Envy-free Policy Teaching to Multiple Agents

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study envy-free policy teaching. A number of agents independently explore a common Markov decision process (MDP), but each with their own reward function and discounting rate. A teacher wants to teach a target policy to the diverse group of agents, by way of modifying the agents' reward functions, providing additional bonus to certain behaviors or penalizing others. These reward modifications are personalized for each agent. An important question in this setting concerns how a teaching program can be designed so that the agents think that they are treated fairly. We adopt the fairness notion of envy-freeness (EF) to formalize this question and define three different EF notions, each imposing stronger requirements than the previous one. Using these notions, we then investigate several fundamental questions, including the existence of EF solutions in the policy teaching setting, the computation of cost-minimizing solutions, and the price of fairness (PoF), i.e., the increase in cost due to consideration of fairness. We show that an EF solution may not exist when penalties are not allowed, but exists otherwise. Depending on the cost measures, computing a cost-minimizing EF solution can be formulated as convex or linear programming and hence solved efficiently. Asymptotically, the PoF increases but at most linearly with the geometric sum of the discount factor in general, the size of the MDP, and the number of agents involved. Thus, fairness can be incorporated in multi-agent teaching without significant computational or price-of-fairness burdens.

1 Introduction

Incentive design is an important approach to influencing rational agents' behavior. In reinforcement learning (RL), the incentive of an agent is expressed through their reward function, which plays a key role in determining the policy the agent learns [1]. One can thus teach a desired policy to an agent by modifying the agent's reward function, in a way that makes the target policy optimal with respect to the new rewards. This can be very useful in many scenarios. In safe RL, for example, penalties can be imposed on dangerous actions to prevent an agent from executing them [2]. In many cases, personalized teaching programs are useful when agents to be taught are heterogeneous: they might have very different innate reward functions, or apply different discount factors when evaluating a policy. As a result, the agents may find them rewarded/penalized differently for performing the same action in the same situation (see Figure 1). Concerns of fairness may arise due to these differences, leading us to the question of how to design personalized teaching programs so that agents being taught think that they are treated fairly.

For example, in a classroom teaching setting, a teacher wants to teach a group of students to accomplish a task that involves performing a sequence of actions. Personalize teaching schemes might be used concerning the diverse background of the students, and it is usually important that the students feel they are treated equally by the teacher. Similarly, in a principal-agent setting, a company outsources a task to different contractors. Rewards or penalties are stipulated via personalized

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

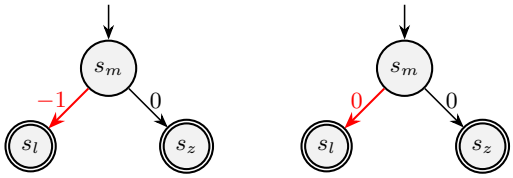

Figure 1: To teach the agent to choose the action leading to state s_l, an additional reward 1 is necessary for an agent whose innate reward function is illustrated on the left, whereas an agent with the reward function on the right already find this target policy optimally. When the personalized teaching programs are used, the agent on the right may think they are treated unfairly as they get no bonus for following the target policy while the agent on the left gets bonus 1.

contracts to ensure that the contractors abide by a desired policy when fulfilling the task. Meanwhile fairness might also be an important consideration as a beneficial factor for long-term partnerships.

Towards answering this question, the first step is to understand what it means to be fair in the setting of policy teaching. Indeed, in a world with growing awareness of equality and transparency, fairness has discussed and evaluated in a wide range of applications. Various fairness notions and concepts have been proposed and used across many domains in technology [3]. We borrow the well-studied fairness notion of *envy-freeness* (EF) from the literature on fair division, which naturally applies to the policy teaching setting. The notion is widely used for settling disputes over property divisions, or finding resource allocations that make participants with different valuations unenvious of each other (e.g., the fair cake-cutting problem [4]). By adapting EF to policy teaching, we aim to find a set of personalized teaching programs for a group of agents, such that no agent would prefer to switch the program they receive with another agent. At the same time, as a basic requirement of policy teaching, each program should also incentivize the corresponding agent who receives it to use the target policy. Besides this basic notion, we also consider two stronger variants of it—one allows an agent to further deviate from the target policy when evaluating the potential benefit had they received another agent's teaching program; the other simply requires all teaching programs to be identical, which gives complete fairness in a sense.

Main Results We investigate several fundamental questions related to the above laid out objective.

