Group Fairness in Reinforcement Learning and Large Language Models

Submitted in partial fulfillment of the requirements for

the degree of

Master of Science

in

Computer Science

Kefan Song

University of Virginia

October 2024

APPROVAL SHEET

This

is submitted in partial fulfillment of the requirements
for the degree of

Author:

Advisor:

Committee Member:

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

To my parents, for their love and support.

Acknowledgements

I would like to express my gratitude to my advisor, Professor Shangtong Zhang, for his continous support through my research. I am grateful to him for guiding me into doing Reinforcement Learning and Large Language Model research for the first time. His guidance and encouragement were instrumental in bringing this work to completion.

I would also like to thank my other thesis committee members, Professor Chen-Yu Wei and Professor Yu Meng, for their detailed critiques and suggestions, which have provided me with a better understanding of the research problems at hand and ideas for improvement.

I am fortunate to learn from and collaborate with Jin Yao, whose problem-solving mindset is truly inspiring. A big thanks to Shuze Liu for introducing me to W&B and slurm servers, and to Haolin Liu for our fun and insightful discussions on reinforcement learning research. I am fortunate to have Lei Gong as a friend and always enjoy the get-togethers he hosts. I truly appreciate the great conversations with Ye Ma and Licheng Luo about research and life in general. Special thanks to Amir Shariatmadari for the cherished times spent solving math problems together. And, lastly, I'm grateful to Jiuqi Wang for organizing the UVA Ping Pong Club every Friday—it's been a weekly tradition I've thoroughly enjoyed and looked forward to.

Abstract

This thesis addresses an important societal consideration in the application of Reinforcement Learning (RL): the equitable distribution of its benefits across different demographic groups. Specifically, we investigate how to incorporate group fairness into reinforcement learning algorithms to ensure that their societal impact is just and fair. The thesis is organized around two key contributions to group fairness in RL.

The first contribution focuses on multi-task group fairness in reinforcement learning. In many practical applications, such as recommender systems or fine-tuning large language models, a single policy is required to perform multiple tasks in real-world environments. In this thesis, we introduce a multi-task fairness constraint and propose a novel algorithm to solving this problem based on constrained optimization. Through experiments in Mujoco, we demonstrate that our method better ensures group fairness compared to the previous approach that lacks this multi-task fairness constraint.

The second contribution studies group fairness in the context of fine-tuning large language models (LLMs) through Reinforcement Learning with Human Feedback (RLHF). Current approaches to address bias in LLMs largely concentrate on mitigating harmful language and often overlook group fairness considerations. In this work, we emphasis on demographic parity, a key group fairness definition that aligns with the broader fair machine learning research. In this work, we identify reward models as a potential source of bias in the RLHF process and propose a novel evaluation method based on arXiv meta-data for group fairness in reward models. Our experiment on fine-tuning the Phi-1.5 model further demonstrates that biases in reward models can propagate into the fine-tuned LLMs during RLHF training.

Contents

List of Tables

List of Figures

Chapter 1

Group Fairness in Reinforcement Learning

In this chapter, we will first introduce the group fairness problem in Reinforcement Learning, followed by the two contributions under this problem. Then we describe the outline of this paper.

1.1 Importance of Ensuring Group Fairness in Reinforcement Learning

Learning-based algorithms have been applied more to real-world high-stakes social problems, such as bank loans, medical interventions, and school admissions. They are also applied in less high-stake scenarios, by making movie recommendations, suggesting products to buy, and question answering. In both cases, the deployment of learning-based algorithms will make automated decisions that have a direct impact on our society. Therefore, one critical issue is to ensure the algorithm has low social biases and delivers fair outcomes for people from all demographic groups. However, since these social problems are long-term in nature, an unconstrained algorithm may create a feedback loop over time and enlarge the discrepancy between people from different social groups. Reinforcement Learning has demonstrated a superior performance in many of these tasks, which are sequential-decision making problems in nature. When the fairness requirement is accounted for in the RL algorithm, it has the promise of addressing the long-term fairness issue and thus has an advantage over fair machine learning algorithms.

In this thesis, we study the problem of group fairness reinforcement learning. We focus on the aspect of group fairness. It requires the algorithm to deliver similar outcomes for people from different social groups, categorized by their sensitive information such as gender, education, or social-economic status.

1.2 Multi-Task Group Fairness in Reinforcement Learning

For our first contribution, we developed a novel algorithm for multi-task group fairness in reinforcement learning, and demonstrated that policies trained with this approach achieve a smaller disparity in average rewards across social groups and therefore better ensures group fairness.

Many real-world applications are multi-task in nature and it is critical to ensure group fairness is achieved for all tasks. To the best of our knowledge, this is the first work that accounts for the fairness problem in multi-task reinforcement learning problem. To further motivate our problem, we discuss two application scenarios in the following.

Scenario 1: RL based Recommender systems. Consider an example in multi-task recommender systems, where an RL policy recommends contents catered to the preference of the users and aims to achieve multiple tasks such as a high long-term user-engagement and a high click-through-rate for advertisement content. The users' sensitive information, such as age, social economic status and gender, is taken by the policy as input features to make recommendations. When there is no group fairness consideration during the algorithm development, it is more likely for the algorithm to maximize the user-engagement of the majority social groups, and creating a feedback loop by increasing the size of the majority social groups. This is a realistic concern, as It is known that some social media platforms such as TikTok has predominately younger demographics, while Twitter and

Reddit have more male users. Ensuring the group fairness criteria over multiple tasks requires to limit the user-engagement of the majority groups. With a more balanced social group ratios in the system's user pool, more content creators need to include the minority group as their targeting audience and thus creating content that is more inclusive and engaging for the minority group, potentially breaking the feedback loop.

Scenario 2: Finetuning LLM with RL In fine-tuning Large Language Models with Reinforcement Learning from Human Feedback(RLHF) over multiple tasks such as common sense reasoning, question answering, and explanation generation, the training data is a collection of human prompts as inputs for the LLM, which can also be regarded as the states for the RL policy. Since the prompts are collected from people from diverse social groups, inherent imbalance exists with respect to specific demographics in the dataset. Consequently, the LLM fine-tuned with RL may disproportionally improve the quality of responses of prompts from majority groups. A feedback loop exists if the deployed updated LLM results in more active users from the majority groups, who may generate new data for further LLM fine-tuning.

Motivated by these use cases, we designed a algorithm with a multi-task fairness constraint that can achieve a smaller return difference between different social groups and provide a proof for zero multi-task fairness constraint violation in the tabular case.

1.3 Group Fairness in Reward Models for Fine-tuning LLMs with RLHF

For our second contribution, we introduced demographic parity into the group fairness evaluation of reward models for fine-tuning LLMs with RLHF, and demonstrated that all reward models evaluated are significantly biased.

As the state-of-the-art Large Language Models (LLM) acquire advanced capabilities and are already assisting a large population of human users (Hu, 2023), it is important to ensure the benefits from LLMs are equally, fairly and broadly distributed among people from various demographic groups. This concern aligns with the broader ethical and responsible AI initiatives (Goellner et al., 2024) and is listed as the top priority objective in OpenAI's charter (OpenAI, 2024), which seeks to prevent the concentration of technological benefits from LLM towards specific groups of people that could otherwise exacerbate existing societal disparities.

We formulate this concern as the *group fairness* problem in LLMs, which aims to ensure that LLM outputs do not disproportionately benefit specific demographic groups. However, current bias and fairness research in LLMs faces significant limitations when addressing group fairness in the following two aspects.

One of the limitations is that it does not account for the cases where different groups have different questions for the LLMs. Traditionally, group fairness in fair machine learning classification is evaluated in settings where non-sensitive attributes, i.e. input features of a person without the group attribute, are expected to be different for people from different demographic groups. In the case of LLMs, the prompt questions without the group attribute, are the non-sensitive attributes. Users from different demographic groups often generate distinct prompts due to diverse interests and challenges in their daily lives, as evidenced by prior research on demographic-driven web searches (Weber and Castillo, 2010).

However, common methods in LLMs typically require users from different demographic groups to share the same prompt questions for fairness evaluation (Nangia et al., 2020; Webster et al., 2021; Wang and Cho, 2019). Existing work has primarily focused on ensuring that sensitive attributes (e.g. gendered pronouns like "he" or "she"), when added to the same prompt, do not lead to different model outputs. While these efforts reduce stereotypical language toward specific groups, they do not account for the challenges posed by different prompt questions from various demographic groups. Furthermore, in real-world interactions with LLMs, users often do not explicitly state their sensitive attributes in prompts, making these methods inapplicable when demographic information is implicit rather than explicitly mentioned.

In this regard, it's essential to focus on how non-sensitive attributes in the prompt questions may contribute to unfairness in LLM outputs. Although we can no longer directly compare model outputs when the prompt questions are different, the comparisons can be made based on the quality of the LLM output. In the context of LLMs, the model outcomes are less about high-stakes decisions and more about user-perceived helpfulness, correctness, and coherence of generated text. These attributes can be conveniently measured by the reward model used in Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) as shown in Figure **??**.

A second limitation in current research is the focus on evaluating fairness solely based on the LLM's final output. These approaches fail to identify the specific sources of bias within the model's development pipeline. Bias in LLM outputs can originate from various stages of LLM training, including both the pre-training and fine-tuning phases. During fine-tuning, bias can be introduced through the RLHF procedure, or the learned reward model itself. A more granular evaluation of group fairness across various components of the LLM training pipeline, such as pre-training, supervised fine-tuning, reward modeling, and reinforcement learning, could offer valuable insights into the origins of these biases and help develop more effective mitigation strategies toward LLMs that can benefit for people from all demographic groups equitably.

Recognizing the above limitations, in this thesis, we first benchmark the group fairness in reward models and demonstrate that the group unfairness in reward models propagates to LLMs during the RLHF fine-tuning process.

1.4 Outline

The remainder of this thesis is organized as follows: Chapter 2 provides an overview of key concepts, including Group Fairness, Constrained Markov Decision Processes, and Reinforcement Learning from Human Feedback (RLHF). Chapter 3 introduces the first major contribution: the algorithm of Multi-Task Group Fairness in Reinforcement Learning. Chapter 4 presents the second contribution, focusing on the evaluation of reward models and the observed increase in bias in RLHF fine-tuned LLMs.

Chapter 2

Background and Related Work

2.1 Infinite-horizon Discounted Markov Decision Process

We formulate the long-term fairness problem as an infinite-horizon discounted Markov Decision Process (MDP), defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \gamma, \mu, r, P \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action state, $\mu : \mathcal{S} \to [0, 1]$ is the initial state distribution, $\gamma \in [0, 1)$ is the discount factor, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, and $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition function.

In this setting, a stationary policy π is defined as $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$. The trajectory $\tau = \{(s_t, a_t)\}_{t=1}^{\infty}$ is sampled from $p_\pi(\tau)$, which is defined as $p_\pi(\tau) = \mu(s_1)\Pi_{t=1}^{\infty}\pi(a_t|s_t)$ $P(s_{t+1}|s_t, a_t)$. The infinite-horizon discounted return of policy π and reward r is defined as $J(\pi; \mu, P, r) \doteq \mathrm{E}_{\tau \sim p_\pi(\tau)}[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)]$. The value function is defined as $V^\pi(s; \mu, P, r) = \mathrm{E}_{\tau \sim p_\pi(\tau)}[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)|s_t = s]$, and the state-action value function is defined as $Q^\pi(s, a; \mu, P, r) = \mathrm{E}_{\tau \sim p_\pi(\tau)}[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t)|s_t = s, a_t = a]$. The advantage function is then defined by $A^\pi(s, a; \mu, P, r) = Q^\pi(s, a; \mu, P, r) - V^\pi(s, a; \mu, P, r)$.

In this thesis, we address the multi-task reinforcement learning problem, where a collection of tasks share the same state and action spaces, discount factor, and transition function, but have different reward functions $r \in \{r_n\}_{n=1}^{N}$.

2.2 Constrained Markov Decision Process

The focus of the Constrained Markov Decision Process (CMDP) is to find a policy that maximizes return, only from the set of policies that obey the constraints. The constraints in CMDP are specified by a set of constraint reward functions $\{C_m\}_{m=1}^{M}$, where $C_m : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and a set of corresponding scalar constraint tolerance $\{\theta_m\}_{m=1}^{N}$. The set of policies that obey the constraints is denoted by

$$\Pi_C \doteq \{\pi \in \Pi : \forall m, J(\pi; \mu, P, C_m) \leq \theta_m\} \tag{2.1}$$

and to find an optimal policy in a CMDP is to solve the following optimization problem

$$\pi^* = \arg\max_{\pi \in \Pi_C} J(\pi) \tag{2.2}$$

2.3 Group Fairness

For our definition of fairness, we adopt the demographic parity notion, also commonly known as group fairness. It requires the outcomes experienced by individuals to be independent of their particular social group membership, where each social group is denoted as $z \in \mathcal{Z}$.

In the long-term group fairness problem, we ensure the expected return to be equal across all groups. We assume all groups share the same state and action spaces, discount factor, and reward functions, but each group has a different initial state distribution μ_z and a different transition function P_z, and the long-term group fairness for a single task r is defined as

$$J(\pi_i; \mu_i, P_i, r) = J(\pi_j; \mu_j, P_j, r) \tag{2.3}$$

In practice, we relax this constraint by introducing a positive slack variable $\epsilon > 0$ and ensure the difference in return is within this tolerance

$$|J(\pi_i; \mu_i, P_i, r) - J(\pi_j; \mu_j, P_j, r)| < \epsilon, \qquad \forall i \leq j; (i, j) \in \mathcal{Z}^2 \tag{2.4}$$

2.4 Reinforcement Learning from Human Feedback (RLHF)

The RLHF pipeline typically involves three key stages: supervised fine-tuning, reward modeling, and reinforcement learning.

Stage 1: Supervised Finetuning (SFT). In the first stage, a pre-trained language model is fine-tuned using supervised learning on task-specific datasets, such as dialogue, summarization, or instruction following, to create a reference policy denoted as π_{ref}.

Stage 2: Reward Modeling. The second stage, reward modeling, seeks to capture human preferences of LLMs responses. Let x be a prompt given to an LLM and y be the model's output response for the prompt. For each given input x, LLM will generate a pair of responses and human annotators are asked to express their preference between two output responses, with y_0 and y_1 denote the chosen and rejected responses respectively. These human preference data are used to train a reward model $r_\theta(x, y)$, which learns to predict which response is better according to human judgment. Formally, the reward model's loss derived from the Bradley-Terry (BT) preference model (Bradley and Terry, 1952) can be expressed as:

$$\text{loss}(r_\theta) = -\mathbb{E}_{(x, y_0, y_1) \sim D} \left[\log \left(\sigma \left(r_\theta(x, y_0) - r_\theta(x, y_1) \right) \right) \right],$$

where σ is the logistic function, and D is the dataset of human-annotated preferences.

Stage 3: Reinforcement Learning Finally, in the third stage, the learned reward model is used in reinforcement learning to further optimize the model denoted as π_ϕ, where ϕ is the weights of the LLM. The policy is trained to maximize the reward from the human feedback model while controlling for divergence from the initial supervised policy. The objective function of the reinforcement learning stage is usually given by:

$$\max_{\phi} \mathbb{E}_{y \sim \pi_{\phi}(\cdot|x)} r(x, y) - \beta D_{\mathrm{KL}}(\pi_{\phi}(y|x) \| \pi_{\mathrm{ref}}(y|x)), \tag{2.5}$$

where β controls the learned policy's deviation from the pretrained LLM as an initial reference policy π_{ref}.

2.5 Bias in Large Language Models

Most research on fairness and bias in LLMs has focused on reducing harm and risk in LLM generation through bias mitigation techniques. Techniques such as counterfactual data augementation (Lu et al., 2020), data filtering and selection (Garimella et al., 2022), designing specific prompting triggers (Venkit et al., 2023) and incorporating the notion group fairness in constructing a bias evaluation dataset Bi et al. (2023), have proven effective to reduce stereotypical or harmful language targeted at various demographic groups. Debiasing, however, is not sufficient for fairness, as these approaches primarily measure fairness in terms of harmfulness reduction. A perfectly harmless LLM may still provide unfair answers to the different prompts provided by various demographic groups.

Specifically, the evaluation and mitigation of counterfactual bias, often operationalized by switching group attributes (e.g., gender) at the prompt level, is a prevalent approach in assessing the fairness of large language models (LLMs). However, this approach typically require users from different demographic groups to share the same prompt questions for fairness evaluation (Nangia et al., 2020; Webster et al., 2021; Wang and Cho, 2019). While these efforts reduce stereotypical language toward specific groups, they do not account for the challenges posed by different prompt questions from various demographic groups. Furthermore, in real-world interactions with LLMs, users often do not explicitly state their sensitive attributes in prompts, making these methods inapplicable when demographic information is implicit rather than explicitly mentioned.

Chapter 3

Multi-task Group Fairness in Reinforcement Learning

We present our first contribution of the thesis in this chapter, and started by formulating the multi-task group fairness problem in reinforcement learning as a constrained Markov decision problem.

