonomics in Design*, 28(3):23–31. 5

Zhao, C., Chen, F., and Thuraisingham, B. (2021). Fairness-aware online meta-learning. 6, 7

Zheng, X., Xie, T., Tan, X., Yener, A., Zhang, X., Payani, A., and Lee, M. (2024). Profl: Performative robust optimal federated learning. *arXiv preprint arXiv:2410.18075*. 8