- *Existence of a Solution.* The first question is about the existence of an EF solution under the three EF notions of interest. We show that an EF solution always exists and one can be obtained simply by penalizing undesired actions by a sufficiently large value. Nevertheless, the reverse does not hold true: one cannot hope to find an EF solution by rewarding actions desired by the target policy, no matter how large the rewards are set. We demonstrate instances that do not admit any EF solution even with the weakest EF notion when only bonuses are allowed, but also prove that this non-existence issue is resolved if the agents have an identical discount factor.
- *Cost Minimization.* Since reward modification can be very costly, we are also interested in finding out a solution with the least cost. We consider two cost measures in this paper, which consider the norm of the reward modifications made and the cumulative cost the teacher actually pays, respectively. Depending on the choice of the cost measure, we show that computing a cost-minimizing EF solution can be formulated as convex or linear programming and can hence be solved efficiently.
- *Price of Fairness.* Finally, we analyze the price of fairness (PoF), a quantity that measures the (multiplicative) increase of cost due to consideration of fairness—in a similar spirit of the PoA (price of anarchy) in game theory. We show that, asymptotically, the PoF increases linearly with the geometric sum of the discount factor in general, while it may also grow, at most linearly, with the size of the MDP and the number of agents involved depending on the specific EF notion considered.

In summary, our results indicate that the consideration of fairness in addition to the original goal of policy teaching may result in non-existence of workable solutions, but the existence is guaranteed in a fairly wide range of important settings. It does not appear to complicate the problem in terms of computational complexity. The additional cost it incurs grows moderately with the size of the

problem. Thus, fairness can be incorporated in multi-agent teaching without significant computational or price-of-fairness burdens.

Related Work Our work lies at the intersection of policy teaching and envy-free resource allocation.

Without the fairness constraints, our model can be seen as a policy teaching problem for each individual agent in the model. A number of studies have looked at this problem [5, 6]. The problem can be computationally harder though when the target is to hit one in a set of policies rather than a single target [7]. When the teacher is targeting a malicious policy, policy teaching can also be interpreted as reward poisoning [8, 9, 10, 11, 12]. From a technical point of view, these two problems are almost identical and can be solved by using the same techniques. However, conceptually, it is less likely that one would take fairness into consideration when designing a poisoning attack. More broadly, policy teaching can be seen as a sub-field of reward design, a broader area that studies how to influence agents' behaviors thought tweaking the reward function. The objectives of these studies are not limited to inducing a target policy. A notable example is *reward shaping* [13, 14, 15], which aims to accelerate an agent's learning process through reward design. Indeed, while our focus is on policy teaching, the same question of how to design rewards fairly can be asked with other objectives as well. These can be potential directions for future work.

The study of fair division dates back to the early work of Foley [16], and the formal concept of envy-freeness appeared even earlier [17]. Research on fair division has since evolved into a large body of work, with focuses on allocation of divisible or indivisible items [18, 19, 20]. Our work is in particular related to fair allocation of indivisible goods with subsidies [21], as external benefits are provided to change the agents' original incentives in both settings. Our work considers the additional goal of teaching the target policy; indeed, without this goal, an envy-free solution can be achieved trivially by no making any modification to the original reward functions.

We note that there are also other studies on machine teaching settings with multiple agents, or multiple teachers [22, 23, 24], though with very different models from ours. From a mechanism design perspective, our model can also be viewed as one version of the contract design problem [25], where a principal offers an agent a contract for performing a target policy, but might be uncertain about the agent's type (i.e., the original reward function). Our EF solutions correspond exactly to truthful mechanisms that elicit the agent's true type.

2 Preliminaries

There is a set of n agents $1, \ldots, n$. Let $[n] = \{1, \ldots, n\}$. Each agent $i \in [n]$ faces an MDP $\mathcal{M}_i = \langle S, A, R_i, P, \mathbf{z}, \gamma_i \rangle$ with shared state space S, action space A, transition function $P : S \times A \times S \to [0, 1]$ describing the state transition dynamics, and initial state distribution $\mathbf{z}$. Moreover, there is a personal reward function $R_i : S \times A \to \mathbb{R}$ and discount factor γ_i. Whenever agent i takes an action a in state s, a reward $R_i(s, a)$ is generated for this agent; meanwhile the state transitions to a next one $s' \in S$ with probability $P(s, a, s')$. We consider the setting where each agent is concerned with the (expected) cumulative reward, i.e., the discounted sum of rewards with respect to γ_i, obtained over an infinite horizon. Without loss of generality these are actions generated by following some deterministic policy $\pi : S \to A$, and the cumulative reward of this policy for agent i is

$$\rho_i^\pi := \mathbb{E}\left[\sum_{t=0}^{\infty} (\gamma_i)^t \cdot R_i(s_t, a_t) | s_0 \sim \mathbf{z}, \pi\right],$$

where the expectation is taken over the trajectory $(s_t, a_t)_{t=0}^{\infty}$ resulting from an initial state s_0 sampled from $\mathbf{z}$ and the agent executing π subsequently. The agent aims to find an optimal policy, which maximizes ρ_i^π, and this can usually be handled by standard planning and reinforcement learning algorithms.

Note that in the setting we consider throughout this paper, the agents operate independently in separate environments.

2.1 Single-agent Policy Teaching

Consider the situation where we want agent i to execute a target policy $\pi^\star$, but the agent finds a different policy π' optimal for $\mathcal{M}_i$. We hope to influence the agent's decision making and incentivize them to use $\pi^\star$. A typical way is by modifying the agent's reward function by providing additional

rewards. We follow the literature and consider only deterministic target policies—indeed, in general, one cannot hope to incentivize an agent to use precisely a non-deterministic policies only through tweaking the reward function.