3.1 Formulating the Multi-Task Group Fairness in RL as a CMDP Problem

We are now ready to formulate the Group Fairness in Multi-Task Reinforcement Learning problem. We aim to ensure the long-term outcome experienced by different social groups to be equal, so we are not restricted to using a single policy for all social groups. Let $\boldsymbol{\pi}$ denotes a list of policies π. A social-group specific policy $\boldsymbol{\pi}_z$ is used to solve for each social group's specific transition P_z, and our goal is to find a list of optimal policies $\boldsymbol{\pi}^*$ that obey the relaxed group fairness constraint across all tasks $r_n \in \{r_n\}_{n=1}^{N}$

$$\begin{aligned} \boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi}} & \sum_{i}\sum_{r_n} J(\boldsymbol{\pi}_i; \mu_i, P_i, r_n) \\ & \text{s.t.} \max_{r_n} |J(\boldsymbol{\pi}_i; \mu_i, P_i, r_n) - J(\boldsymbol{\pi}_j; \mu_j, P_j, r_n)| \leq \epsilon, \qquad \forall i \leq j; (i,j) \in \mathcal{Z}^2, r_n \in \{r_n\}_{n=1}^{N} \end{aligned} \tag{3.1}$$

To practically tackle this problem, we first frame it as a CMDP problem and then use the constrained policy optimization algorithm to solve it.

In practice, instead of finding the list of all policies at the same time, each group's policy π_i is updated individually using a block coordinate descent approach, which may not give us the optimal solution of the original problem. Under this setting, the objective function can be simplified to only include the return of group i. Since the policies of other social groups are not updated, the returns of reward function r_n for other social groups remain constant, denoted as $\bar{J}_j(r_n) \doteq J(\boldsymbol{\pi}_j; \mu_j, P_j, r_n)$, which can be excluded from the objective function. Note that ensuring the maximum difference in return to be less than ϵ is equivalent to ensuring all differences in return to be less than ϵ, so the constraint in (5) can be written into N number of inequalities. Therefore, the objective and constraints can be rewritten as the following

$$\begin{aligned} &\boldsymbol{\pi}_i^* = \arg\max_{\pi} \sum_n J(\boldsymbol{\pi}_i; \mu_i, P_i, r_n) \\ &\quad s.t. \quad |J(\boldsymbol{\pi}_i; \mu_i, P_i, r_n) - \bar{J}_j| \leq \epsilon, \qquad \forall i \leq j; (i,j) \in \mathcal{Z}^2, r_n \in \{r_n\}_{n=1}^N \end{aligned} \tag{3.2}$$

All tasks in each social group will share the same distribution of trajectories $p_\pi(\tau)$ because each social group shares the same policy, initial state distribution, and transition function. Therefore, pulling out $p_\pi(\tau)$ as the common factor, the objective function can be written as

$$\sum_n J(\boldsymbol{\pi}_i; \mu_i, P_i, r_n) = \sum_n \mathrm{E}_{\tau \sim p_\pi(\tau)}[\sum_{t=1}^{\infty} \gamma^t r_n(s_t, a_t)] \tag{3.3}$$

$$= \sum_n \sum_\tau \mu_i(s_1) \Pi_{t=1}^{\infty} \pi(a_t|s_t) P_i(s_{t+1}|s_t, a_t) [\gamma^t r_n(s_t, a_t)] \tag{3.4}$$

$$= \sum_\tau \mu_i(s_1) \Pi_{t=1}^{\infty} \pi(a_t|s_t) P_i(s_{t+1}|s_t, a_t) [\gamma^t \sum_n r_n(s_t, a_t)] \tag{3.5}$$

$$= J(\boldsymbol{\pi}_i; \mu_i, P_i, \sum_n r_n) \tag{3.6}$$

Removing the absolute value in the constraint functions, the number of inequalities in the constraint will double from N to $2N$ as the following

$$
\begin{aligned}
&J(\boldsymbol{\pi}_i; \mu_i, P_i, r_n) \leq \epsilon + \bar{J}_j(r_n), && (3.7)\\
&-J(\boldsymbol{\pi}_i; \mu_i, P_i, r_n) \leq \epsilon - \bar{J}_j(r_n) \\
\iff\ &J(\boldsymbol{\pi}_i; \mu_i, P_i, -r_n) \leq \epsilon - \bar{J}_j(r_n), \qquad \forall i, j \text{ with } i \leq j;\ (i,j) \in \mathcal{Z}^2, r_n \in \{r_n\}_{n=1}^{N},
\end{aligned}
$$

by the linearity of expectation.

Our problem can be formulated into a CMDP problem:

To formulate our problem into a CMDP problem, let the constraint reward function be $C_m(s,a) = r_n(s,a)$ for the first N inequalities where $m \in 1, 2, ..., N$, and the corresponding constraint tolerance $\theta_m = \epsilon + \bar{J}_j(r_n)$. For the second N inequalities, we define the constraint reward function as $C_m(s,a) = -r_n(s,a)$ and set the constraint tolerance to $\theta_m = \epsilon - \bar{J}_j(r_n)$, where $m \in N+1, N+2, ..., 2N$.

Then, finding the optimal policy for a specific social group i is to solve the following CMDP problem

$$
\pi_i^* = \arg\max_{\pi \in \Pi_C} J(\pi; \mu_i, P_i, \sum_n r_n), \tag{3.8}
$$

where

$$
\Pi_C \doteq \{\pi \in \Pi : \forall m, i \leq j; (i,j) \in \mathcal{Z}^2, J(\pi, P_j, C_m) \leq \theta_m\} \tag{3.9}
$$

3.2 Constrained Policy Optimization Methodology

Constrained Policy Optimization (CPO) is one method that solves the CMDP problem. It has the advantage of maintaining constraint satisfaction throughout training, whereas other methods such as Primal-Dual Optimization Chow et al. (2015) only achieve constraint satisfaction after policy converges. As one of the trust region methods, CPO aims to maximize the next updated policy's performance improvement from the old policy of the current iteration: $J(\pi^{k+1}) - J(\pi^k)$, while keeping the new policy's costs within the tolerances, $J(\pi^{k+1}; \mu, P, C_m) \leq d_m$ for all cost function C_m and all tolerances d_m. To avoid the problem of off-policy evaluation for π^{k+1}, in practice, only a lower bound for the performance difference and an upper bound of the cost of the new policy that's dependent on d^{π} is used in the optimization.

The proposed CPO method is as follows

$$\begin{aligned} \pi^{k+1} = \underset{\pi_\theta \in \Pi_\theta}{\arg\max} & \underset{\substack{s \sim d^{\pi_k} \\ a \sim \pi_\theta}}{\mathrm{E}} [A^{\pi^k}(s, a; \mu, P, r)] \\ \text{s.t. } & J(\pi^k; \mu, P, C_m) + \frac{1}{1-\gamma} \underset{\substack{s \sim d^{\pi^k} \\ a \sim \pi_\theta}}{\mathrm{E}} [A^{\pi^k}(s, a; \mu, P, C_m)] \leq d_m \quad \forall m \\ & \underset{s \sim \pi^k}{\mathrm{E}} \left[D_{KL} \left(\pi_\theta(\cdot|s) || \pi^k(\cdot|s) \right) \right] \leq \delta. \end{aligned} \tag{3.10}$$

The original CPO algorithm relies on second-order Taylor approximation and inverting a high-dimensional Fisher information matrix. A first-order method, FOCOPS, is proposed by Zhang et al. (2020) for the CPO problem. To solve the group fairness problem, FOCOPS is required to handle more than one constraint. In the following Algorithm 1, we extended the FOCOPS algorithm for multiple constraints. In Algorithm 2, the multi-objective group fairness reinforcement learning algorithm is proposed.

Algorithm 1: First Order Constrained Optimization in Policy Space (FOCOPS) for M constraints

```
Input: Initial policy parameters $\theta^0$, initial value function parameters $\phi^0$, initial cost value function parameters $\{\psi_m^0\}_{m=1}^M$, Cost functions $\{C_m\}_{m=1}^M$, Cost tolerances $\{b_m\}_{m=1}^M$.
Output: Final policy parameters $\theta^{\text{final}}$, Final value function parameters $\phi^{\text{final}}$, Final cost value function parameters $\{\psi_m^{\text{final}}\}_{m=1}^M$.
Hyperparameters: Discount rates $\gamma$, GAE parameter $\beta$; Learning rates $\alpha_\nu, \alpha_V, \alpha_\pi$; Temperature $\lambda$; Initial cost constraint parameter $\nu$; Cost constraint parameter bound $\nu_{\max}$. Trust region bound $\delta$.
while Stopping criteria not met do
    Generate batch data of $H$ episodes of length $T$ of $\left(s_{i,t}, a_{i,t}, r_{i,t}, s_{i,t+1}, \{c_{m,i,t}\}_{m=1}^M\right)$ from $\pi_\theta$, where $i = 1 \dots, H, t = 1, \dots, T$.
    for $m \in 1, 2, \dots, M$ do
        For cost function $m$, estimate cost-return by averaging over $C$-return for all episodes:
            $$\hat{J}_{C_m} = \frac{1}{M}\sum_{i=1}^{M}\sum_{t=0}^{T-1}\gamma^t c_{m,i,t}$$
    Store old policy $\theta' \leftarrow \theta$ Estimate advantage functions $\hat{A}_{i,t}$ and $\{\hat{A}_{i,t}^{C_m}\}_{m=1}^M, i = 1, \dots, H, t = 1, \dots, T$ using GAE.
    Get $V_{i,t}^{\text{target}} = \hat{A}_{i,t} + V_\phi(s_{i,t})$ and $V_{i,t}^{C_m,\text{target}} = \hat{A}_{i,t}^{C_m} + V_{\psi_m}^{C_m}(s_{i,t})$, for $m \in 1, 2, \dots, M$.
    for $m \in 1, 2, \dots, M$ do
        Update $\nu_m$ by: $\nu_m \leftarrow \underset{\nu_m}{\text{proj}}\left[\nu_m - \alpha_{\nu_m}\left(b - \hat{J}_{C_m}\right)\right]$
    for $K$ epochs do
        for each minibatch $\left\{s_j, a_j, A_j, \{A_j^{C_m}\}_{m=1}^M, V_j^{target}, \{V_j^{C_m,target}\}_{m=1}^M\right\}$ of size $B$ do
            Update value loss functions: $\mathcal{L}_V(\phi) = \frac{1}{2N}\sum_{j=1}^{B}\left(V_\phi(s_j) - V_j^{\text{target}}\right)^2$
            for $m \in 1, 2, \dots M$ do
                $$\mathcal{L}_{V_m^C}(\psi_m) = \frac{1}{2N}\sum_{j=1}^{B}\left(V_{\psi_m^{C_m}}(s_j) - V_j^{C_m,\text{target}}\right)^2$$
            Update value networks: $\phi \leftarrow \phi - \alpha_V \nabla_\phi \mathcal{L}_V(\phi)$
            for $m \in 1, 2, \dots, M$ do
                $$\psi_m \leftarrow \psi_m - \alpha_V \nabla_{\psi_m}\mathcal{L}_{V^{C_m}}(\psi_m)$$
            Update policy: $\theta \leftarrow \theta - \alpha_\pi \hat{\nabla}_\theta \mathcal{L}_\pi(\theta)$, where
                $$\hat{\nabla}_\theta \mathcal{L}_\pi(\theta) \approx \frac{1}{B}\sum_{j=1}^{B}\left[\nabla_\theta D_{\text{KL}}\left(\pi_\theta \| \pi_{\theta'}\right)[s_j] - \frac{1}{\lambda}\frac{\nabla_\theta \pi_\theta(a_j \mid s_j)}{\pi_{\theta'}(a_j \mid s_j)}\left(\hat{A}_j - \sum_{m=1}^{M}\nu_m \hat{A}_j^{C_m}\right)\right]\mathbf{1}_{D_{\text{KL}}(\pi_\theta \| \pi_{\theta'})[s_j] \le \delta}$$
                if $\frac{1}{HT}\sum_{i=1}^{H}\sum_{t=0}^{T-1} D_{\text{KL}}\left(\pi_\theta \| \pi_{\theta'}\right)[s_{i,t}] > \delta$ then
                Break out of inner loop
```

Algorithm 2: Outline of the Multi-Task Fairness RL Algorithm

```
Input: Initial policy parameters $\theta_z^0, \forall z \in |\mathcal{Z}|$, initial value function parameters $\phi_z^0$, initial cost value function parameters
       $\psi_{m,z}^0, \forall z \in |\mathcal{Z}|, m \in 1, 2, ..., M$, where $M = (|Z| - 1)2N$.
Output: Final policy parameters $\theta_z^{\text{final}}$, final value function $\phi_z^{\text{final}}$, and final cost function parameters $\{\psi_{m,z}^{\text{final}}\}_{m=1}^M$,
for each group $z$
Initialize: Group fairness threshold $\epsilon$, M constraint funcitons $\mathbf{C}$, M constraint thresholds $\mathbf{b}$, $m = 1$. for $k = 0, 1, 2, \ldots$ do
    // Calculate performance estimates of policies for all groups.
    for $z \in |\mathcal{Z}|$ do
        for $z_1 \in |\mathcal{Z}|$ do
            for $i \in 1, 2, \ldots, H$ do
                Sample the $i$th trajectory of length $T$ for group $z_1$: $(s_{i,t}, a_{i,t}, \{r_{n,i,t}\}_{n=1}^N, s_{i,t+1})$, for $t = 1, \ldots, T$.
                for $n \in 1, 2, ..., N$ do
                    Use the Monte Carlo Method to estimate the return $\bar{J}_z(r_n)$ of policy $\pi_z$ at reward function $r_n$:
                        $$\bar{J}_z(r_n) = \frac{1}{H} \sum_{i=1}^{H} \sum_{t=0}^{T-1} \gamma^t r_{n,i,t}$$
                    if $z \neq z_1$ then
                        Set the cost functions as the reward function and the negative reward function:
                            $$\mathbf{C}[m] = r_n$$
                            $$\mathbf{C}[m+1] = -r_n$$
                        Calculate the thresholds for M constraints:
                            $$\mathbf{b}[m] = \epsilon + \bar{J}_z(r_n)$$
                            $$\mathbf{b}[m+1] = \epsilon - \bar{J}_z(r_n)$$
                            $$m = m + 1$$
        Update the parameters for policy, value function, and cost functions of group $z$ by
            $\theta_z^{k+1}, \phi_z^{k+1}, \{\psi_{m,z}^{k+1}\}_{m=1}^M = \mathbf{FOCOPS}(\theta_z^k, \phi_z^k, \{\psi_{m,z}^k\}_{m=1}^M, \mathbf{C}, \mathbf{b})$.
```

18 3.3 Experimental Results

In the experiments, we compare our Multi-Task Group Fairness algorithm (MTGF) to a Group Fairness in Reinforcement Learning (GFRL) algorithm. The original GFRL algorithm imposes a fairness constraint on only one task, as it was designed for single-task settings. Applying this algorithm to multiple tasks leaves other tasks unconstrained, leading to violations of the fairness threshold. To ensure a fair comparison, we alternate the single-task constraint across the two tasks during training.

For the environments for experiments, we use the customized environment from the work from (Satija et al., 2023), which are modified from the Half-Cheetah-v3 environment from the OpenAI gym Brockman et al. (2016) to create three additional subgroups with different dynamics: one BigFoot HalfCheetah with 2× the feet size of the default Half-Cheetah-v3, , one TenFriction HalfCheetah with 10× friction than the default setting, and another HugeGravity HalfCheetah with 1.5x gravity than the default setting.

Three experiments were conducted between two social groups as detailed in Table 3.1. In each of the experiment, the algorithms is trained to control the locomotion of the original HalfCheetah, and one of the following three customized HalfCheetahs: HugeGravity HalfCheetah, TenFricion HalfCheetah, and BigFoot HalfCheetah.

Table 3.1: Summary of Experiments with HalfCheetah Variants Across Social Groups and Tasks

Experiment	Social Group A	Social Group B	Tasks
1	Original HalfCheetah	HugeGravity HalfCheetah	Backward, Forward Running
2	Original HalfCheetah	BigFoot HalfCheetah	Backward, Forward Running
3	Original HalfCheetah	TenFriction HalfCheetah	Backward, Forward Running

Results show that the Multi-Task Group Fairness algorithm achieved a smaller fairness gap while having a comparable mean reward to the GFRL algorithm. Specifically, the fairness gap for the original HalfCheetah and HugeGravity HalfCheetah for the backward running task presented in Figure 3.5, and for the forward running task presented in Figure 3.6; the fairness gap for the original HalfCheetah and BigFoot HalfCheetah for the backward running task presented in Figure 3.11, and for the forward running task presented in Figure 3.12; the fairness gap for the original HalfCheetah and TenFriction HalfCheetah for the backward running task presented in Figure 3.17, and for the forward running task presented in Figure 3.18 together demonstrate that our MTGF algorithm better ensures fairness than the GFRL algorithm.

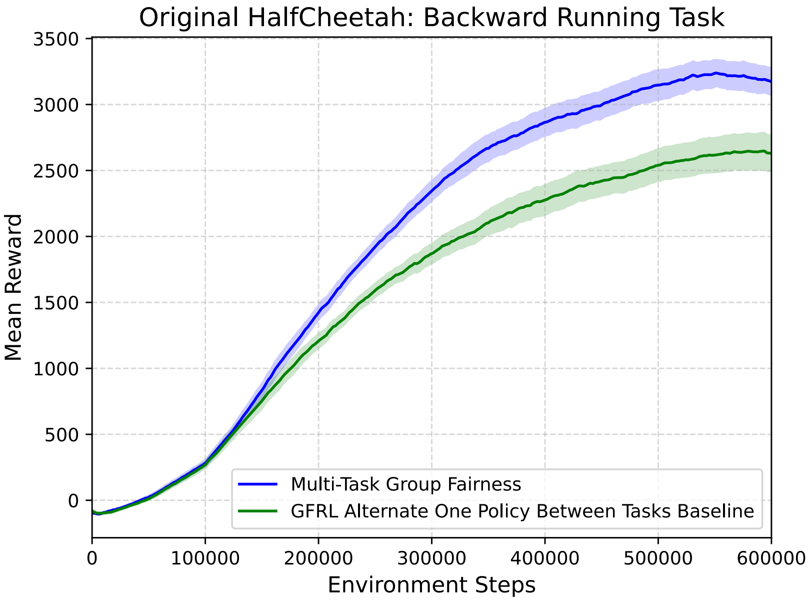

Figure 3.1: Training the Original HalfCheetah on the Backward Running Task.

