Envy-free Policy Teaching to Multiple Agents

Jiarui Gan
University of Oxford
jiarui.gan@cs.ox.ac.uk

Rupak Majumdar
MPI-SWS
rupak@mpi-sws.org

Goran Radanovic
MPI-SWS
gradanovic@mpi-sws.org

Adish Singla
MPI-SWS
adish@mpi-sws.org

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 this diverse group of agents, by means of modifying the agents' reward functions: providing additional bonuses to certain actions, or penalizing them. When personalized reward modification programs are used, an important question is how to design the programs so that the agents think they are treated fairly. We adopt the notion of envy-freeness (EF) from the literature on fair division to formalize this problem and investigate several fundamental questions about the existence of EF solutions in our setting, the computation of cost-minimizing solutions, as well as the price of fairness (PoF), which measures the increase of cost due to the consideration of fairness. We show that 1) an EF solution may not exist if penalties are not allowed in the modifications, but otherwise always exists. 2) Computing a cost-minimizing EF solution can be formulated as convex optimization and hence solved efficiently. 3) The PoF increases but at most quadratically with the geometric sum of the discount factor, and at most linearly with the size of the MDP and the number of agents involved; we present tight asymptotic bounds on the PoF. These results indicate that fairness can be incorporated in multi-agent teaching without significant computational or PoF 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 [1]. One can thus teach a desired policy to an agent by modifying their reward function, in a way that makes the target policy optimal with respect to the modified rewards. 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 against heterogeneous agents, who might have very different innate reward functions or apply different discounting rate. 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 arise, and we ask the question of how to design fair personalized teaching programs so that the agents think that they are treated fairly.

To be more concrete, consider a language teaching setting modeled as an MDP. Each state of the MDP represents the overall skill of a student (agent) and is encoded as the student's performance on different components such as listening, reading, speaking, and writing. Actions available to the students are defined by the levels of effort they put into the components, and it is desired that they always put more effort into the components that they are currently weaker at, which is also the target

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

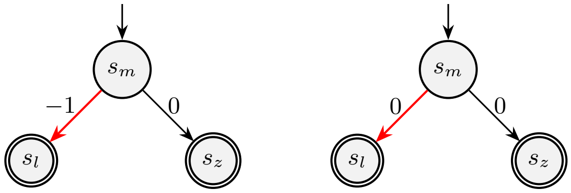

Figure 1: To teach agents to choose the action leading to state s_l, an additional reward 1 is necessary for an agent whose innate reward function is the one on the left, whereas an agent with the reward function on the right already finds this target policy optimal, so no additional reward is needed. When these two agents are being taught together, the agent on the right would think they are treated unfairly as they get no bonus for following the target policy while the agent on the left gets bonus 1.

policy the teacher aims to teach. The students' innate reward functions are defined by their interests, which vary across the classroom: some students may be more interested in reading, some enjoy speaking, and some are just not a fan of any of them. The teacher can assign additional credits to incentivize the students to follow the target policy (e.g., credits that can be used to exchange snacks, or that will be considered in the final evaluation). Similar interactions may also happen with other types of training programs in various domains, such as sports training. They can happen both in physical classrooms and virtual classrooms such as language educational apps (e.g., Duolingo uses a credit system where credits can be used to unlock next learning levels). Beyond classroom teaching, examples can also be found in principal-agent settings. For example, a company wants to outsource a task to different contractors. Rewards or penalties are stipulated through customized contracts to ensure that contractors comply with a desired policy when performing the task. Meanwhile, fairness is important as a beneficial factor for long-term partnerships.

1.1 Approach and Results

Our 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 been discussed and evaluated in a wide range of domains. Various concepts and notions of fairness have been proposed and used [3]. We borrow the well-studied fairness notion of *envy-freeness* (EF) from the literature on fair division. It is a notion that has been used for settling disputes over property divisions or deciding how to split an apartment rent [e.g., 4, 5]. Applying EF to policy teaching, we aim to find a set of personalized teaching programs, 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 to use the target policy. Besides the basic version of EF, we also consider two stronger variants: one allows an agent to further deviate from the target policy when evaluating how much they would have got had they been offered another agent's teaching program; the other simply requires all teaching programs to be identical, which is completely fair in a sense.

We investigate several fundamental questions about EF policy teaching.

- *Existence of an EF 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 only by rewarding actions desired by the target policy. We demonstrate instances that do not admit any EF solution when penalties are not allowed even with the weakest EF notion; we also prove that this non-existence issue is resolved if the agents have the same discount factor.
- *Cost Minimization.* Since reward modification can be very costly, we are also interested in finding out an EF solution with the least cost. We consider the norm of the modification and show that computing a cost-minimizing EF solution can be formulated as convex optimization 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 the cost due to consideration of fairness and is in a similar spirit of the *price of anarchy* (PoA) in game theory [6]. We present tight asymptotic bounds on the PoF. The

PoF increases at most quadratically with the geometric sum of the discount factor and linearly with the size of the MDP in general, while it may also grow linearly with 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 increase the computational complexity of policy teaching, while the additional cost it incurs grows moderately with the size of the problem. The results indicate that fairness can be incorporated in multi-agent teaching without significant computational or PoF burdens.

1.2 Related Work

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

Policy Teaching 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 [7, 8]. The problem can be computationally harder though when the target is to hit one in a set of policies rather than a single target [9]. When the teacher is targeting a malicious policy, policy teaching can also be interpreted as reward poisoning [10, 11, 12, 13, 14]. 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 [15, 16, 17], 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.