Specifically, the teacher chooses a *reward adjustment function* $\delta_i : S \times A \to \mathbb{R}$, or *adjustment* for short, whereby an additional reward $\delta_i(s, a)$ is provided whenever the agent takes an action $a \in A$ in a state $s \in S$. Effectively, the adjustment changes the agent's reward function to $\widetilde{R}_i(s, a) = R_i(s, a) + \delta_i(s, a)$. The agent then optimizes their policy with respect to $\widetilde{R}_i$, and will be incentivized to use $\pi^\star$ if it offers the maximum *cumulative reward* with respect to $\widetilde{R}_i$. The cumulative reward of agent i obtained by executing a policy π is defined as as the sum of discounted rewards as follows; we view it as a function of δ_i:

$$\rho_i^\pi(\delta_i) := \mathbb{E}\left[\sum_{t=0}^{\infty}(\gamma_i)^t \cdot \widetilde{R}_i(s_t, a_t)\Big| s_0 \sim \mathbf{z}, \pi\right].$$

We can define the V-function and Q-function of π conditioned on adjustment δ_i as follows. For all $s \in S$ and $a \in A$:

$$V_i^\pi(s|\delta_i) = Q_i^\pi(s, \pi(s)|\delta_i),$$
$$\text{and} \quad Q_i^\pi(s, a|\delta_i) = \widetilde{R}_i(s, a) + \gamma_i \cdot \mathbb{E}_{s' \sim P(s,a,\cdot)} V_i^\pi(s'|\delta_i).$$

The V-function captures the cumulative reward the agent obtains by starting from s and following π, whereas the Q-function captures the cumulative reward by starting from s and taking action a at the first step. We have $\rho_i^\pi(\delta_i) = V_i^\pi(\mathbf{z}|\delta_i) := \mathbb{E}_{s_0 \sim \mathbf{z}} V_i^\pi(s_0|\delta_i)$. Using these two concepts, the Bellman equation further gives the optimal policy in the MDP: a policy π is optimal if and only if $Q_i^\pi(s, \pi(s)|\delta_i) \geq Q_i^\pi(s, a|\delta_i)$ for all $s \in S$ and $a \in A$.

Incentive Constraint Hence, we require that $\pi^\star$ satisfies the condition described by the Bellman equation. Since the agent may find multiple policies optimal, a robustness guarantee $\epsilon > 0$ is considered to strictly incentivize the agent to use $\pi^\star$. The incentive constraint requires that

$$Q_i^{\pi^\star}(s, \pi^\star(s)|\delta_i) \geq Q_i^{\pi^\star}(s, a|\delta_i) + \epsilon \quad \text{for all } a \neq \pi^\star(s),$$

so that $\pi^\star$ remains optimal even if there is a small noise in the agent's perception of the values.

Cost Measures The goal of the teacher is to find the most cost-efficient way of teaching. In particular, we consider the following cost measures in this paper:

$$\text{cost}(\delta_i) = \|\delta_i\| = \left(\sum_{s \in S, a \in A} (\delta_i(s, a))^2\right)^{1/2}, \tag{1}$$

which measures the norm of the adjustment; or

$$\text{cost}(\delta_i) = \mathbb{E}\left[\sum_{t=0}^{\infty} \overline{\gamma}^t \cdot \delta_i(s_t, a_t)\Big| s_0 \sim \mathbf{z}, \pi\right], \tag{2}$$

which measures the cumulative payment, where $\overline{\gamma}$ is a discount factor from the teacher's perspective. The cumulative payment measure applies in particular when the teacher uses non-negative rewards (i.e., subsidies) as additional incentives for the agent.

3 Teaching Multiple Agents and Fairness

In the multi-agent setting, the teacher needs to provide adjustments to all the agents in $[n]$. We call a collection of adjustments $(\delta_i)_{i \in [n]}$ an *adjustment scheme*. A basic approach for this setting is to deal with each agent separately, by solving a single-agent teaching problem for each of them. The solution obtained via this approach provides personalized adjustments to the agents and it minimizes the teacher's overall cost. Nevertheless, it might not be fair as we showed in the example of Figure 1. To be more specific, we define three fairness notions, each being stronger than the previous one. We start with the following weak EF notion.

Definition 3.1 (Weak envy-freeness (WEF)). An adjustment scheme $(\delta_i)_{i \in [n]}$ is *envy-free* if it holds for all $i \in [n]$ that

$$\rho_i^{\pi^\star}(\delta_i) \geq \rho_i^{\pi^\star}(\delta_j) \qquad \text{for all } j \in [n]. \tag{3}$$

In other words, no agent i would prefer the adjustment for another agent j to their own.