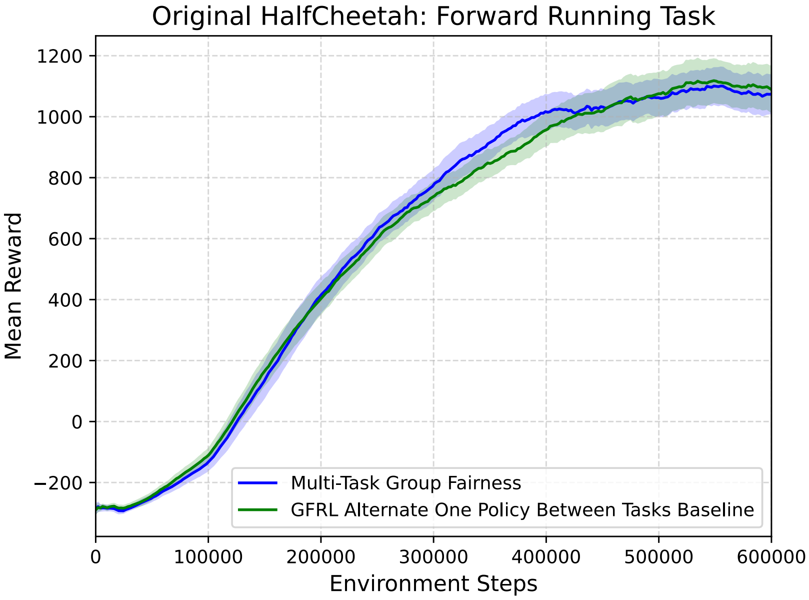

Figure 3.2: Training the Original HalfCheetah on the Forward Running Task.

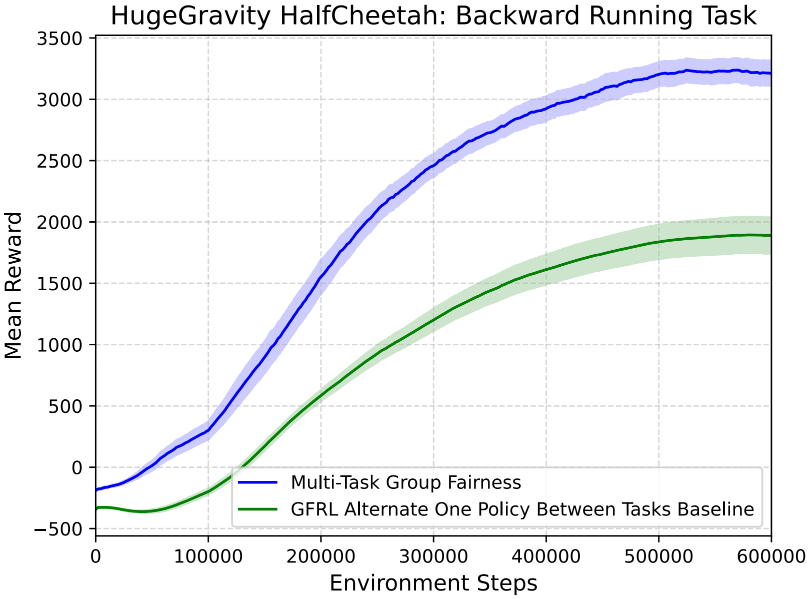

Figure 3.3: Training the HugeGravity HalfCheetah on the Backward Running Task.

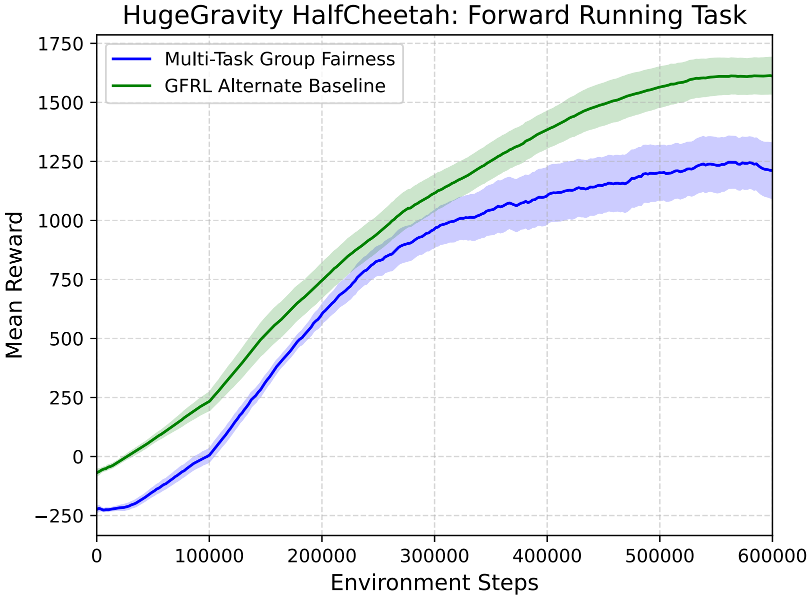

Figure 3.4: Training the HugeGravity HalfCheetah on the Forward Running Task.

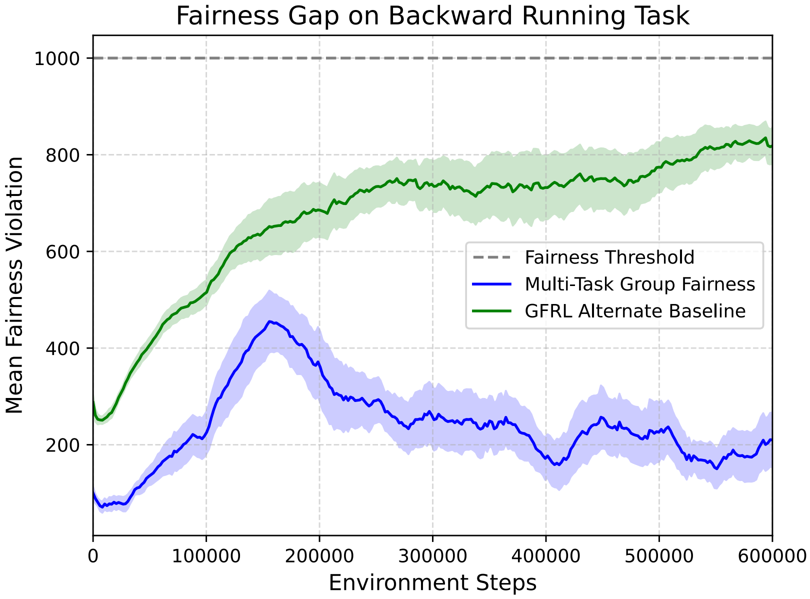

Figure 3.5: Fairness Gap on the Backward Running Task between the Original HalfCheetah and HugeGravity HalfCheetah .

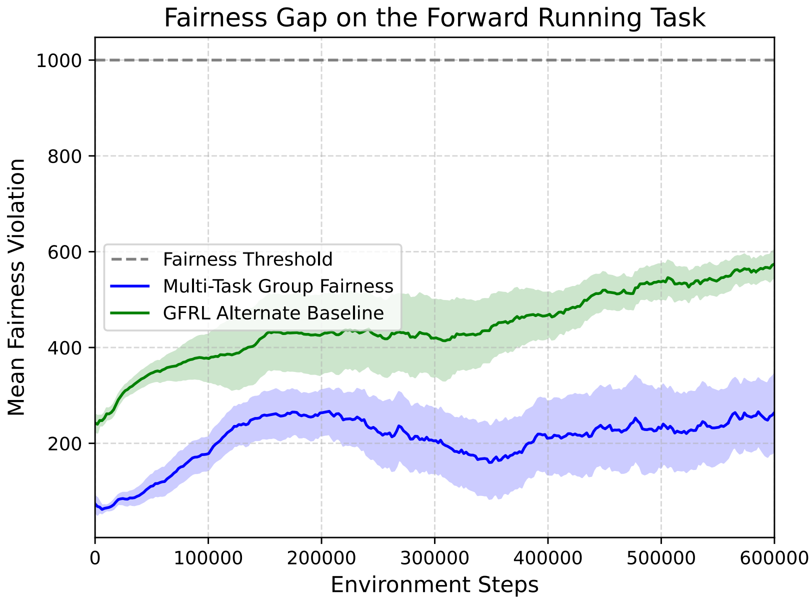

Figure 3.6: Fairness Gap on the Forward Running Task between the Original HalfCheetah and HugeGravity HalfCheetah.

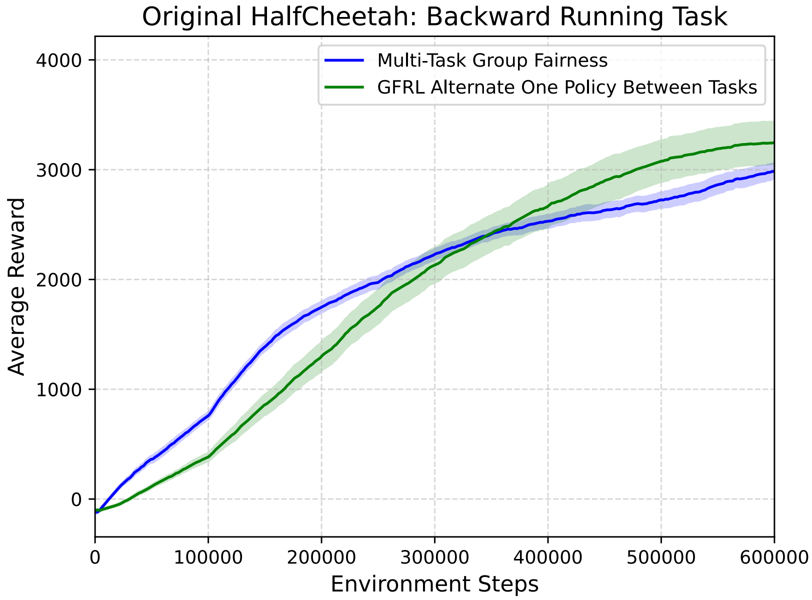

Figure 3.7: Training the Original HalfCheetah on the Backward Running Task.

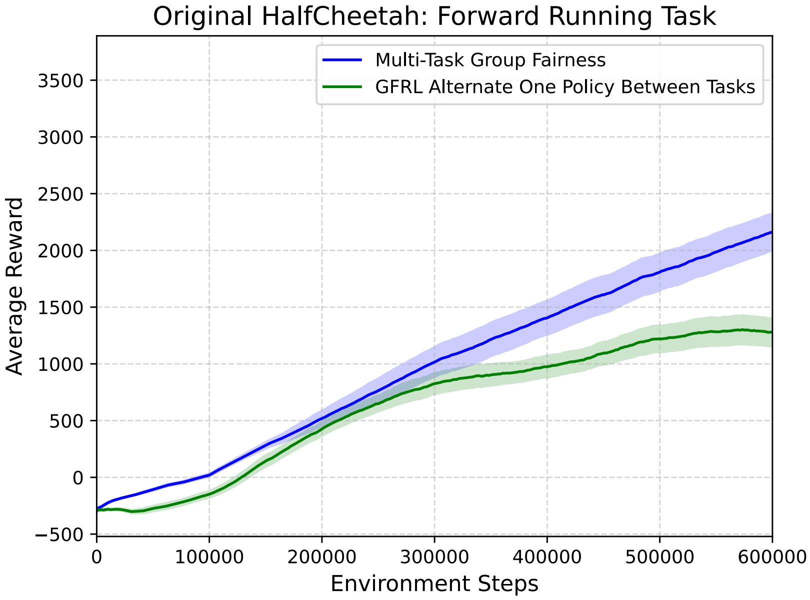

Figure 3.8: Training the Original HalfCheetah on the Forward Running Task.

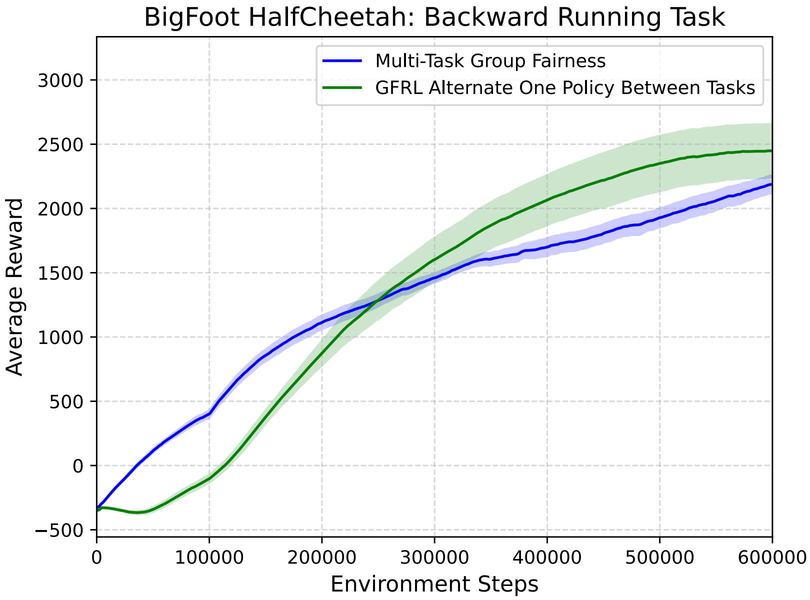

Figure 3.9: Training the BigFoot HalfCheetah on the Backward Running Task.

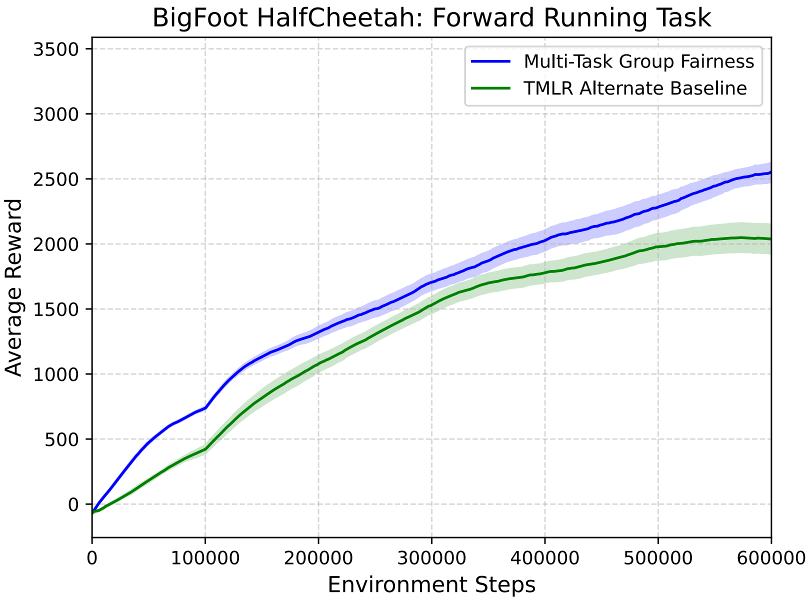

Figure 3.10: Training the BigFoot HalfCheetah on the Forward Running Task.

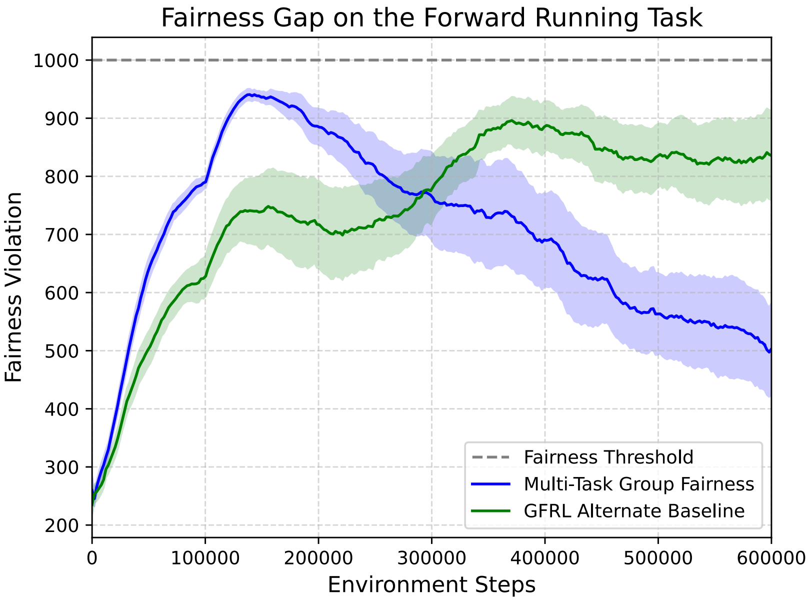

Figure 3.11: Fairness Gap on the Forward Running Task between the Original HalfCheetah and BigFoot HalfCheetah.