The above notion only compares the agents' benefits under $\pi^\star$. Indeed, we require that δ_i incentivizes agent i to use $\pi^\star$, so the left side of (3) is exactly also the highest possible extra benefit i can obtain given adjustment δ_i. But this is not true for the adjustment on the right side: $\pi^\star$ need not be optimal for agent i with respect to $R_i + \delta_j$; a higher cumulative reward might be attainable if the agent switches to another policy. In some scenarios, this higher potential reward may be a legitimate consideration when fairness is evaluated. The following stronger notion takes this aspect into consideration.

Definition 3.2 (Envy-freeness (EF)). An adjustment scheme $(\delta_i)_{i\in[n]}$ is *strongly envy-free* if it holds for all $i \in [n]$ that:

$$\max_\pi \rho_i^\pi(\delta_i) \geq \max_\pi \rho_i^\pi(\delta_j) \quad \text{for all } j \in [n]. \tag{4}$$

An even stronger fairness notion defined below simply requires the same adjustment to be applied to all the agents.

Definition 3.3 (Strong envy-freeness (SEF)). An adjustment scheme $(\delta_i)_{i\in[n]}$ is *completely fair* if $\delta_1 = \cdots = \delta_n$.

Let $\mathcal{D}_{\text{WEF}}$, $\mathcal{D}_{\text{EF}}$, and $\mathcal{D}_{\text{SEF}}$ denote the sets of adjustment schemes complying with the above fairness notions respectively. It is not hard to see that: $\mathcal{D}_{\text{WEF}} \supseteq \mathcal{D}_{\text{EF}} \supseteq \mathcal{D}_{\text{SEF}}$.

Besides the fairness guarantees, we also want the adjustment scheme to—as our initial objective—incentivize the agents to use the target policy, i.e., to satisfy the incentive constraints we defined above; we call such adjustment schemes *feasible* schemes (Definition 3.4). Indeed, without the feasibility guarantee, all the fairness notions above can be trivially satisfied by providing zero additional reward to every agent. In addition, sometimes only non-negative additional rewards are allowed, e.g., when we can provide the agents with subsidies but cannot penalize them. Hence, we define *non-negative* adjustment schemes (Definition 3.5).

Definition 3.4 (Feasibility). An adjustment scheme $(\delta_i)_{i\in[n]}$ is *feasible* (with respect to a given robustness guarantee $\epsilon > 0$) if for all $i \in [n]$, it holds that

$$\widetilde{Q}_i^{\pi^\star}(s, \pi^\star(s)) \geq \widetilde{Q}_i^{\pi^\star}(s, a) + \epsilon \quad \text{for all } a \neq \pi^\star(s). \tag{5}$$

Definition 3.5 (Non-negativity). An adjustment scheme $(\delta_i)_{i\in[n]}$ is *non-negative* if $\delta_i(s, a) \geq 0$ for all $i \in [n]$, $s \in S$, and $a \in A$.

Similarly to the single-agent teaching problem, we are interested in obtaining a cost-minimizing solution. We consider the sum of the costs and define the cost of an adjustment scheme δ as

$$\text{cost}(\delta) = \sum_{i\in[n]} \text{cost}(\delta_i).$$

Next, before we delve into the computation of a cost-minimizing solution, we first investigate the existence of a solution with respect to the above defined fairness notions.

4 Existence of a Fair Feasible Solution

Throughout this section, we assume that the original rewards are bounded in the interval $[-h, h]$, i.e., $R_i(s, a) \in [-h, h]$ for all s, a, and i. Our first result shows that a fair and feasible solution always exists under all of the above fairness notions, in particular under the strongest notion SEF.

Theorem 4.1. *For any robustness guarantee $\epsilon > 0$, an SEF and feasible adjustment scheme always exists.*[1]

Proof sketch. The idea is to penalize actions off the target policy by a sufficiently large value. We construct an adjustment scheme $(\delta_i)_{i\in i}$ where

$$\delta_i(s, a) = \begin{cases} 0, & \text{if } a = \pi^\star(s) \\ -\max_{i'\in[n]} \frac{2h}{1-\gamma_{i'}} - \epsilon, & \text{otherwise} \end{cases}$$

[1] Full proofs and omitted proofs can all be found in the appendix.

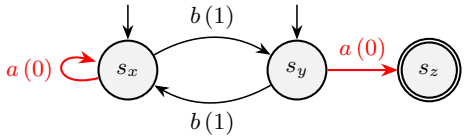

Figure 2: There are two agents, whose discount factors are $\gamma_1 = 0.9$ and $\gamma_2 = 0.5$, respectively. $S = \{s_x, s_y, s_z\}$, $A = \{a, b\}$, and all transitions are deterministic. Originally, the agents' reward functions are the same, with $R_1(s, a) = R_2(s, a) = 0$ and $R_1(s, b) = R_2(s, b) = 1$ for all $s \in S$, as annotated on the edges. The states s_x and s_y are chosen as the initial state with equal probability. The target policy $\pi^\star$, highlighted as the red edges, is such that $\pi^\star(s) = a$ for all $s \in S$, i.e., it always chooses action a.

for all $s \in S$ and $i \in [n]$. The scheme is obviously SEF as δ_i is the same for all the agents. It can also be verified that it is feasible. Intuitively, the penalty is so large such that once the agent is penalized, the subsequent cumulative rewards cannot compensate this penalty even if the highest rewards are attained at every step. □

Nevertheless, if we only allow non-negative schemes, the existence of a feasible solution cannot be taken for granted. As we prove in Theorem 4.2, the example illustrated in Figure 2, albeit involving only two agents, does not admit any EF feasible solution (and hence neither an SEF one).