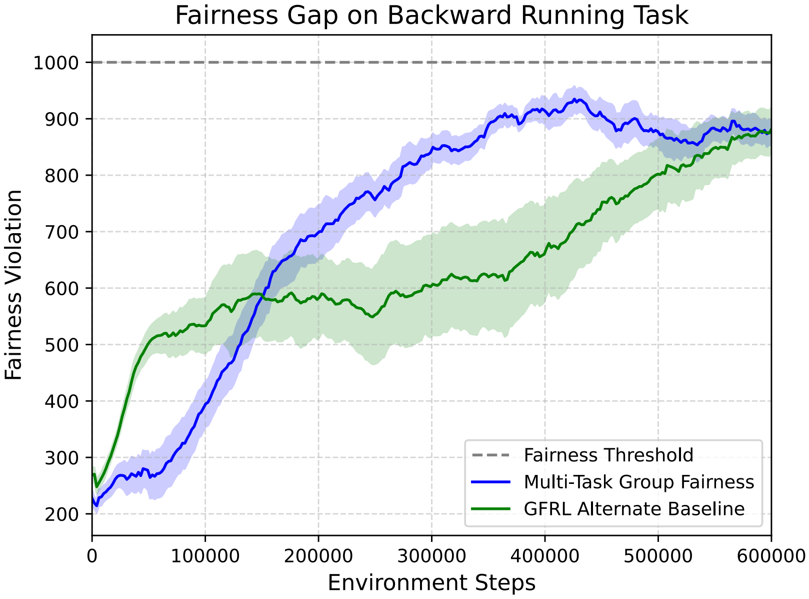

Figure 3.12: Fairness Gap on the Backward Running Task between the Original HalfCheetah and BigFoot HalfCheetah.

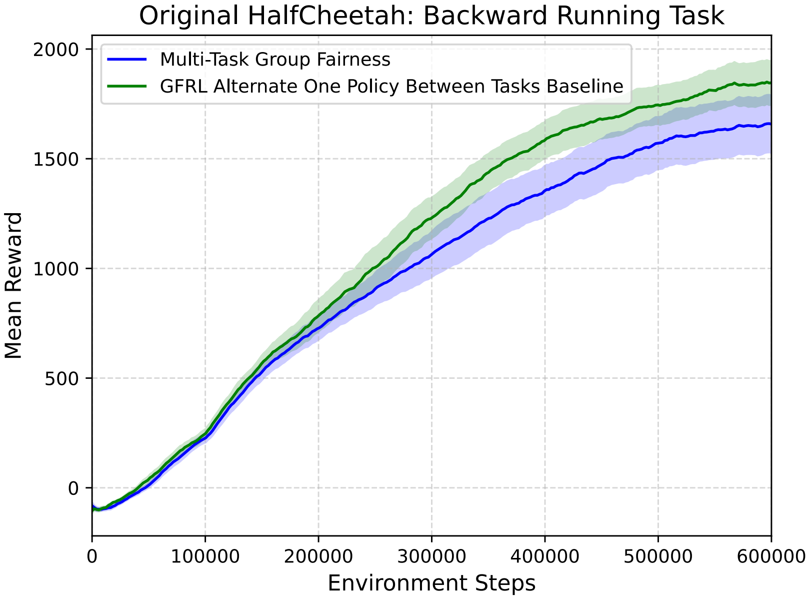

Figure 3.13: Training the Original HalfCheetah on the Backward Running Task.

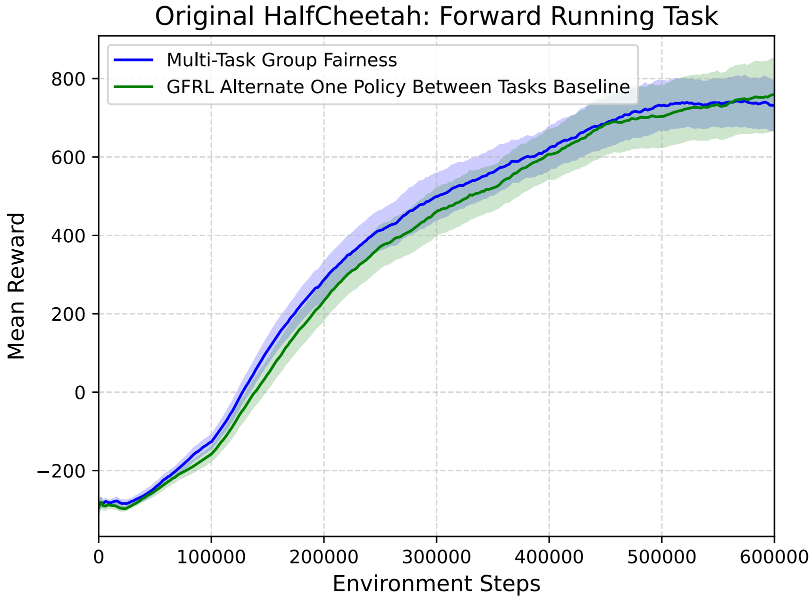

Figure 3.14: Training the Original HalfCheetah on the Forward Running Task.

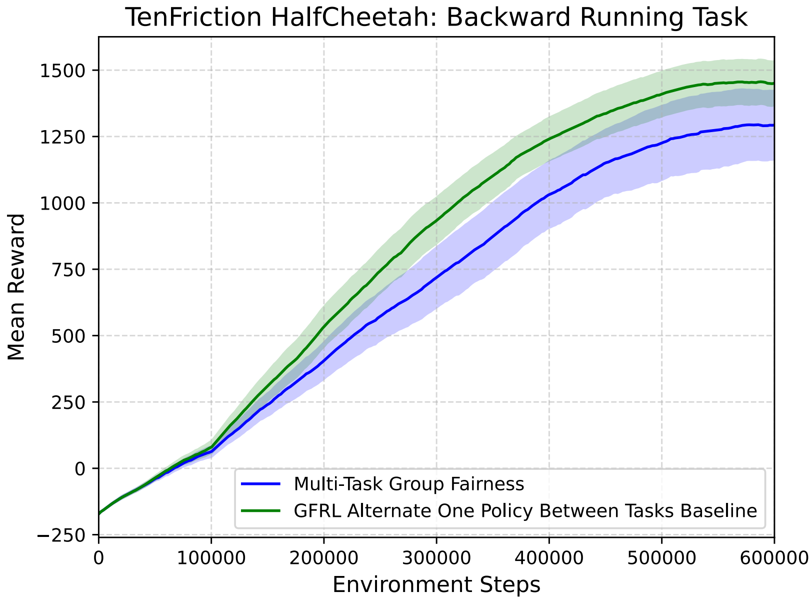

Figure 3.15: Training the TenFriction HalfCheetah on the Backward Running Task.

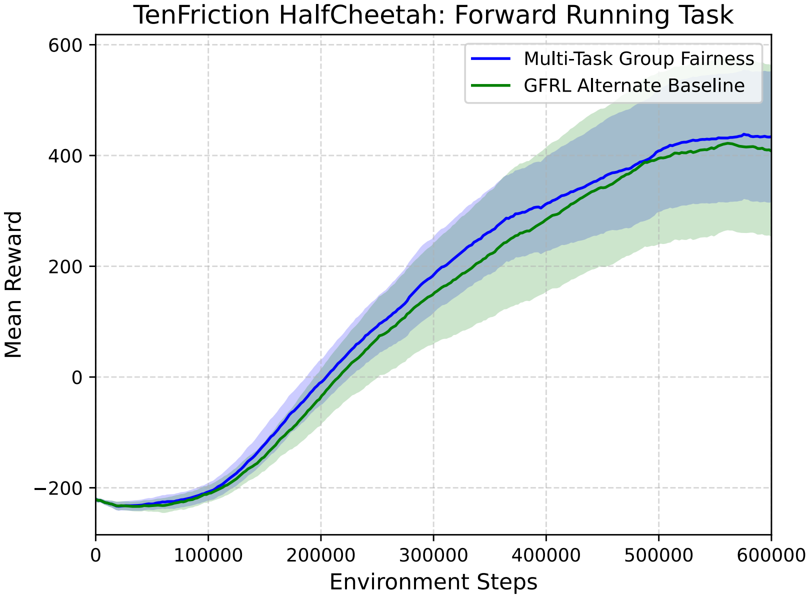

Figure 3.16: Training the TenFriction HalfCheetah on the Forward Running Task.

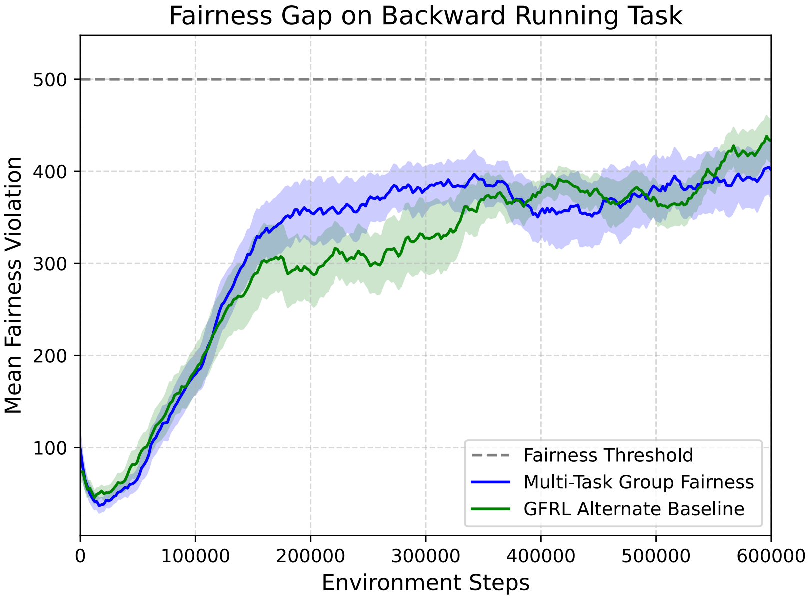

Figure 3.17: Fairness Gap on the Backward Running Task between the Original HalfCheetah and TenFriction HalfCheetah.

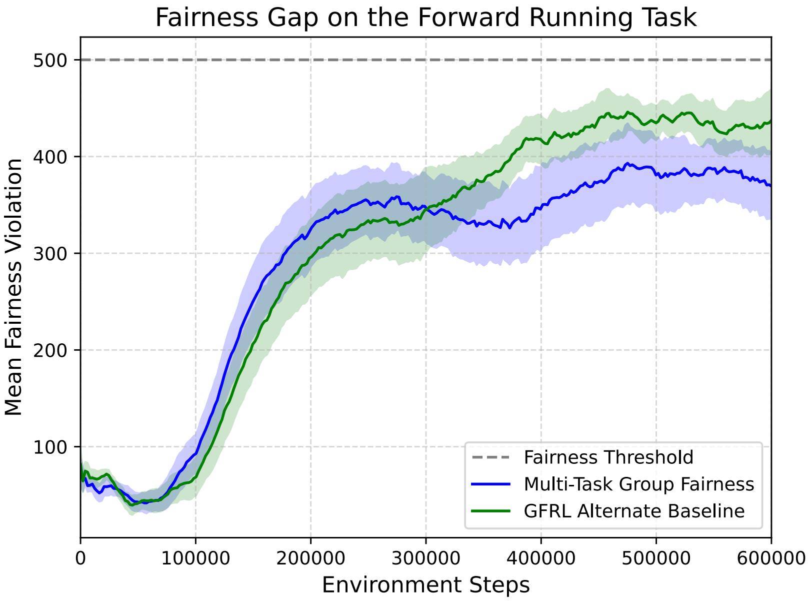

Figure 3.18: Fairness Gap on the Forward Running Task between the Original HalfCheetah and TenFriction HalfCheetah.

3.4 Zero Multi-Task Fairness Constraint Violation

In this section, we present a result of zero multi-task fairness constraint Violation, stating that assuming we have access to an initial fair policy, the fairness guarantees for any of the subgroups throughout the learning duration for all tasks would not be violated with high probability.

A multi-task finite-horizon Markov Decision Process (MDP) is defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, \{r_m\}_{m=1}^{M}, \mu)$, where $\mathcal{S}$ is the state space, H number of steps in each episode, and $P(\cdot|s,a) \in \Delta_{\mathcal{S}}^{H}, \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$, where $\Delta_{\mathcal{S}}$ is the $|\mathcal{S}|$-dimensional probability simplex. The tasks within MDP are characterized by distinct reward functions $\{r_m\}_{m=1}^{M}$, where $r_m : \mathcal{S} \times \mathcal{A} \leftarrow [0,1]$ specifies the reward function for each task m, and M is the total number of tasks. The algorithm samples a total of K episodes from the environment. We assume the initial state distribution μ is known to the agent and reward functions are deterministic.

In the context of group fairness, each group is denoted as $z \in \mathcal{Z}$ and we assume all groups share the same state space, action space, and reward functions, but each group has a different initial state distribution μ_z and a different transition function P_z. The return of policy π under transition P_z and initial transition distribution μ_z is denoted as $J_z^{\pi}(r, P_z)$. The fairness threshold for the acceptable performance difference between any two groups is denoted as $\epsilon : \epsilon \in (0, H]$.

The multi-task group fairness RL problem is formulated as finding a list of optimal policies $\boldsymbol{\pi}^*$ that obey the group fairness constraint across all tasks $m \in [M]$:

$$\boldsymbol{\pi}^* = \arg\max_{\pi} \sum_{z} \sum_{r_m} J^{\pi}(r_m, P_z) \tag{3.11}$$

$$\text{s.t.} \qquad \max_{m}(|J_i^{\pi}(r_m, P_i) - J_j^{\pi}(r_m, P_j)|) \leq \epsilon, \quad \forall i \geq j;\ (i, j) \in \mathbb{Z}^2, \forall m \in [M]. \tag{3.12}$$

To ensure the above problem is feasible, we assume there exist an initial strictly fair policy π_0 that our algorithm can use to safely sample data from the environment.

Assumption 1.1 (Initial strictly fair policy). The algorithm has access to a policy π that satisfies the fairness constraints in Equation (3.12). We also assume $\left|J_i^{\pi^0}(r_m, P_i) - J_j^{\pi^0}(r_m, P_j)\right|$ $\leq \epsilon^0 < \epsilon, \forall (i, j) \in \mathbb{Z}^2, \forall m \in [M]$ and the value of ϵ^0 is known to the algorithm.

One key objective of our work is to ensure that our algorithm does not violate the group fairness constraint in Equation (3.12) during training, where the fairness gap is calculated by the absolute difference between the returns of two groups. We seek to construct a set of policies that obey the group fairness constraint, and then find the policy with maximum return within the set. However, the true transition P is unknown to our algorithm and we can only estimate the fairness gap by sampling from the true environment to evaluate the returns of different groups. A poor estimation of the fairness gap may result in selecting a policy whose true fairness gap violates the fairness constraint by a large margin.

To address this issue, we aim to construct a conservative set of policies that will achieve zero-fairness-constraint violation with high probability. Following the techniques from Satija et al. (2023), we design an optimistic estimation of the fairness gap and then select policies whose optimistic fairness gap is less than or equal to the fairness threshold ϵ to construct the conservative set of policies.

Designing the optimistic fairness gap requires an optimistic reward $\bar{r}^k_{m,h}$ and a pessimistic reward $\underline{\mathrm{r}}^k_{m,h}$ defined as:

$$\bar{r}^k_{m,h}(s,a) \doteq r^k_h(s,a) + |\mathcal{S}|H\beta^k_h(s,a) \tag{3.13}$$

$$\underline{\mathrm{r}}^k_{m,h}(s,a) \doteq r^k_h(s,a) - |\mathcal{S}|H\beta^k_h(s,a), \tag{3.14}$$

where $\beta^k_h(s,a)$ is the confidence radius to account for the uncertainties from the transition probabilities.

Taking a model-based policy evaluation approach, the return of the policy is evaluated using an estimated transition $\hat{P}^k_z$. The optimistic and pessimistic reward estimates then allow us to calculate the difference between an optimistic return from one group and a pessimistic return from the other group, which gives us the optimistic fairness gap. Selecting policies that obey the fairness threshold for every task m, a set of safe policies can be constructed as follows:

$$\mathbf{\Pi}^k_F \left\{ \pi : \begin{array}{l} J^\pi_i(\bar{r}^k_m, \hat{P}^k_i) - J^\pi_j(\underline{\mathrm{r}}^k_m, \hat{P}^k_j) \leq \epsilon, \quad \forall i \geq j;\ (i,j) \in \mathbb{Z}^2, \forall m \in [M]. \\ J^\pi_j(\bar{r}^k_m, \hat{P}^k_j) - J^\pi_i(\underline{\mathrm{r}}^k_m, \hat{P}^k_i) \leq \epsilon, \quad \forall i \geq j;\ (i,j) \in \mathbb{Z}^2, \forall m \in [M]. \end{array} \right\} \tag{3.15}$$

When the transitions are poorly estimated, it is possible that no policy obeys the constraint. In case the above policy set is empty, we can simply use the strictly fair policy π_0 that will not violate the fairness constraint in the true MDP to sample more data for better-estimated transitions $\hat{P}_z$. Executing π_0 under the condition in the following is sufficient to guarantee that $\mathbf{\Pi}^k_F$ is non-empty in the otherwise condition.

Let the conservative set of policies $\mathbf{\Pi}^k$ be defined as follows:

$$\mathbf{\Pi}^k = \begin{cases} \{\pi^0\}, & \begin{cases} \text{if } J_i^{\pi^0}(\bar{r}_m^k, \hat{P}_i^k) - J_j^{\pi^0}(\underline{\mathrm{r}}_m^k, \hat{P}_j^k) > (\epsilon + \epsilon^0)/2, & \forall i \geq j; (i,j) \in \mathbb{Z}^2, \exists m \in [M] \\ \text{or } J_j^{\pi^0}(\bar{r}_m^k, \hat{P}_j^k) - J_i^{\pi^0}(\underline{\mathrm{r}}_m^k, \hat{P}_i^k) > (\epsilon + \epsilon^0)/2, & \forall i \geq j; (i,j) \in \mathbb{Z}^2, \exists m \in [M] \end{cases} \\ \mathbf{\Pi}_F, & \text{otherwise.} \end{cases} \tag{3.16}$$

We now present a result stating that policies chosen from $\mathbf{\Pi}^k$ do not violate the fairness guarantees for any of the subgroups throughout the learning duration with high probability.

Theorem 1.1(Fairness violation) Given an input confidence parameter $\delta \in (0,1)$ and an initial fair policy π^0, the construction of $\mathbf{\Pi}^k$ ensures that there are no fairness violations at any episode in the learning procedure in the true environment with high probability $(1-\delta)$, i.e., for any $\pi \in \mathbf{\Pi}^k$,

$$\Pr(|J_i^{\pi}(r_m, P_i) - J_j^{\pi}(r_m, P_j)| \leq \epsilon) \geq 1 - \delta, \quad \forall m \in [M], \forall k \in [K], \forall i \geq j; (i,j) \in \mathcal{Z}^2.$$

Besides zero fairness violation, we also care about achieving sub-linear regret. Under the principle of optimism under the face of uncertainty, we set another exploration bonus for reward function of each task m to achieve efficient exploration.

$$\ddot{r}_{m,h}^k(s,a) = r_{m,h}^k(s,a) + \alpha\beta_h^k(s,a), \tag{3.17}$$

where $\alpha = |\mathcal{S}|H + \frac{4|\mathcal{S}|H}{\epsilon - \epsilon^0} 2H$.

At each episode k, we will solve the following optimization problem:

$$\pi^k \in \arg\max_{\pi \in \Pi^k} \sum_{m}^{M} \sum_{z \in \mathcal{Z}} J^{\pi}(\ddot{r}_m, \hat{P}_z) \tag{3.18}$$

Theorem 1.2 (Regret Bound). For any $\delta \in (0,1)$, with probability $1-\delta$, for any task m, executing π^k from Equation (3.18) at every episode $k \in [K]$ incurs in a regret of at most

$$\mathrm{Reg}(K; r_m) = \tilde{\mathcal{O}}\left(\frac{|\mathcal{Z}|H^3}{(\epsilon - \epsilon^0)} \sqrt{|\mathcal{S}|^3|\mathcal{A}|K} + \frac{|\mathcal{Z}|^2 M H^5 |\mathcal{S}|^3 |\mathcal{A}|}{\min\{(\epsilon - \epsilon^0), (\epsilon - \epsilon^0)^2\}} \right),$$

where $\tilde{O}(\cdot)$ hides polylogarithmic terms.

3.5 High Probability Good Event

Our subsequent analysis on performance guarantees depends on establishing a high probability "good" event $\boldsymbol{\mathcal{E}}$.

Let $\{\mathcal{F}_k\}_{k\leq 0}$ denotes the filtration with $\mathcal{F}_k = \sigma\left((S^{k'}_{z,h}, A^{k'}_{z,h}, R^{k'}_{m,z,h})_{z\in\mathcal{Z},h\in[H],m\in[M],k'\in[k]}\right) \forall k \in [K]$, and $\mathcal{F}_0$ denotes the trivial sigma algebra. The sequence of deployed policy $\{\pi^k\}_{k\in[K]}$ is predictable with respect to the filtration $\{\mathcal{F}_k\}_{k\leq 0}$.

$N^k_{z,h}(s,a)$ denotes the number of times the state-action tuple (s,a) for group z was observed at time step h in the episodes [1, . . . , k-1]. The expectation operator $E_{\mu_z,P_z,\pi}[\cdot]$ is the expectation with respect to the stochastic trajectory $(S_h, A_h)_{h\in[H]}$ generated according to the markov chain induced by (μ_z, P_z, π).

For each $(z,s,a,h) \in \mathcal{Z}\times\mathcal{S}\times\mathcal{A}\times[H]$, the empirical estimates of the transition is defined as:

$$\hat{P}^k_{z,h}(s'|s,a)\frac{\sum_{k'=1}^{k-1} 1(S^{k'}_{z,h} = s, A^{k'}_{h} = a, S^{k'}_{z,h+1} = s')}{\max(N^k_{z,h}(s,a),1)} \tag{3.19}$$

We define the event $\mathcal{E}_{\mathcal{G}}$ for the event sequence $\mathcal{G}_k \in \mathcal{F}_{k-1}, \forall k \in [K]$:

$$\mathcal{E}_{\mathcal{G}}(\delta)\doteq \big\{\forall K' \in [K]. \tag{3.20}$$

$$\sum_{k=1}^{K'}\sum_{h=1}^{H}\sum_{z,s,a}\frac{1(\mathcal{G}_k)d^{\pi^k}_{z,h}(s,a)}{\max(N^k_{z,h}(s,a),1)} \leq 4H|\mathcal{Z}||S||A| + 2H|\mathcal{Z}||S||A|\ln K'_{\mathcal{G}} + 4\ln\frac{2HK}{\delta}, \tag{3.21}$$

$$\sum_{k=1}^{K'}\sum_{h=1}^{H}\sum_{z,s,a}\frac{1(\mathcal{G}_k)d^{\pi^k}_{z,h}(s,a)}{\sqrt{\max(N^k_{z,h}(s,a),1)}} \leq 6H|\mathcal{Z}||S||A| + 2H\sqrt{|\mathcal{Z}||S||A|\ln K'_{\mathcal{G}}} + 2H|\mathcal{Z}||S||A|\ln K'_{\mathcal{G}} + 5\ln\frac{2HK}{\delta}, \big\}, \tag{3.22}$$

where $K'_{\mathcal{G}}\doteq\sum_{k=1}^{K'} 1(\mathcal{G}_k)$ and $d^{\pi^k}_{z}$ is the occupancy measure of policy π^k such that $d^{\pi^k}_{z,h}(s,a) = E_{\mu_z,P_z,\pi^k}[1(S_{z,h} = s, A_h = a|\mathcal{F}_{k-1})]$.

Let $\mathcal{E}_{\Omega}(\delta)$ be the event with the event sequence $\mathcal{G}_k = \Omega, \forall k \in [K]$, where Ω is the sample space. let $\mathcal{E}_0(\delta)$ denote $\mathcal{E}_{\mathcal{G}'}$, for the event that we choose the strictly safe policy π^0, with the event sequence

$$\mathcal{G}'_{1:K} = \left\{ J^{\pi^0}(\hat{P}_i^k, \bar{r}_m^k) - J^{\pi^0}(\hat{P}_j^k, \underline{\mathrm{r}}_m^k) \leq (\epsilon + \epsilon^0)/2, \forall i, j \in \mathcal{Z}^2, m \in [M] \right\} \tag{3.23}$$

Our subsequent analysis on performance guarantees depends on establishing a high probability "good" event $\boldsymbol{\mathcal{E}}$.

Good Event $\mathcal{E}$ is defined as:

$$\mathcal{E} \doteq \Big\{ \forall k \in [K], \forall h \in [H], \forall z \in \mathcal{Z}, \forall s \in \mathcal{S}, \forall a \in \mathcal{A}, \tag{3.24}$$

$$|P_{z,h}^k(s'|s,a) - \hat{P}_{z,h}^k(s'|s,a)| \leq \beta_{z,h}^k(s,a), \forall s' \in \mathcal{S} \Big\} \cap \mathcal{E}_\Omega(\delta/4) \cap \mathcal{E}_0(\delta/4), \tag{3.25}$$

where $\hat{\beta}_{z,h}^k(s,a) \sqrt{\frac{1}{\max(N_{z,h}^k(s,a),1)} \mathrm{C}}$ and $C \log(2|\mathcal{Z}||S|^2|A|HK/\delta)$

Lemma C.1 *Fix any $\delta \in (0,1)$, the good event $\mathcal{E}$ occurs with probability at least $1 - \delta$.*

Proof of Lemma C.1 For each $(z, s, a, h) \in \mathcal{Z} \times \mathcal{S} \times \mathcal{A} \times [H]$, we take K mutually independent samples of next states from the distribution specified by the true MDP model:

$$\{S_z^n(s,a,h)\}_{n=1}^K. \tag{3.26}$$

Let $\hat{P}_{z,h}^n$ be running empirical means for the samples

$$\{S_z^i(s,a,h)\}_{i=1}^n. \tag{3.27}$$

We can define the failure event:

$$F_n^P \doteq \{\exists z, s, a, s', h : |P_{z,h}(s'|s,a) - \hat{P}_{z,h}^n(s'|s,a)| \geq \beta(n)\}, \tag{3.28}$$

We define a generated event $\mathcal{E}^{gen}$,

$$\mathcal{E}^{gen} \doteq \left(\cup_{n=1}^K (F_n^P) \right)^C \cap \mathcal{E}_\Omega(\delta/4) \cap \mathcal{E}_0(\delta/4) \tag{3.29}$$

Let $n_{z,k}(s,a,h)$ denote the quantity $N_{z,h}^k(s,a) + 1$. Then the problem in our setting can be simulated as follows: for group z, at an episode k, taking action a in state s at time-step

h, we get the sample $(S_z^{n_{z,k}(s,a,h)}(s,a,h))$. Therefore, the set

$$\{S_z^n(s,a,h)\}_{n=1}^K \tag{3.30}$$

already contains all the samples drawn in the learning problem and the sample averages calculated by the algorithms are:

$$\hat{P}_{z,h}^k(s'|s,a) = P_z^{n_k(z,\tilde{s},a,h)}(\cdot|s,a,h). \tag{3.31}$$

As a result, the $\mathcal{E}^{gen}$ implies $\mathcal{E}$, and it is sufficient to show that $\mathcal{E}^{gen}$ occurs with probability at least $1-\delta$.

Using Lemma 8 and union bound, $\mathcal{E}_\Omega(\delta/4) \cap \mathcal{E}_0(\delta/4)$ occurs with probability at least $1-\delta/2$. To see this, let A denotes $\mathcal{E}_\Omega(\delta/4)$ and let B denotes $\mathcal{E}_0(\delta/4)$. By Lemma 5, $\Pr(A) = 1-\delta/4$ and $\Pr(B) = 1-\delta/4$

$$\begin{aligned}
\Pr(A \cap B) &= \Pr(A) + \Pr(B) - \Pr(A \cup B) && (3.32)\\
&\geq \Pr(A) + \Pr(B) - 1 && (3.33)\\
&= 1-\delta/4+1-\delta/4-1 && (3.34)\\
&= 1-\delta/2 && (3.35)
\end{aligned}$$

For the failure event F_n^P, by Hoeffding's inequality in Lemma 3 and Union Bound, we have:

$$\begin{aligned}
\Pr(\cup_{n=1}^K F_n^P) &\leq \sum_n^K \sum_{z\in\mathcal{Z}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_h^H \sum_{s'\in\mathcal{S}} \exp(-n(\beta(n))^2) && (3.36)\\
&= \sum_n^K \sum_{z\in\mathcal{Z}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_h^H \sum_{s'\in\mathcal{S}} \exp\left(-n \cdot \sqrt{\frac{1}{\max(n,1)\log(2|\mathcal{Z}||S|^2|A|HK/\delta)}}^2\right) && (3.37)\\
&= K|\mathcal{Z}||\mathcal{S}|^2|\mathcal{A}|H\frac{\delta}{2|\mathcal{Z}||S|^2|A|HK} && (3.38)\\
&= \delta/2 && (3.39)
\end{aligned}$$

The event $(\cup_{n=1}^{K} F_n^P)^C$ occurs with probability at least $1 - \delta/2$. Combining the results we have $\Pr(\mathcal{E}^{\text{gen}}) = \Pr((\cup_{n=1}^{K} F_n^P)^C \cap \mathcal{E}_{\Omega}(\delta/4) \cap \mathcal{E}_0(\delta/4)) \leq 1 - \delta$, which implies $\mathcal{E}$ occurs with probability at least $1 - \delta$.

3.5.1 Proof for Theorem 1.1

Now, we are ready to present the proof for Theorem 1.1. Without loss of generality, let $\{i, j\}$ denote any pair of subgroups in $\mathcal{Z}^2$. $\mathbf{\Pi}^k$ consists of either the singleton set $\{\pi^0\}$ or the selected policies $\mathbf{\Pi}_{\mathrm{F}}^k$ defined in Equation (3.15). For π^0, we have $|J_i^{\pi^0}(r_m, P_i) - J_j^{\pi^0}(r_m, P_j)| \leq \epsilon, \forall m \in [M]$ by definition of initial fair policy (Assumption 1.1). We will now show that our construction of $\mathbf{\Pi}_F^k$ also satisfies the zero constraint violation property for any such pair of subgroups. For $\pi \in \mathbf{\Pi}_{\mathrm{F}}^k$, to show $|J_i^{\pi}(r_m, P_i) - J_j^{\pi}(r_m, P_j)| \leq \epsilon, \forall m \in [M]$ holds under the good event, we will first show $J_i^{\pi}(r_m, P_i) - J_j^{\pi}(r_m, P_j) \leq \epsilon, \forall m \in [M]$, i.e. the return of group i is no more than the return of group j by ϵ for all tasks m in part 1, and then show $J_j^{\pi}(r_m, P_j) - J_i^{\pi}(r_m, P_i) \leq \epsilon, \forall m \in [M]$, i.e. the return of group j is no more than the return of group i by ϵ for all tasks m in part 2.

Part 1: In the first part of the proof, we will show that on the good event $\mathcal{E}$, for any $k \in [K]$ and policy $\pi \in \mathbf{\Pi}_F^k$,

$$J_i^{\pi}(r_m, P_i) - J_j^{\pi}(r_m, P_j) \leq \epsilon, \forall m \in [M]. \tag{3.40}$$

Proof. Using Lemma 1, we have:

$$J_i^{\pi}(r_m, P_i) \leq J_i^{\pi}(\bar{r}_m^k, \hat{P}_i^k), \forall m \in [M]. \tag{3.41}$$

Similarly, using Lemma 2, we get

$$-J_j^{\pi}(r_m, P_j) \leq -J_j^{\pi}(\underline{r}_m^k, \hat{P}_j^k), \forall m \in [M]. \tag{3.42}$$

Combining Equation (3.41) and Equation (3.42), we have:

$$J_i^{\pi}(r_m, P_i) - J_j^{\pi}(r_m, P_j) \leq J_i^{\pi}(\bar{r}_m^k, \hat{P}_i^k) - J_j^{\pi}(\underline{r}_m^k, \hat{P}_j^k), \forall m \in [M]. \tag{3.43}$$

Note that from the definition of Π_F^k in Equation (3.15), we know any policy in $\pi \in \Pi_F^k$ satisfies the constraint:

$$J_i^\pi(\bar{r}_m^k, \hat{P}_i^k) - J_j^\pi(\underline{r}_m^k, \hat{P}_j^k) \leq \epsilon, \forall m \in [M]. \tag{3.44}$$

Therefore, we have the following relation:

$$J_i^\pi(r_m, P_i) - J_j^\pi(r_m, P_j) \leq J_i^\pi(\bar{r}_m^k, \hat{P}_i^k) - J_j^\pi(\underline{r}_m^k, \hat{P}_j^k) \leq \epsilon, \forall m \in [M]. \tag{3.45}$$

Part 2: In the first part of the proof, we will show that on the good event $\mathcal{E}$, for any $k \in [K]$ and policy $\pi \in \mathbf{\Pi}_F^k$,

$$J_j^\pi(r_m, P_j) - J_i^\pi(r_m, P_i) \leq \epsilon, \forall m \in [M]. \tag{3.46}$$

Proof. Using Lemma 1, we have:

$$J_j^\pi(r_m, P_j) \leq J_j^\pi(\bar{r}_m^k, \hat{P}_j^k), \forall m \in [M]. \tag{3.47}$$

Similarly, using Lemma 2, we get

$$-J_i^\pi(r_m, P_i) \leq -J_i^\pi(\underline{r}_m^k, \hat{P}_i^k), \forall m \in [M]. \tag{3.48}$$

Combining Equation (3.47) and Equation (3.48), we have:

$$J_j^\pi(r_m, P_j) - J_i^\pi(r_m, P_i) \leq J_j^\pi(\bar{r}_m^k, \hat{P}_j^k) - J_i^\pi(\underline{r}_m^k, \hat{P}_i^k), \forall m \in [M]. \tag{3.49}$$

Note that from the definition of Π_F^k in Equation (3.15), we know any policy in $\pi \in \Pi_F^k$ satisfies the constraint:

$$J_j^\pi(\bar{r}_m^k, \hat{P}_j^k) - J_i^\pi(\underline{r}_m^k, \hat{P}_i^k) \leq \epsilon, \forall m \in [M]. \tag{3.50}$$

Therefore, we have the following relation:

$$J_j^\pi(r_m, P_j) - J_i^\pi(r_m, P_i) \leq J_j^\pi(\bar{r}_m^k, \hat{P}_j^k) - J_i^\pi(\underline{r}_m^k, \hat{P}_i^k) \leq \epsilon, \forall m \in [M]. \tag{3.51}$$

Combining the results of $\mathbf{\Pi}^k$ being the singleton set $\{\pi^0\}$ or $\mathbf{\Pi}_{\mathrm{F}}^k$, we have for $\pi \in \mathbf{\Pi}^k$,

$$|J_i^\pi(r_m, P_i) - J_j^\pi(r_m, P_j)| \leq \epsilon, \forall m \in [M], \forall k \in [K], \tag{3.52}$$

which holds for any pair of group $\{i, j\} \in \mathcal{Z}^2$. Extending to all pairs of groups:

$$|J_i^{\pi}(r_m, P_i) - J_j^{\pi}(r_m, P_j)| \leq \epsilon, \forall m \in [M], \forall k \in [K], \forall i \geq j; (i, j) \in \mathcal{Z}^2. \tag{3.53}$$

3.5.2 Proof for Theorem 1.2

From the definition of the conservative set of policies in Equation (3.16), we will apply π^0 when there exist one pair of groups $(i, j) \in \mathcal{Z}^2$ and one task $m \in [M]$ such that the return difference under a optimistic MDP and a pessimistic MDP is greater than or equal to $\frac{\epsilon+\epsilon^0}{2}$. In this case, $|\Pi^k| = |\{\pi^0\}| = 1$. By the Assumption 1.1, we have $\epsilon^0 < \epsilon$ and therefore $\frac{\epsilon+\epsilon^0}{2} < \epsilon$. When the return difference of applying π^0 for all pair of groups $(i, j) \in \mathcal{Z}^2$ and for all task $m \in [M]$ is less than or equal to $\frac{\epsilon+\epsilon^0}{2}$, which is strictly less than ϵ, then there exist infinitely many policies that are close to π^0 that can result in a return difference less than ϵ and thus satisfy the constraint in Equation (3.15). In this case, $|\Pi^k| = |\Pi_F^k| > 1$.