Theorem 4.2. *For any robust guarantee $\epsilon \geq 0$, a feasible adjustment scheme that is WEF and non-negative may not exist, even when there are only two agents and the agents' reward functions are the same (but discount factors are different).*

Proof. We show that there exists no feasible adjustment scheme that is WEF and non-negative in the example illustrated in Figure 2. Suppose for the sake of contradiction that there exists a scheme (δ_1, δ_2) which is EF, non-negative, and feasible.

Without loss of generality, we can assume that $\delta_1(s, b) = \delta_2(s, b) = 0$ for all $s \in S$. Indeed, it is not hard to see that if there exists a WEF and feasible scheme with some or all of these values being strictly positive, it will remain WEF and feasible if these values are reset to 0. Hence, it remains to pin down the values for action a in the adjustment scheme. For ease of description, let $x_i = \delta_i(s_x, a)$ and $y_i = \delta_i(s_y, a)$ for $i \in \{1, 2\}$.

We first argue that the following two inequalities hold:

$$y_1 \geq y_2, \quad \text{and} \quad x_2 \geq x_1. \tag{6}$$

To see this, consider the WEF constraints. The adjustment scheme considered is WEF, so it satisfies (3), which implies that $\rho_i^{\pi^\star}(\delta_i) \geq \rho_i^{\pi^\star}(\delta_{-i})$, where $-i$ means the other index in $\{1, 2\}$. Hence,

$$0.5 \cdot V_i^{\pi^\star}(s_x|\delta_i) + 0.5 \cdot V_i^{\pi^\star}(s_y|\delta_i) \geq 0.5 \cdot V_i^{\pi^\star}(s_x|\delta_{-i}) + 0.5 \cdot V_i^{\pi^\star}(s_y|\delta_{-i}). \tag{7}$$

It is easy to derive the values of s_x and s_y as neither of them depends on the values of any other states. When agent i uses $\pi^\star$, we have

$$V_i^{\pi^\star}(s_x|\delta_j) = Q_i^{\pi^\star}(s_x, a|\delta_j) = \tfrac{1}{1-\gamma_i} \cdot y_j, \tag{8}$$

$$\text{and} \quad V_i^{\pi^\star}(s_y|\delta_j) = Q_i^{\pi^\star}(s_y, a|\delta_j) = x_j. \tag{9}$$

Plugging these two equations back into (7) gives

$$0.5 \cdot \tfrac{1}{1-\gamma_i} \cdot y_i + 0.5 \cdot x_i \geq 0.5 \cdot \tfrac{1}{1-\gamma_{-i}} \cdot y_{-i} + 0.5 \cdot x_{-i},$$

where 0.5 is the probability of the initial states. Replacing γ_i with the corresponding values gives

$$10 \cdot y_1 + x_1 \geq 10 \cdot y_2 + x_2 \tag{10}$$

$$2 \cdot y_2 + x_2 \geq 2 \cdot y_1 + x_1 \tag{11}$$

It follows that (10)+(11) gives $y_1 \geq y_2$, and (10)+5·(11) gives $x_2 \geq x_1$.

Next, we turn to the feasibility constraints. The assumption that δ_i is feasible means that

$$Q_i^{\pi^\star}(s_x, a|\delta_i) \geq Q_i^{\pi^\star}(s_x, b|\delta_i) + \epsilon = 1 + \gamma_i \cdot \widetilde{V}_i^{\pi^\star}(s_y) + \epsilon$$
$$\text{and} \quad Q_i^{\pi^\star}(s_y, a|\delta_i) \geq Q_i^{\pi^\star}(s_y, b|\delta_i) + \epsilon = 1 + \gamma_i \cdot \widetilde{V}_i^{\pi^\star}(s_x) + \epsilon$$

Plugging the value function and Q-function ((8) and (9)) into the above two equations gives:

$$1 + \tfrac{\gamma_i}{1-\gamma_i} \cdot y_i \leq x_i \leq \tfrac{1}{\gamma_i(1-\gamma_i)} \cdot y_i - \tfrac{1}{\gamma_i}. \tag{12}$$

Hence, applying (12) and (6), we have

$$9 \cdot y_1 + 1 \leq x_1 \leq x_2 \leq 4 \leq 4 \cdot y_1 - 2,$$

This means that $y_1 < 0$ and contradicts the assumption that δ is a non-negative scheme. □

It turns out that the agents' discount factors also play a crucial role: an identical discount factor is sufficient for ensuring the existence of a feasible SEF solution. We present this result below but leave the proof to the appendix.