We can follow Liu et al. (2021) and break down the regret according to the above two cases: $|\Pi^k| = 1$ and $|\Pi^k| > 1$. For all task m, the regret can be broken down in into three terms. Providing upper bounds for each of the three terms by Lemma A.1, Lemma A.2 and Lemma A.3 will conclude our regret analysis.

$$\begin{aligned}
\mathrm{Reg}(K; r_m) = &\sum_{k=1}^{K} 1(|\Pi^k| = 1)(J^{\pi^*}(r_m, P) - J^{\pi^0}(r_m, P)) && \text{(I)}\\
&+ \sum_{k=1}^{K} 1(|\Pi^k| > 1)(J^{\pi^*}(r_m, P) - J^{\pi^k}(\ddot{r}_m, \hat{P}^k)) && \text{(II)}\\
&+ \sum_{k=1}^{K} 1(|\Pi^k| > 1)(J^{\pi^k}(\ddot{r}_m^k, \hat{P}^k) - J^{\pi^k}(r_m, P)) && \text{(III)}
\end{aligned}$$

Lemma A.1(Similar to lemma C.6 in Satija et al. (2023)) *On good event $\mathcal{E}$,*

$$\sum_{k}^{K} 1(|\Pi^k| = 1) \leq \tilde{O}\left(\frac{|\mathcal{Z}|^2 M^4 H^4 |\mathcal{S}|^3 |\mathcal{A}|}{(\epsilon - \epsilon^0) \min\{1, (\epsilon - \epsilon^0)}\right). \tag{3.54}$$

Proof. For part I, we want to obtain an upper bound for $\sum_k^K (|\Pi^k| = 1)$. We start by giving an upper bound for $\sum_k^K 1(|\Pi^k| = 1; (i, j), m)$, which denotes when two particular groups i, j led to the fairness violation in task m. In this case, when the fairness constraint is violated with respect to π^0, either group i's return is much larger than group j's return as in the following Case $A_{(i,j),m}$, or group j's return is much larger than group i's return as in Case $B_{(i,j),m}$.

Case $A_{(i,j),m}$

$$J_i^{\pi^k}(r_m, P_i) - J_j^{\pi^k}(r_m, P_j) \geq (\epsilon + \epsilon^0)/2 \tag{3.55}$$

Case $B_{(i,j),m}$

$$J_j^{\pi^k}(r_m, P_j) - J_i^{\pi^k}(r_m, P_i) \geq (\epsilon + \epsilon^0)/2 \tag{3.56}$$

$$\tag{3.57}$$

We define $K' = \sum_k^K 1(|\Pi^k| = 1; (i,j), m)$.

$$\left(\frac{\varepsilon - \varepsilon^0}{2}\right) K' = \sum_{k=1}^{K} 1(|\Pi^k| = 1; (i,j), m) \left(\frac{\varepsilon - \varepsilon^0}{2}\right) \tag{3.58}$$

$$= \sum_{k=1}^{K} 1(|\Pi^k| = 1; (i,j), m) \left(\frac{\varepsilon + \varepsilon^0}{2} - \varepsilon^0\right) \tag{3.59}$$

$$\leq \sum_{k=1}^{K} 1(|\Pi^k| = 1; A_{(i,j),m}) \left(\frac{\varepsilon + \varepsilon^0}{2} - \varepsilon^0\right) + \sum_{k=1}^{K} 1(|\Pi^k| = 1; B_{(i,j),m}) \left(\frac{\varepsilon + \varepsilon^0}{2} - \varepsilon^0\right) \tag{3.60}$$

For Case $A_{(i,j),m}$:

$$\sum_{k=1}^{K} 1(|\Pi^k| = 1; A_{(i,j),m}) \left(\frac{(\varepsilon + \varepsilon^0)}{2} - \varepsilon^0\right) \tag{3.61}$$

$$\leq 1(|\Pi^k| = 1; A_{(i,j),m}) \left((J^{\pi_i^k}(r^k, \hat{P}^k) - J^{\pi_j^k}(r^k, \hat{P}^k)) - (J^{\pi_i^k}(r, P) - J^{\pi_j^k}(r, P)) \right) \tag{3.62}$$

$$= \underbrace{1(|\Pi^k| = 1; A_{(i,j),m})(J^{\pi_i^k}(r^k, \hat{P}^k) - J^{\pi_i^k}(r, P))}_{(A.1)} + \underbrace{1(|\Pi^k| = 1; A_{(i,j),m})(J^{\pi_j^k}(r, P) - J^{\pi_j^k}(r^k, \hat{P}^k))}_{(A.2)}, \tag{3.63}$$

For the first term, we use Lemma 5 with the designed optimistic reward function from Equation (3.13)

$$\begin{aligned} |\bar{r}^k_{m,h} - r_{m,h}| &= |\alpha \beta^k_h| && (3.64)\\ &\leq (|\mathcal{S}|H)\beta^k_h, && (3.65) \end{aligned}$$

Plugging in $\alpha = |\mathcal{S}H|$ in Lemma 5, the first term A.1 is bounded by

$$A.1 = \tilde{\mathcal{O}}(H^4|\mathcal{S}|^3|\mathcal{A}| + H^2\sqrt{|\mathcal{S}|^3|\mathcal{A}|K'}) \tag{3.66}$$

For the second term A.2, we use the following relation from the designed pessimistic reward function from Equation (3.14)

$$\begin{aligned} |r_h - \underline{r}^k_h| &= |-(-\alpha\beta^k_h)| && (3.67)\\ &\leq (|\mathcal{S}|H)\beta^k_h && (3.68) \end{aligned}$$

Applying Lemma 5,

$$A.2 = \tilde{\mathcal{O}}(H^4|\mathcal{S}|^3|\mathcal{A}| + H^2\sqrt{|\mathcal{S}|^3|\mathcal{A}|K'}) \tag{3.69}$$

Therefore,

$$\sum_{k=1}^{K} 1(|\Pi^k| = 1; A_{(i,j),m}) \left(\frac{(\varepsilon + \varepsilon^0)}{2} - \varepsilon^0\right) = \tilde{\mathcal{O}}(H^4|\mathcal{S}|^3|\mathcal{A}| + H^2\sqrt{|\mathcal{S}|^3|\mathcal{A}|K'}) \tag{3.70}$$

Since case A and case B are symmetric with respect to the two groups i and j, we can follow the above steps and obtain the same big O notation for case B.

$$\sum_{k=1}^{K} 1(|\Pi^k| = 1; B_{(i,j),m}) \left(\frac{(\varepsilon + \varepsilon^0)}{2} - \varepsilon^0\right) = \tilde{\mathcal{O}}(H^4|\mathcal{S}|^3|\mathcal{A}| + H^2\sqrt{|\mathcal{S}|^3|\mathcal{A}|K'}) \tag{3.71}$$

Combining results for Case A and Case B, we will have the same big O notation for K'.

$$\frac{(\epsilon + \epsilon^0)}{2}K' = \tilde{\mathcal{O}}(H^4|\mathcal{S}|^3|\mathcal{A}| + H^2\sqrt{|\mathcal{S}|^3|\mathcal{A}|K'}) \tag{3.72}$$

By Lemma 7 (Lemma D.6 in Liu et al. (2021)),

$$K' = \sum_{k}^{K} 1(|\Pi^k| = 1; (i,j), m) \leq \tilde{\mathcal{O}}\left(\frac{H^4|\mathcal{S}|^3|\mathcal{A}|}{(\epsilon-\epsilon^0)\min\{1, (\epsilon-\epsilon^0)}\right) \tag{3.73}$$

Now, to obtain the upper bound for fairness violation by any possible pairs of groups and for all tasks $\sum_k^K 1(|\Pi^k| = 1)$, by the union bound we have

$$\sum_{k}^{K} 1(|\Pi^k| = 1) \leq \sum_{i,j \in \mathcal{Z}^2} \sum_{m}^{M} \sum_{k}^{K} 1(|\Pi^k| = 1; (i,j), m) \tag{3.74}$$

$$\leq |\mathcal{Z}|^2 MK' \tag{3.75}$$

$$\leq \tilde{\mathcal{O}}\left(\frac{|\mathcal{Z}|^2 MH^4|\mathcal{S}|^3|\mathcal{A}|}{(\epsilon-\epsilon^0)\min\{1, (\epsilon-\epsilon^0)}\right). \tag{3.76}$$

Lemma A.2 *For $\alpha_l = |S|H + \frac{8M^2|S|H^2}{\epsilon-\epsilon^0}$, on good event $\mathcal{E}$,*

$$\sum_{k=1}^{K} 1(|\Pi^k| > 1)(J^{\pi^*}(r_m, P) - J^{\pi^k}(\bar{r}_m, \hat{P}^k)) \leq 0 \tag{3.77}$$

Proof. When $\pi^* \in \Pi^k$, the inequality holds because of the reward bonus and the fact that π^k maximizes the optimistic CMDP from **??**.

When $\pi^* \notin \Pi^k$, we first show the difference in cost is less or equal to 0 for any pair of groups i, j, then it holds for all groups.

Let B_{γ_k} denote an independent Bernoulli distributed random variable with mean γ_k. We can define a probability mixed policy $\tilde{\pi}^k$ as:

$$\tilde{\pi} = B_{\gamma_k}\pi^* + (1 - B_{\gamma_k})\pi^0 \tag{3.78}$$

Let $\gamma_k \in [0,1]$ be the largest coefficient that ensures the constraint is not violated by the mixed policy $\tilde{\pi}^k$,

$$J^{\tilde{\pi}_i}(\bar{r}_m^k, \hat{P}_i^k) - J^{\tilde{\pi}_j}(\bar{r}_m^k, \hat{P}_j^k) \leq \epsilon \tag{3.79}$$

If $J_i^{\pi^*}(\bar{r}_m^k, \hat{P}_i^k) - J_j^{\pi^*}(\underline{r}_m, \hat{P}_j^k) < \epsilon$, then $\gamma_k = 1$. Else, we will obtain a γ_k that make the equality hold for (3.79).

Denote the pessimistic cost of the difference in value between the two groups as:

$$\tilde{J}_{i,j}^{\pi} := J_i^{\pi}(\bar{r}_m^k, \hat{P}_i^k) - J_j^{\pi}(\underline{\mathrm{r}}_m^k, \hat{P}_j^k), \tag{3.80}$$

where π could be π^* or π^0, and denote the difference in value in the true MDP as

$$J_{i,j}^{\pi} := J_i^{\pi}(r_m, P_i) - J_j^{\pi}(r_m, P_j) \tag{3.81}$$

When the equality holds, we have

$$\begin{aligned}
\epsilon &= \gamma_k \tilde{J}_{i,j}^{\pi^*} + (1-\gamma_k)\tilde{J}_{i,j}^{\pi^0} \\
&\leq \gamma_k \tilde{J}_{i,j}^{\pi^*} + (1-\gamma_k)\frac{\epsilon+\epsilon^0}{2} \\
&= \gamma_k(\tilde{J}_{i,j}^{\pi^*} - J_{i,j}^{\pi^*}) + \gamma_k J_{i,j}^{\pi^*} + (1-\gamma_k)\frac{\epsilon+\epsilon^0}{2} \\
&\leq \gamma_k(\tilde{J}_{i,j}^{\pi^*} - J_{i,j}^{\pi^*}) + \gamma_k \epsilon + \frac{\epsilon+\epsilon^0}{2} - \gamma_k \frac{\epsilon+\epsilon^0}{2} \\
&\leq \gamma_k(\tilde{J}_{i,j}^{\pi^*} - J_{i,j}^{\pi^*} + \frac{\epsilon-\epsilon^0}{2}) + \frac{\epsilon+\epsilon^0}{2}
\end{aligned}$$

Using Lemma 1 and Lemma 2, we have

$$J^{\pi_i}(r_m, P_i) \leq J^{\pi_i}(\bar{r}_m^k, \hat{P}_i^k), \tag{3.82}$$

and

$$-J^{\pi_j^*}(r_m^k, \hat{P}_j^k) \leq -J^{\pi_j^*}(\underline{\mathrm{r}}_m, P_j). \tag{3.83}$$

Adding (3.82) and (3.83),

$$\tilde{J}_{i,j}^{\pi^*} - J_{i,j}^{\pi^*} \geq 0. \tag{3.84}$$

Since $\epsilon > \epsilon^0$, $\tilde{J}_{i,j}^{\pi^*} - J_{i,j}^{\pi^*} + \frac{\epsilon-\epsilon^0}{2} \geq 0$. Therefore,

$$\gamma_k \geq \frac{\epsilon-\epsilon^0}{\epsilon-\epsilon^0+2(\tilde{J}_{i,j}^{\pi^*} - J_{i,j}^{\pi^*})} \tag{3.85}$$

Using Lemma 3 and Lemma 4 , we have

$$J_i^{\pi}(\bar{r}_m, \hat{P}_i^k) - J_i^{\pi}(r_m, P_i) \leq 2(|\mathcal{S}|H) J_i^{\pi}(\beta_h^k(s,a), \hat{P}_i^k), \tag{3.86}$$

and

$$J_j^{\pi}(r_m, P_j) - J_j^{\pi}(\underline{\mathrm{r}}_m, \hat{P}_j^k) \leq 2(|\mathcal{S}|H) J_j^{\pi}(\beta_h^k(s,a), \hat{P}_j^k). \tag{3.87}$$

Adding (3.86) and (3.87),

$$\tilde{J}_{i,j}^{\pi} - J_{i,j}^{\pi} \leq 2(|\mathcal{S}|H) \left(J_i^{\pi}(\beta_h^k(s,a), \hat{P}^k) + J_j^{\pi}(\beta_h^k(s,a), \hat{P}_j^k) \right). \tag{3.88}$$

Because π^k is the optimal policy in the optimistic CMDP, we have:

$$\begin{aligned}
J_i^{\pi^k}(\ddot{r}_m, \hat{P}_i^k) + J_j^{\pi^k}(\ddot{r}_m, \hat{P}_j^k) &\geq J_i^{\hat{\pi}^k}(\ddot{r}_m, \hat{P}_i^k) + J_j^{\hat{\pi}^k}(\ddot{r}_m, \hat{P}_j^k) && (3.89)\\
&= J_i^{\tilde{\pi}^k}(\ddot{r}_m, \hat{P}_i^k) + J_j^{\tilde{\pi}^k}(\ddot{r}_m, \hat{P}_j^k) && (3.90)\\
&= \gamma_k(J_i^{\pi^{*k}}(\ddot{r}_m, \hat{P}_i^k) + J_j^{\pi^{*k}}(\ddot{r}_m, \hat{P}_j^k)) + \underbrace{(1-\gamma_k)(J_i^{\pi^{0^k}}(\ddot{r}_m, \hat{P}_i^k) + J_j^{\pi^{0^k}}(\ddot{r}_m, \hat{P}_j^k))}_{\geq 0} && (3.91)\\
&\geq \gamma_k(J_i^{\pi^{*k}}(\ddot{r}_m, \hat{P}_i^k) + J_j^{\pi^{*k}}(\ddot{r}_m, \hat{P}_j^k)) && (3.92)\\
&\geq \frac{\epsilon - \epsilon^0}{\epsilon - \epsilon^0 + 2(\tilde{J}_{i,j}^{\pi^*} - J_{i,j}^{\pi^*})} (J_i^{\pi^{*k}}(\ddot{r}_m, \hat{P}_i^k) + J_j^{\pi^{*k}}(\ddot{r}_m, \hat{P}_j^k)) && (3.93)\\
&\geq \frac{\epsilon - \epsilon^0}{\epsilon - \epsilon^0 + 4|\mathcal{S}|H\left(J_i^{\pi}(\beta_h^k(s,a), \hat{P}_i^k) + J_j^{\pi}(\beta_h^k(s,a), \hat{P}_j^k)\right)} (J_i^{\pi^{*k}}(\ddot{r}_m, \hat{P}_i^k) + J_j^{\pi^{*k}}(\ddot{r}_m, \hat{P}_j^k)) && (3.94)
\end{aligned}$$