Theorem 4.3. *When the agents have an identical discount factor, a feasible adjustment scheme that is also SEF and non-negative always exists, for any robustness guarantee $\epsilon > 0$.*

Proof sketch. Suppose that $\gamma_1 = \cdots = \gamma_n = \gamma$. Let $H = \frac{2\cdot\gamma}{1-\gamma} \cdot h + 2h + \epsilon$. We construct the following scheme $\delta = (\delta_i)_{i\in[n]}$:

$$\delta_i(s,a) = \begin{cases} H \cdot \sum_{s' \in S \setminus S^{\mathrm{T}}} P(s,a,s') + \frac{1}{1-\gamma} \cdot H \cdot \sum_{s' \in S^{\mathrm{T}}} P(s,a,s'), & \text{if } a = \pi^\star(s) \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

for all $s \in S$ and $i \in [n]$, where S^{T} denotes the set of terminal states in S.

The scheme is obviously non-negative and SEF. Moreover, we can also show that it is feasible. Intuitively, δ_i results the agent getting a reward that is sufficiently large (and is roughly the same) at every step if they follow $\pi^\star$. The rewards are adjusted by a factor of $1/(1-\gamma)$ at terminal states so that it is as if the process continues forever. Hence, under δ_i, the process is roughly the same as an infinite process where the agent gets a constant positive reward at every step if they take an action stipulated by $\pi^\star$, and reward 0 otherwise—the optimal choice for the agent in such a process is to always follow $\pi^\star$. □

5 Computing an Optimal Fair Solution

In this section, we investigate the computational problems of fair policy teaching. Our main result is that for each of the fair notions we defined the set of fair solutions lie in a convex polytope defined by polynomially many linear constraints. To find out a cost minimizing solution from this convex set can then be formulated as a convex optimization problem, given that the two types of cost functions we consider (i.e., (1) and (2)) are convex functions of the adjustment scheme. We show how to write down the various types of constraints considered as linear constraints next. We start from the feasibility constraints.

Feasibility Constraints A feasible scheme δ is characterized by the following linear constraints, where we add an auxiliary variable $V_i(s)$ for each $s \in S$, and $Q_i(s,a)$ for pair $(s,a) \in S \times A$ — these auxiliary variables correspond to the V- and Q-functions of the target policy when δ is applied.

$$V_i(s) = Q_i(s, \pi^\star(s)) \qquad \forall i, s \tag{14a}$$
$$Q_i(s,a) = R_i(s,a) + \delta_i(s,a) + \gamma_i \textstyle\sum_{s' \in S} P(s,a,s') \cdot V_i(s') \qquad \forall i, s, a \tag{14b}$$
$$Q_i\left(s, \pi^\star(s)\right) \geq Q_i(s,a) + \epsilon \qquad \forall i, s, a \neq \pi^\star(s) \tag{14c}$$

Indeed, the first two lines above follow from the Bellman equation and captures the values $V_i^{\pi^\star}(s|\delta_i)$ and $Q_i^{\pi^\star}(s,a|\delta_i)$; the last line is the incentive constraints and enforces the feasibility of δ.

We then consider each of the fairness notions.

	PoWEF	PoEF	PoSEF
Norm (1)	$\Theta(\lambda \cdot \sqrt{m})$	$\Theta(\lambda \cdot n \cdot \sqrt{m})$	$\Theta(\lambda \cdot n \cdot \sqrt{m})$
Cumulative (2)	$\Theta(\lambda \cdot n)$	$\Theta(\lambda \cdot n)$	$\Theta(\lambda \cdot n)$

Table 1: Summary of the price of fairness.

SEF Constraints To enforce SEF simply amounts to the following constraints for each pair of agents $i, j \in [n]$, which enforces the schemes to be identical:

$$\delta_i(s,a) = \delta_j(s,a) \qquad \forall s, a. \tag{15}$$

WEF Constraints To enforce WEF, we add additional variables $V_{i,j}$ and $Q_{i,j}$ to capture the values $V_i^{\pi^\star}(s|\delta_j)$ and $Q_i^{\pi^\star}(s,a|\delta_j)$, i.e., the values agent i would have got had they been offered the adjustment for agent j. We then add the following constraints—the same format as the Bellman equation—to force these additional variables to acquire the desired values.

$$V_{i,j}(s) = Q_{i,j}(s, \pi^\star(s)) \qquad \forall i, j, s \tag{16a}$$
$$Q_{i,j}(s,a) = R_i(s,a) + \delta_j(s,a) + \gamma_i \textstyle\sum_{s' \in S} P(s,a,s') \cdot V_{i,j}(s') \qquad \forall i, j, s, a \tag{16b}$$