To make $J_i^{\pi^k}(\ddot{r}_m, \hat{P}_i^k) + J_j^{\pi^k}(\ddot{r}_m, \hat{P}_j^k) \leq J_i^{\pi^*}(r_m, P_i) + J_j^{\pi^*}(r_m, P_j)$, it is sufficient to show

$$\frac{\epsilon - \epsilon^0}{\epsilon - \epsilon^0 + 4|\mathcal{S}|H\left(J_i^{\pi}(\beta_h^k(s,a), \hat{P}_i^k) + J_j^{\pi}(\beta_h^k(s,a), \hat{P}_j^k)\right)} \left(J_i^{\pi^k}(\ddot{r}_m, \hat{P}_i^k) + J_j^{\pi^k}(\ddot{r}_m, \hat{P}_j^k)\right) \geq J_i^{\pi^*}(r_m, P_i) + J_j^{\pi^*}(r_m, P_j), \tag{3.95}$$

which is equivalent to

$$\begin{aligned}
(\epsilon - \epsilon^0)\left(\left(J_i^{\pi}(\ddot{r}_{m,h}(s,a), \hat{P}_i^k) + J_j^{\pi}(\ddot{r}_{m,h}(s,a), \hat{P}_j^k)\right) - \left(J_i^{\pi^*}(r_m, P_i) + J_j^{\pi^*}(r_m, P_j)\right)\right) && (3.96)\\
\geq 4|\mathcal{S}|H\left(J_i^{\pi}(\beta_h^k(s,a), \hat{P}_i^k) + J_j^{\pi}(\beta_h^k(s,a), \hat{P}_j^k)\right)\left(J_i^{\pi^*}(r_m, P_i) + J_j^{\pi^*}(r_m, P_j)\right) && (3.97)
\end{aligned}$$

From the value difference lemma (Lemma 6), for any group $z \in \mathcal{Z}$,

$$J_z^{\pi^*}(\ddot{r}_m, \hat{P}_z^k) - J_z^{\pi^*}(r_m, P_z) \tag{3.98}$$

$$= E\left[\sum_{h=1}^{H}\left(\ddot{r}_m(s_h, a_h) - r_m(s_h, a_h) + \sum_{s'}(\hat{P}^k_{z,h} - P_{z,h})(s'|s_h, a_h)V^{\pi^*_z}_{h+1}(s'; \sum_m r_m, P_{z,h})\right)\Big|\mathcal{F}_{k-1}\right] \tag{3.99}$$

$$\geq E\left[\sum_{h=1}^{H}(\alpha_l - |\mathcal{S}|H)\beta_h^k(s_h, a_h)\Big|\mathcal{F}_{k-1}\right] \tag{3.100}$$

$$= (\alpha_l - |\mathcal{S}|H)J_z^{\pi^*}(\beta^k, \hat{P}_z^k). \tag{3.101}$$

Using the above result for group i and j seperately, we have

$$J_i^{\pi^*}(\ddot{r}_m, \hat{P}_i^k) - J_i^{\pi^*}(r_m, P_i) \geq (\alpha_l - |\mathcal{S}|H)J_i^{\pi^*}(\beta^k, \hat{P}_i^k). \tag{3.102}$$

$$J_j^{\pi^*}(\ddot{r}_m, \hat{P}_j^k) - J_j^{\pi^*}(r_m, P_j) \geq (\alpha_l - |\mathcal{S}|H)J_j^{\pi^*}(\beta^k, \hat{P}_j^k). \tag{3.103}$$

Adding the above two inequalities,

$$\left(J_i^{\pi}(\ddot{r}_{m,h}(s,a), \hat{P}_i^k) + J_j^{\pi}(\ddot{r}_{m,h}(s,a), \hat{P}_j^k)\right) - \left(J_i^{\pi^*}(r_m, P_i) + J_j^{\pi^*}(r_m, P_j)\right) \geq (\alpha_l - |\mathcal{S}|H)(J_i^{\pi^*}(\beta^k, \hat{P}_i^k) + J_j^{\pi^*}(\beta^k, \hat{P}_j^k)) \tag{3.104}$$

Letting $\alpha_l = |\mathcal{S}|H + \frac{4|\mathcal{S}|H}{\epsilon - \epsilon^0}2H$,

$$\left(J_i^{\pi}(\ddot{r}_{m,h}(s,a), \hat{P}_i^k) + J_j^{\pi}(\ddot{r}_{m,h}(s,a), \hat{P}_j^k)\right) - \left(J_i^{\pi^*}(r_m, P_i) + J_j^{\pi^*}(r_m, P_j)\right) \geq \frac{4|\mathcal{S}|H}{\epsilon - \epsilon^0}(J_i^{\pi^*}(\beta^k, \hat{P}_i^k) + J_j^{\pi^*}(\beta^k, \hat{P}_j^k))2H. \tag{3.105}$$

Since $J_i^{\pi^*}(r_m, P_i) + J_j^{\pi^*}(r_m, P_j) \leq 2H$, the inequality (3.96) is satisfied. Now we've shown the difference in cost is less or equal to 0 for any pair of groups i, j, which is $J_i^{\pi^k}(\ddot{r}_m, \hat{P}_i^k) + J_j^{\pi^k}(\ddot{r}_m, \hat{P}_j^k) \leq J_i^{\pi^*}(r_m, P_i) + J_j^{\pi^*}(r_m, P_j)$

Using the above result for consecutive pairs of subgroups $\{(1,2), (2,3), \ldots, (|\mathcal{Z}| - 1, |\mathcal{Z}|), (|\mathcal{Z}|, 1)\}$, and adding them together we get

$$2\sum_{z=1}^{|\mathcal{Z}|} J_z^{\pi^k}(\ddot{r}_m^k, \hat{P}^k) \geq 2\sum_{z=1}^{|\mathcal{Z}|} J_z^{\pi^*}(r_m, P), \tag{3.106}$$

which is

$$\sum_{z\in\mathcal{Z}} \left(J_z^{\pi^*}(r_m, P) - J_z^{\pi^k}(\ddot{r}_m^k, \hat{P}^k)\right) \leq 0 \tag{3.107}$$

In our setting, we iterate through every group z from $\mathcal{Z}$, therefore we have:

$$\sum_{k=1}^{K} 1(|\Pi^k| > 1)\left(J^{\pi^*}(r_m, P) - J^{\pi^k}(\ddot{r}^k, \hat{P}^k)\right) = \sum_{k=1}^{K} 1(|\Pi^k| > 1) \sum_{z\in\mathcal{Z}} \left(J_z^{\pi^*}(r_m, P_z) - J_z^{\pi^k}(\ddot{r}_m^k, \hat{P}_z^k)\right) \tag{3.108}$$

$$= \sum_{k=1}^{K} 1(|\Pi^k| > 1) \sum_{z\in\mathcal{Z}} \left(J_z^{\pi^*}(r_m, P_z) - J_z^{\pi^k}(\ddot{r}_m^k, \hat{P}_z^k)\right) \tag{3.109}$$

$$\leq 0. \tag{3.110}$$

Lemma A.3 *On good event $\mathcal{E}$,*

$$\sum_{k=1}^{K} 1(|\Pi^k| > 1)(J(\pi^k, \mu, \sum_m \ddot{r}_m^k, \hat{P}^k) - J(\pi^k, \mu, \sum_m r_m, P)) = \tilde{\mathcal{O}}\left(\frac{|\mathcal{Z}|H^3}{(\epsilon-\epsilon^0)}\sqrt{|\mathcal{S}|^3|\mathcal{A}|K} + \frac{|\mathcal{Z}|H^5|\mathcal{S}|^3|\mathcal{A}|}{(\epsilon-\epsilon^0)}\right) \tag{3.111}$$

Proof. Since we build the optimistic reward with bonus, we have $|\ddot{r}_{m,h} - r_{m,h}| \leq \alpha_l \beta_h^k$. By applying Lemma B.1,

$$\sum_{k=1}^{K} 1(|\Pi^k| > 1)(J^{\pi^k}(\ddot{r}_m, \hat{P}^k) - J^{\pi^k}(r_m, P)) \tag{3.112}$$

$$\leq \sum_{k=1}^{K} \sum_{z\in\mathcal{Z}} J^{\pi^k}(\ddot{r}_m^k, \hat{P}_z^k) - J^{\pi^k}(r_m, P_z) \tag{3.113}$$

$$= \tilde{\mathcal{O}}\left(|\mathcal{Z}|(\alpha_l + \sqrt{2|\mathcal{S}|H})(H\sqrt{|\mathcal{S}||\mathcal{A}|K}) + \alpha_l|\mathcal{Z}|H^3|\mathcal{S}|^2|\mathcal{A}|\right) \tag{3.114}$$

$$= \tilde{\mathcal{O}}\left(\frac{|\mathcal{Z}|H^3}{(\epsilon-\epsilon^0)}\sqrt{|\mathcal{S}|^3|\mathcal{A}|K} + \frac{|\mathcal{Z}|H^5|\mathcal{S}|^3|\mathcal{A}|}{(\epsilon-\epsilon^0)}\right) \tag{3.115}$$

Combining the results for term (I), term(II) and term(III), we have

$$\mathrm{Reg}(K; r_m) = \sum_k [J(\pi_i^*, \mu, P, r_m) - J(\pi^k, \mu, P, r_m)] = \tilde{\mathcal{O}}\left(\frac{|\mathcal{Z}|H^3}{(\epsilon - \epsilon^0)} \sqrt{|\mathcal{S}|^3|\mathcal{A}|K} + \frac{|\mathcal{Z}|^2 M H^5 |\mathcal{S}|^3 |\mathcal{A}|}{\min\{(\epsilon - \epsilon^0), (\epsilon - \epsilon^0)^2\}} \right) \tag{3.116}$$

Chapter 4

Group Fairness in Reward Models for Fine-tuning LLMs with RLHF

4.1 Group Fairness in Reward Models

To define group fairness in reward models, we first present the definitions for social groups and protected groups.

Definition 1 (SOCIAL GROUP). A social group $G \subseteq \mathbb{G}$ is the population that shares an identity trait, which may be fixed, contextual, or socially constructed. Examples include demographic attributes collected through the census, including age, gender, and occupation.

Definition 2 (PROTECTED ATTRIBUTE). A protected attribute is the shared identity trait that determines the group identity of a social group.

In traditional group fairness in machine learning classification, M could be accuracy, true positive rate, or false positive rate. Instead of making fair high-stakes decisions, quality of generation matters in the case of LLM. Helpfulness, correctness, and coherence need to be the outcome a good fairness algorithm aims to equalize over. Therefore, a reward model is a natural candidate for estimating such outcomes. We define group fairness, or demographic parity, in the following:

Definition 3 (Group Fairness of Reward Models). Consider a model $\mathcal{M}$ that evaluates the quality of generated outputs from an LLM. Assume we have access to a set of prompts X_G, where the ground-truth quality of each prompt $x \sim X_G$ is equal. Let $\mathbb{E}_{x\sim X_G}[\mathcal{M}(x;\theta)]$ be the outcome measured by the reward model given a distribution of prompts X_G specific to group $G \in \mathcal{G}$, where $\mathcal{G}$ represents a set of social groups, and each group G has a different distribution of prompts X_G. Group fairness requires (approximate) parity in the average reward scores across all groups $G \in \mathcal{G}$, up to ϵ, as measured by the reward model $\mathcal{M}$:

$$\left|\mathbb{E}_{x\sim X_G}[\mathcal{M}(x;\theta)] - \mathbb{E}_{x\sim X_{G'}}[\mathcal{M}(x;\theta)]\right| \leq \epsilon.$$

4.2 Benchmarking Reward Models

4.2.1 Constructing the Evaluation Dataset from The arXiv Metadata

The arXiv Metadata dataset, which use is under the Creative Commons CC0 1.0 Universal (Public Domain Dedication) license, offers significant advantages to our fairness study. The dataset primarily consists of titles and abstracts from expert-written papers. The expert authorship ensures that the abstracts are high in quality, therefore receiving full scores on attributes such as correctness and coherence should be a minimum requirement. The reward model that satisfies group fairness should consistently deliver equal average reward scores for prompts and responses across all social groups.

Selecting Social Groups arXiv naturally has human experts written papers from different domains such as Physics, economics, and computer science. Identifying social groups by occupation, such as physicists, economists, and computer scientists, a total of 8 demographic groups exist in the 8 categories from arXiv: physics, mathematics, computer science, economics, electrical engineering and system science quantitative-biology, quantitative finance.

Evaluation Prompts and Responses We use expert-written texts from arXiv Metadata to benchmark the group fairness in reward models. We utilize the title and abstract of a paper from the arXiv Metadata to construct the prompts and responses required for evaluating the reward model. Specifically, the title of the paper is used to construct the prompt with the question "Write an abstract for a paper with title <Title of the Paper>", and we assume the expert-written abstract of the paper is the ground-truth response to the prompt. A fair reward model would achieve the same average score for prompts and responses for all 8 different categories.

The original arXiv Metadata dataset contains 200 thousand papers, of which less than 400 papers are from the economics category. To obtain larger sample size for each social group, we use arXiv API to fetch metadata for each arXiv category. In constructing the dataset, we only include paper listed as one category only. Papers under multiple categories are removed for a more distinct comparison of the differences in group means. A total of 2000 titles and abstracts are curated for each category.

4.2.2 Experimental Setup

Simplifying the Distributions of Prompts To simplify the evaluation, we only do inference on prompts and responses that are unique to a specific group, assuming other groups never raise these questions as prompts to LLMs. In addition, we assume the distribution of prompts that all groups share is the same, therefore we are not evaluating on these shared common prompts as they will not affect the difference in group mean.

Models We only include reward models that can compute a reward score based on a single prompt and response message. LLM-as-a-Judge (Zheng et al., 2024) and pairwise reward models are not included, as they require comparing two messages. The following 8 models from the RewardBench (Lambert et al., 2024) are selected in the evaluation: GRM-llama3-8B-sftreg (Yang et al., 2024), ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024b,a), Eurus-RM-7b (Yuan et al., 2024), FsfairX-LLaMA3-RM-v0.1 (Dong et al., 2023; Xiong et al., 2024), Mistral-RM-for-RAFT-GSHF-v0 (Dong et al., 2023; Xiong et al., 2023), RM-Mistral-7B (Dong et al., 2023; Xiong et al., 2024), Nemotron-4-340B-Reward (Wang et al., 2024c), and tulu-v2.5-13b-preference-mix-rm (Ivison et al., 2024).

Recourses for Model Inference For the evaluation of the models, we utilized two

NVIDIA A100 GPUs with 80 GB of memory for the tulu-v2.5-13b-preference-mix-rm model. API calls were employed for the Nemotron-4-340B-Reward model, leveraging external compute resources. For models with fewer than 8 billion parameters, such as GRM-llama3-8B-sftreg and ArmoRM-Llama3-8B-v0.1, we used NVIDIA RTX 6000 GPUs. Each model's evaluation was completed within a maximum compute time of 3 hours.

Group Fairness Metrics

Normalized Maximum Group Difference The reward models are not trained to predict scores on the same scale. Therefore, directly computing the difference in group means is not a fair comparison. With this in mind, we propose a normalized maximum group difference score as a metric for group fairness. For each reward model, we compute the maximum difference in average rewards between any two social groups. This difference is then normalized by dividing it by the mean of the reward scores across all social groups.

ANOVA as a Group Fairness Metric

To rigorously assess group fairness in the performance of reward models, we employ Analysis of Variance (ANOVA) as a statistical method to determine whether there are statistically significant differences between the means of rewards across different demographic groups defined in our study. ANOVA is instrumental in identifying whether variations in reward scores are due to inherent differences among the groups or are a result of random variations. This is critical in our context as it helps ensure that any observed difference in reward outcomes are attributable to the model's unfairness across different groups.

4.2.3 Results Analysis

The plot for the average reward score of the selected 8 top-performing reward models from RewardBench is shown in Figure 4.1. Notice that not all reward models are on the same scale. For example, in the model design of ArmoRM-Llama3-8B-v0.1, a gating layer is applied to the outputs of the regression layer, resulting average rewards for all social

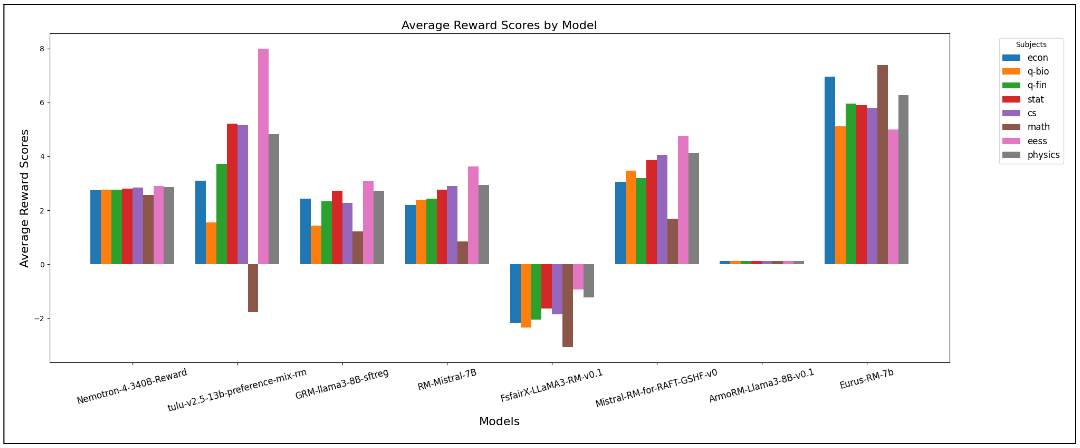

Figure 4.1: Average Reward Scores by Model and Subject across various domains.

groups close to zero.

Through a thorough analysis of the experiment results, we have made the following conclusions:

The group unfairness in all reward models is statistically significant. The F-statistics and p-values from the ANOVA test, detailed in Table 4.1, indicate that all reward models have F-statistics exceeding 70, and p-values significantly lower than 0.0001, confirming substantial differences in group means. Notably, ArmoRM-Llama3-8B-v0.1, the second highest ranked model on the RewardBench leaderboard, has the smallest F-statistic of 70.44. Although the lowest among the models tested, this value is considerably high, indicating significant group differences given that an F-statistic of 1 would mean no group difference. Furthermore, the Nemotron-4-340B-Reward model, which exhibits the second lowest normalized maximum group difference, displays the second highest in the F-statistics. This indicates that it has a low within-group variance, which means there exists a significant group difference in the reward model's output. These findings affirm that the disparities are not due to randomness but reflect a significant systemic bias within the models. Further exploration of these disparities' specific characteristics could guide targeted improvements to enhance group fairness in reward model training. Such

Table 4.1: ANOVA results for various reward models, assessing the significance of group differences in rewards.

Reward Model	F-Statistics	p-Value	RewardBench Rank
ArmoRM-Llama3-8B-v0.1	**70.44**	9.46×10^{-101}	2
GRM-llama3-8B-sftreg	134.63	1.75×10^{-193}	8
Eurus-RM-7b	156.11	5.15×10^{-224}	16
FsfairX-LLaMA3-RM-v0.1	232.98	$< 1 \times 10^{-300}$	12
RM-Mistral-7B	270.06	$< 1 \times 10^{-300}$	22
tulu-v2.5-13b-preference-mix-rm	384.86	$< 1 \times 10^{-300}$	19
Nemotron-4-340B-Reward	427.88	$< 1 \times 10^{-300}$	1
Mistral-RM-for-RAFT-GSHF-v0	518.15	$< 1 \times 10^{-300}$	23

findings confirm that the disparities are not merely by the randomness in the data but reflect a significant systemic bias within the model.

The best performing reward models are the fairer reward models. To compare the group fairness in the reward models, the normalized maximum group difference is computed. The results are shown in percentages in Table 4.2. The top 2 models from RewardBench Leaderboard, namely NemoTron-4-340B-reward and ArmoRM-Llama3-8B-v0.1 exhibit smaller Normalized Maximum Group Differences, substantially outperforming other models evaluated in this study, suggesting that the best reward models also exhibit the better group fairness.

Table 4.2: Differences in average rewards between the maximum and minimum values for each reward model, expressed as percentages. The score with the lowest absolute value is in bold.

Model	Normalized Maximum Group Difference (%)	RewardBench Rank
Nemotron-4-340B-Reward	12.49%	1
tulu-v2.5-13b-preference-mix-rm	262.89%	19
GRM-llama3-8B-sftreg	82.09%	8
RM-Mistral-7B	110.63%	22
FsfairX-LLaMA3-RM-v0.1	-111.52%	12
Mistral-RM-for-RAFT-GSHF-v0	87.46%	23
ArmoRM-Llama3-8B-v0.1	**9.78%**	2
Eurus-RM-7b	39.53%	16

Table 4.3: Multiple Comparison of Means by the Tukey HSD Test

Reward Model	Significant Pairs / Total Pairs
GRM-llama3-8B-sftreg	**23/28**
ArmoRM-Llama3-8B-v0.1	**23/28**
Eurus-RM-7b	24/28
FsfairX-LLaMA3-RM-v0.1	26/28
Mistral-RM-for-RAFT-GSHF-v0	26/28
RM-Mistral-7B	25/28
Nemotron-4-340B-Reward	24/28
tulu-v2.5-13b-preference-mix-rm	25/28

Group unfairness exists in most pairs of demographic groups in every reward model. The Tukey HSD Test, a post-hoc Analysis of ANOVA, shows that each reward model has at least or more than 23 pairs of groups that shows significant differences in the group mean out of a total of all 28 possible combinations of pairs for 8 groups. This indicates that the significant findings from ANOVA are not a result of a significant difference between a only few groups, but rather widespread differences in group means across the majority of group comparisons.

A systematic unfairness might exist in reward models To elucidate the variations in average rewards across different demographic groups, we present a standardized comparison of average rewards by subject in Figure 4.2. This analysis reveals a consistent pattern of disparate treatment for all demographic groups across most reward models. For a better illustration, besides ArmoRM-Llama3-8B-v0.1 and Eurus-RM-7b, the 340B Nemotron model exhibits a Pearson correlation of over 0.8 with all of the rest reward models (in some case 0.99), as shown in Table 4.4. The congruence in average reward score disparities across the majority of models suggests a systemic bias that may originate from similar methodologies in their training datasets and algorithms. These models tend to undervalue outputs related to the "math" subject area while favoring those from "electrical engineering and system science". This necessitates a more nuanced approach to reward model training and evaluation, aiming to enhance its group fairness across people from diverse demographic groups.

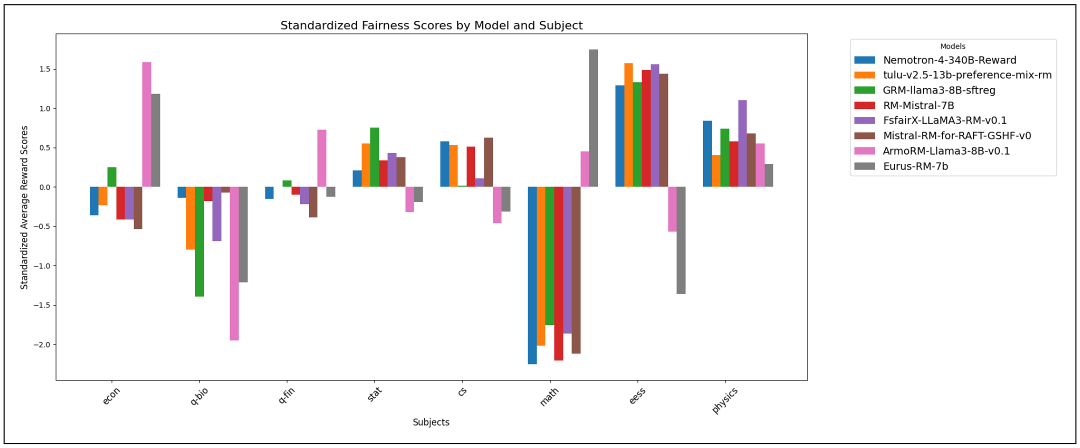

Figure 4.2: Fairness Scores by Model and Subject across various domains.

Table 4.4: Pearson Correlation Coefficients of NVIDIA Nemotron Model with Other Models

Model	Pearson Correlation Coefficient
tulu-v2.5-13b-preference-mix-rm	0.942
RM-Mistral-7B	0.991
Mistral-RM-for-RAFT-GSHF-v0	0.988
FsfairX-LLaMA3-RM-v0.1	0.945
GRM-llama3-8B-sftreg	0.820
Eurus-RM-7b	-0.738
ArmoRM-Llama3-8B-v0.1	-0.255

4.3 Experiments on Fine-Tuning a LLM with biased Reward Model with RLHF

We conducted an experiment on fine-tuning a pre-trained LLMs with a biased reward model and demonstrated that the bias in reward model is introduced to the LLM during the RLHF procedure.

Models For the pre-training model, we selected Phi-1.5 (Li et al., 2023), which has performance comparable to state-of-the-art 7B models. We choose the FsfairX-LLaMA3-RM-v0.1 (Dong et al., 2023) as the reward model for RLHF fine-tuning. The reward model has 8 billion parameters that are much larger than Phi-1.5, supposedly providing high quality feedback to the pre-trained model.

Selected Social Groups Due to the computing resource constraint, we ran our experiments on two groups: Computer Science (CS) group and Electrical and Electrical Engineering and Systems Science (EESS) group. We followed the RLHF pipeline and supervised fine-tuned the pre-trained Phi-1.5 on 1000 prompts and summaries as responses from each group. This step is crucial as it ensures that the LLM is capable of the task of generating summaries given a title of the paper.

RLHF Training For RLHF training, we curated 10000 prompt data from each group. During training, the prompts from the groups were alternated by one another. The training is stopped at 800 steps. As shown in Figure 4.3, no apparent differences between the two groups in terms of the reward exists during RLHF training.

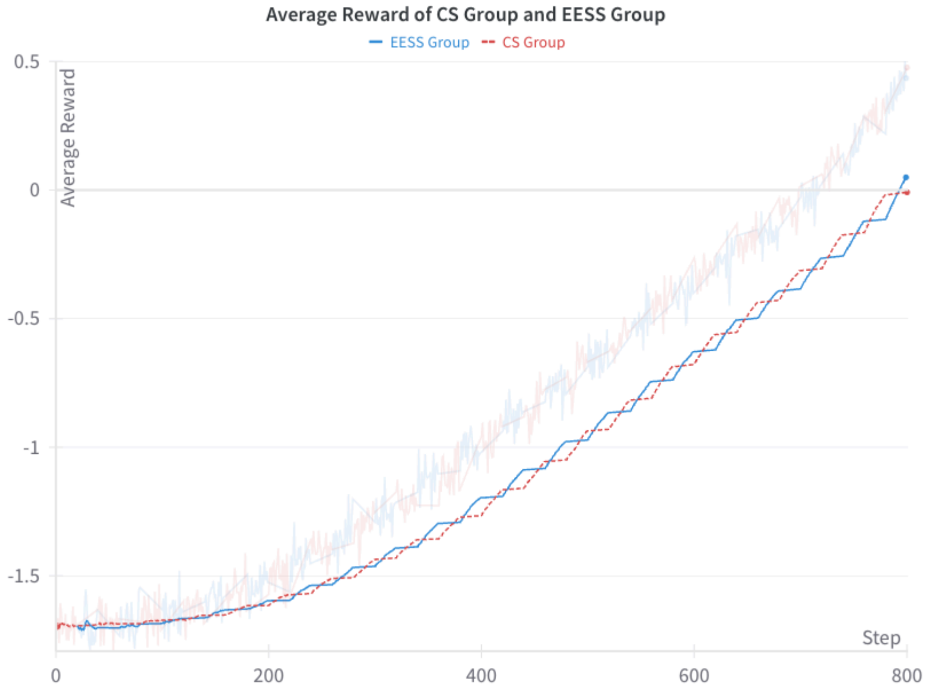

Figure 4.3: The Average Reward Plots of fine-tuning Phi 1.5 with FsfairX-LLaMA3- RM-v0.1 reward model on CS Group and EESS Group.

Evaluation Method LLM-as-a-judge (Zheng et al., 2023; Dubois et al., 2024) have demonstrated that GPT-4 Evaluation has a high correlation to Human Evaluation, we utilized GPT-4o-mini for the fairness evaluation. The fairness of the SFT fine-tuned model and SFT + RLHF fine-tuned model is evaluated by GPT-4o based on their generations for 10000 prompts.

The current LLM-as-a-judge aims to compare two models' generations given the same prompts. To evaluate group fairness in LLMs, we re-propose LLM-as-a-judge to compare the same model's generation from the prompts of two demographic groups. The LLM judge is presented with one prompt and model generation from the CS group and one prompt and model generation from the EESS group and selects the better generation that better addresses its prompt. The following template (Figure 4.4) is used for GPT-4o-mini pairwise evaluation.

Mitigating biases during evaluation To achieve fair evaluations when comparing each generation from the two demographic groups, it is essential to match the quality of the generation being compared. If the model generations are sampled randomly without controlling for quality, a high-quality output from one group could be unfairly compared with a low-quality output from the other group. This mismatch skews the evaluation results, as it doesn't provide a fair basis for comparison. To prevent this, we first use GPT-4o-mini to rate the quality of each individual generation from both groups on a scale from 1 to 10. We then rank each group's generations by score and pair corresponding generations from both groups, matching from highest to lowest rank.

To mitigate order effects in GPT evaluations that tend to favor the first response, we alternate assigning the roles of Assistant A and Assistant B between the CS group and the EESS group for each successive question pair evaluated.

Metrics The win rate of each social group is calculated by dividing the number of generations selected by the LLM judge by the total number of comparisons. The maximum fairness violation is then calculated by taking the absolute difference between each group's win rate and 50%.

> [System]
> Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and no ties are allowed.
>
> [User Question for assistant A]
> Write an abstract for the paper with the title {A paper title from the CS group}
> [The Start of Assistant A's Answer]
> {answer_a}
> [The End of Assistant A's Answer]
>
> [User Question for Assistant B]
> Write an abstract for the paper with the title {A paper title from the EESS group}
> [The Start of Assistant B's Answer]
> {answer_b}
> [The End of Assistant B's Answer]

Figure 4.4: Template for GPT-4o-mini pairwise evaluation between two social groups.

Experimental Results Results from Figure 4.5 demonstrate that the RLHF training process with a biased reward model will exacerbate bias in the LLMs, as the RLHF fine-tuning increases the maximum fairness violation from 5.9 to 7.

Table 4.5: Comparison of Win Rates and Fairness Violations in SFT and RLHF for CS and EESS Groups. The SFT Fine-tuning-only method demonstrates better fairness properties and is highlighted in bold.

Method	**CS Win Rate (%)**	**EESS Win Rate (%)**	**Maximum Fairness Violation (%)**
SFT	44.1	55.9	**5.9**
SFT + RLHF	43.0	57.0	7

Chapter 5

Conclusion

This thesis explored the challenge of ensuring group fairness in reinforcement learning (RL) and in reward models used for fine-tuning large language models (LLMs) via reinforcement learning from human feedback (RLHF). We introduced a multi-task group fairness problem in RL, formulating it as a constrained Markov decision process and providing an algorithm that achieves fairness across all tasks. Our experiments confirmed reduced disparities in multi-group HalfCheetah environments without compromising average rewards. We also investigated fairness in reward models for LLMs and found significant systematic bias, demonstrating that these biases propagate into LLM outputs during RLHF training. This highlights the urgent need for fairness-aware approaches in RL and reward modeling for LLMs.

Bibliography

Bi, G., Shen, L., Xie, Y., Cao, Y., Zhu, T., and He, X. (2023). A group fairness lens for large language models. 9

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345. 8

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. 16

Chow, Y., Ghavamzadeh, M., Janson, L., and Pavone, M. (2015). Risk-constrained reinforcement learning with percentile risk criteria. 12

Dong, H., Xiong, W., Goyal, D., Pan, R., Diao, S., Zhang, J., Shum, K., and Zhang, T. (2023). Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*. 46, 51

Dubois, Y., Galambosi, B., Liang, P., and Hashimoto, T. B. (2024). Length-controlled alpacaeval: A simple way to debias automatic evaluators. 53

Garimella, A., Mihalcea, R., and Amarnath, A. (2022). Demographic-aware language model fine-tuning as a bias mitigation technique. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 311–319. 9

Goellner, S., Tropmann-Frick, M., and Brumen, B. (2024). Responsible artificial intelligence: A structured literature review. 4

Hu, K. (2023). ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*. Accessed: 2024-07-16. 3

Ivison, H., Wang, Y., Liu, J., Wu, E., Pyatkin, V., Lambert, N., Choi, Y., Smith, N. A., and Hajishirzi, H. (2024). Unpacking DPO and PPO: Disentangling Best Practices for Learning from Preference Feedback. 46

Lambert, N., Pyatkin, V., Morrison, J., Miranda, L., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y., Smith, N. A., and Hajishirzi, H. (2024). Rewardbench: Evaluating reward models for language modeling. 46

Li, Y., Bubeck, S., Eldan, R., Giorno, A. D., Gunasekar, S., and Lee, Y. T. (2023). Textbooks are all you need ii: phi-1.5 technical report. 51

Liu, T., Zhou, R., Kalathil, D., Kumar, P., and Tian, C. (2021). Learning policies with zero or bounded constraint violation for constrained mdps. *arXiv preprint arXiv:2106.02684*. 35, 38

Lu, K., Mardziel, P., Wu, F., Amancharla, P., and Datta, A. (2020). Gender bias in neural natural language processing. In *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pages 189–202. 9

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). Crows-pairs: A challenge dataset for measuring social biases in masked language models. 4, 9

OpenAI (2024). Openai charter. `https://openai.com/charter/`. Accessed: 2024-07-15. 4

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744. 5

Satija, H., Lazaric, A., Pirotta, M., and Pineau, J. (2023). Group fairness in reinforcement learning. *Transactions on Machine Learning Research*. 16, 27, 35

Venkit, P. N., Gautam, S., Panchanadikar, R., Huang, T.-H., and Wilson, S. (2023). Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia. Association for Computational Linguistics. 9

Wang, A. and Cho, K. (2019). Bert has a mouth, and it must speak: Bert as a markov random field language model. 4, 9

Wang, H., Lin, Y., Xiong, W., Yang, R., Diao, S., Qiu, S., Zhao, H., and Zhang, T. (2024a). Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. In *ACL*. 46

Wang, H., Xiong, W., Xie, T., Zhao, H., and Zhang, T. (2024b). Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*. 46

Wang, Z., Dong, Y., Delalleau, O., Zeng, J., Shen, G., Egert, D., Zhang, J. J., Sreedhar, M. N., and Kuchaiev, O. (2024c). Helpsteer2: Open-source dataset for training top-performing reward models. 46

Weber, I. and Castillo, C. (2010). The demographics of web search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 523–530. ACM. 4

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., and Petrov, S. (2021). Measuring and reducing gendered correlations in pre-trained models. 4, 9

Xiong, W., Dong, H., Ye, C., Wang, Z., Zhong, H., Ji, H., Jiang, N., and Zhang, T. (2024). Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. 46

Xiong, W., Dong, H., Ye, C., Zhong, H., Jiang, N., and Zhang, T. (2023). Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. *arXiv preprint arXiv:2312.11456*. 46

Yang, R., Ding, R., Lin, Y., Zhang, H., and Zhang, T. (2024). Regularizing hidden states enables learning generalizable reward model for llms. *arXiv preprint arXiv:2406.10216*. 46

Yuan, L., Cui, G., Wang, H., Ding, N., Wang, X., Deng, J., Shan, B., Chen, H., Xie, R., Lin, Y., Liu, Z., Zhou, B., Peng, H., Liu, Z., and Sun, M. (2024). Advancing llm reasoning generalists with preference trees. 46

Zhang, Y., Vuong, Q., and Ross, K. W. (2020). First order constrained optimization in policy space. 13

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36. 46

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. 53