Thus, WEF simply amounts to the following constraints for each pair of agents $i, j \in [n]$ (recall that $\mathbf{z}$ is the distribution of the initial state):

$$\textstyle\sum_{s \in S} z_s \cdot V_i(s) \geq \sum_{s \in S} z_s \cdot V_{i,j}(s) \qquad \forall s, a \tag{17}$$

EF Constraints Similarly to the approach for handling the WEF constraints, we need additional variables to capture the values of each agent i had then been offered adjustment δ_j. Indeed, we also use Constraints in (16) and (17) but replace (16a) with the following one:

$$V_{i,j}(s) \geq Q_{i,j}(s,a) \qquad \forall i, j, s, a \tag{18a}$$

which associates $V_{i,j}(s)$ to the maximum $Q_{i,j}(s,a)$, instead of $Q_{i,j}(s, \pi^\star(s))$.

It should be noted that under these constraints, $V_{i,j}(s)$ in a solution will not necessarily be equal to the V-values of the optimal policy—it will only be an upper bound of them. However, this will not cause any issue to our approach as our goal is not to compute the V-values. Indeed, if the actual V-values does not satisfy (17), then these constraints cannot be satisfied by any upper bounds of the actual V-values, either.

Non-negativity Constraints Finally, to enforce non-negativity, we simply need the additional constraints: $\delta_i(s,a) \geq 0$ for all i, s, and a.

6 Price of Fairness

We consider the price of fairness (PoF) in this section. The PoF measures the increase of teaching cost due to consideration of fairness. It is in a similar spirit of the well-known concept of the price of anarchy in economics and game theory. More specifically, we define the PoWEF, PoEF, and PoSEF for our three fairness notions, which stand for the prices of WEF, EF, and SEF, respectively. Formally, let $\mathcal{I}_{n,m,\lambda}$ be the set of instances with n agents, m state-action pairs (i.e., $m = |S| \cdot |A|$), and $\frac{1}{1-\gamma_i} \leq \lambda$ for all $i \in [n]$. We define

$$\mathrm{PoEF}(n,m,\lambda) := \max_{I \in \mathcal{I}_{n,m,\lambda}} \frac{\min_{\delta\text{: EF and feasible for } I} \mathrm{cost}(\delta)}{\min_{\delta\text{: feasible for } I} \mathrm{cost}(\delta)}.$$

PoWEF and PoSEF can be defined the same way with the corresponding notions. We consider both types of cost functions we defined in Section 2. When cumulative payment is considered (i.e., (2)), we require adjustment schemes to be non-negative for this measure to be meaningful. Since a feasible fair solution may not exist with this requirement, we only consider instances with an identical discount factor for all the agents, in which case the existence of a feasible fair solution is guaranteed (see Theorem 4.3).

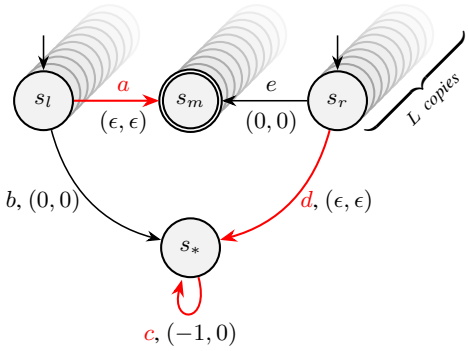

Figure 3: There are n agents with discount factors $\gamma_1 = \cdots = \gamma_n = \gamma$. $A = \{a, b, c, d\}$ and all transitions are deterministic. The original rewards for agents 1 and 2 are annotated on the corresponding edges (numbers in parentheses), which are the same for both agents except the ones for c. Moreover, the reward functions are completely identical for agents $2, \ldots, n$. There are L sets of copies of s_l, s_m, and s_r, each connected to copies in the same set and in addition s_* in the same way illustrated above. The initial state follows a uniform distribution over s_l, s_m, and all their copies. The target policy is highlighted in red: $\pi^\star(s_l) = a$, $\pi^\star(s_r) = d$, and $\pi^\star(s_*) = c$.

We analyze the asymptotic growth of the PoF. The results are summarized in Table 1. The PoF increases linearly with λ in all the cases, and with the number of agents involved, with the only exception of PoWEF when the norm is used as the cost function. It also increases but sub-linearly with the size of the MDP with this cost function but not with the cumulative payment. Our results are all asymptotically tight.

Due to space limit, we leave the detailed proofs of the bounds to the appendix and only provide some intuition about the growth of PoF here. Consider the example illustrated in Figure 3. Without fairness consideration, all agents except agent 1 already find the target policy optimal, whereas agent 1 prefers action e to d at state s_r. Overall, it suffices to give a bonus of 1 on (s_*, c) for agent 1, the cost of which is 1. Now we consider fairness constraints, and suppose that we still provide a bonus $\delta_1(s_*, c) = 1$. The consequence is that agents $2, \ldots, n$ will be envious of this bonus to agent 1. To achieve SEF for example, the same bonus will have to be offered to these agents as well. Nevertheless, while this removes envies, it also incentivizes the agents to take action b instead of a, leading to violation of the feasibility constraint. Inevitably, to construct a feasible and fair in this example, we cannot hope to only modify the reward for action c (or only for agent 1 with EF and SEF). Modifying the other rewards is however much more costly (under the norm cost function) since each one of them has L copies of themselves, which requires the same modification by symmetry.

7 Conclusion

We studied the fairness issue in policy teaching and adopted the notion of envy-freeness to formalize the problem. Several fundamental questions regarding the existence of a fair solution, the computation of cost-minimization solution, and the price of considering fairness have been answered in the paper. For future work, it would be interesting to generalize the model to other reward design settings, where a larger set of design objectives can be considered. In the case of non-negative adjustment schemes, we prove that a fair solution exists for one setting. An interesting future research direction would be to prove this result for a broader class of settings.

Limitations As we mentioned earlier in the paper, policy teaching is equivalent to reward poisoning from a technical point of view. Hence, almost any techniques that applies to policy teaching also applies immediately to solve reward poisoning problems. We note this potential negative social impact of our results but also remark that since our consideration is fairness we are not aware of any scenario where a malicious party intends to launch poisoning attack while considering fairness at the same time. There are many other notions of fairness, equity, and equality. The EF notions we studied are concerned with the additional rewards provided by the adjustment scheme but not with the overall rewards. Hence, they are not applicable if the latter should be the key consideration.

References

[1] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[2] Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *ICLR*, 2018.

[3] Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. This thing called fairness: Disciplinary confusion realizing a value in technology. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019.

[4] Ariel D. Procaccia. Cake cutting: Not just child's play. *Commun. ACM*, 56(7):78–87, jul 2013.

[5] Haoqi Zhang and David C. Parkes. Value-based policy teaching with active indirect elicitation. In *AAAI*, pages 208–214, 2008.

[6] Haoqi Zhang, David C Parkes, and Yiling Chen. Policy teaching through reward function learning. In *EC*, pages 295–304, 2009.

[7] Kiarash Banihashem, Adish Singla, Jiarui Gan, and Goran Radanovic. Admissible policy teaching through reward design. *arXiv preprint arXiv:2201.02185*, 2022.

[8] Yuzhe Ma, Xuezhou Zhang, Wen Sun, and Jerry Zhu. Policy poisoning in batch reinforcement learning and control. In *NeurIPS*, pages 14543–14553, 2019.

[9] Yunhan Huang and Quanyan Zhu. Deceptive reinforcement learning under adversarial manipulations on cost signals. In *GameSec*, pages 217–237, 2019.

[10] Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy teaching in reinforcement learning via environment poisoning attacks. *CoRR*, abs/2011.10824, 2020.

[11] Yanchao Sun, Da Huo, and Furong Huang. Vulnerability-aware poisoning mechanism for online rl with unknown dynamics. In *ICLR*, 2021.

[12] Xuezhou Zhang, Yuzhe Ma, Adish Singla, and Xiaojin Zhu. Adaptive reward-poisoning attacks against reinforcement learning. In *ICML*, pages 11225–11234, 2020.

[13] Maja J Mataric. Reward functions for accelerated learning. In *ICML*, pages 181–189, 1994.

[14] Marco Dorigo and Marco Colombetti. Robot shaping: Developing autonomous agents through learning. *Artificial intelligence*, 71(2):321–370, 1994.

[15] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, pages 278–287, 1999.

[16] Duncan Karl Foley. *Resource allocation and the public sector*. Yale University, 1966.

[17] G Gamow and M Stern. Puzzle-math, edn. *Viking*, 1958.

[18] Hervé Moulin. *Fair division and collective welfare*. MIT press, 2004.

[19] Kenneth J Arrow, Amartya Sen, and Kotaro Suzumura. *Handbook of social choice and welfare*, volume 2. Elsevier, 2010.

[20] Tim Roughgarden. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010.

[21] Daniel Halpern and Nisarg Shah. Fair division with subsidy. In *Proceedings of 12th International Symposium (SAGT '19)*, page 374–389, Berlin, Heidelberg, 2019. Springer-Verlag.

[22] Xiaojin Zhu, Ji Liu, and Manuel Lopes. No learner left behind: On the complexity of teaching multiple learners simultaneously. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI '17*, pages 3588–3594, 2017.

[23] Teresa Yeo, Parameswaran Kamalaruban, Adish Singla, Arpit Merchant, Thibault Asselborn, Louis Faucon, Pierre Dillenbourg, and Volkan Cevher. Iterative classroom teaching. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI '19)*, 2019.

[24] Ritesh Noothigattu, Tom Yan, and Ariel D. Procaccia. Inverse reinforcement learning from like-minded teachers. volume 35, pages 9197–9204, May 2021.

[25] Paul Duetting, Tim Roughgarden, and Inbal Talgam-Cohen. The complexity of contracts. *SIAM Journal on Computing*, 50(1):211–254, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]