Fair Division The study of fair division dates back to the early work of Foley [18], and the formal concept of envy-freeness appeared even earlier [19]. Research on fair division has since evolved into a large body of work, with focuses on allocation of divisible or indivisible items [20, 21, 22, 23]. Our work is in particular related to fair allocation of indivisible goods with subsidies [24], where external benefits are provided to change the agents' original incentives. The difference is that no items are allocated in our model and our goal in addition to achieving fairness is to teach the target policy.

We note that there are also other studies on machine teaching settings involving multiple agents or multiple teachers [25, 26, 27], 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 [28], 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 are n agents $1, \dots, n$. Let $[n] = \{1, \dots, n\}$. Each agent $i \in [n]$ faces an MDP $\mathcal{M}_i = \langle S, A, R_i, P, \mathbf{z}, \gamma_i \rangle$. The MDPs have the same state space S, action space A, transition function $P : S \times A \times S \to [0, 1]$, and initial state distribution $\mathbf{z}$. Moreover, there is a reward function $R_i : S \times A \to \mathbb{R}$ and discount factor γ_i for each agent $i \in [n]$. 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 $s' \in S$ 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 the factor γ_i, obtained over an infinite horizon. More specifically, the cumulative reward of agent i for executing a policy $\pi : S \to \Delta(A)$ is

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

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. Each 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.

Throughout the paper, we consider the setting where the agents operate independently in separate environments. Their payoffs are only determined by their own policies.

2.1 Single-agent Policy Teaching

Consider the situation where we want an agent i to execute a target policy $\pi^\star$, but the agent finds a different policy π' optimal for $\mathcal{M}_i$. To incentivize the agent to use $\pi^\star$, a typical way is to modify the the reward function by providing additional rewards (positive or negative). We follow the literature and consider only deterministic target policies. (Indeed, in general, one cannot hope to incentivize an agent to use a non-deterministic policy only by 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 payoff (cumulative reward) with respect to $\widetilde{R}_i$. We can view each agent's payoff for policy π as a function of δ_i as follows:

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

Moreover, we define the V-function and Q-function of π given adjustment δ_i as:

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

The V-function captures the expected cumulative reward by starting from s and following π. The Q-function captures the expected cumulative reward by starting from s, taking action a at the first step, and following π subsequently. We have

$$\rho_i^\pi(\delta_i) = V_i^\pi(\mathbf{z} \mid \delta_i) := \mathbb{E}_{s_0 \sim \mathbf{z}} V_i^\pi(s_0 \mid \delta_i).$$

Using these two functions, the Bellman equation further characterizes the optimal policy in the MDP: a policy π is optimal if and only if the following Bellman optimality equation holds: $Q_i^\pi(s, \pi(s) \mid \delta_i) \geq Q_i^\pi(s, a \mid \delta_i)$ for all $s \in S$ and $a \in A$.

Incentive Constraints Hence, the goal of policy teaching is to make the target policy $\pi^\star$ a solution to the Bellman optimality equation. Since the agent may find multiple policies optimal, a robustness guarantee $\epsilon > 0$ is imposed to *strictly* incentivize the agent to use $\pi^\star$, and this results in the following incentive constraints:

$$Q_i^{\pi^\star}(s, \pi^\star(s) \mid \delta_i) \geq Q_i^{\pi^\star}(s, a \mid \delta_i) + \epsilon \quad \text{for all } a \neq \pi^\star(s). \tag{1}$$

The constraints ensure the optimality of $\pi^\star$ even if there is a small error in the Q-values.

Cost Measures In addition to incentivizing $\pi^\star$, the teacher also wants to find the most cost-efficient way of teaching. We consider the norm of the adjustment, which means the following cost measure:

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

3 Teaching Multiple Agents and EFness

In the multi-agent setting, the teacher provides an adjustment to every agent 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 agent. The solution obtained via this approach provides personalized adjustments to the agents and it minimizes the teacher's total 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 *weakly 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 $\pi^\star$. When δ_i incentivizes agent i to use $\pi^\star$, the left side of (3) is also exactly the highest possible 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 concern when fairness is evaluated. The following stronger notion takes this aspect into account.

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

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

An even stronger fairness notion defined below simply requires the same adjustment to be applied to all the agents. It is completely fair in a sense.

Definition 3.3 (Strong envy-freeness (SEF)). An adjustment scheme $(\delta_i)_{i\in[n]}$ is *strongly envy-free* if $\delta_i = \delta_j$ for all $i, j \in [n]$.

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

Besides achieving EFness, the original goal of policy teaching is to incentivize the agents to use $\pi^\star$. Hence, we will call adjustment schemes that satisfy equation (1) *feasible* schemes (Definition 3.4). Indeed, the definitions of WEF and SEF would be meaningless without the feasibility requirement, in which case they can be achieved trivially by providing zero additional reward to every agent. (The definition of EF, on the other hand, already incorporates the incentive constraints as equation (3) also includes the case where $i = j$, except that there is no ϵ robustness requirement.)

Definition 3.4 (Feasibility). An adjustment scheme $(\delta_i)_{i\in[n]}$ is *feasible* (with respect to a robustness guarantee $\epsilon > 0$) if equation (1) holds for all $i \in [n]$.

Sometimes only bonuses (non-negative additional rewards) are allowed, e.g., when one can provide the agents with subsidies but cannot penalize them. Hence, we are also interested in finding *non-negative* adjustment schemes defined as follows.

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 policy teaching problem, cost-minimizing solutions are desired. We consider the sum of the teaching costs in the multi-agent setting:

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

4 Existence of Fair Solutions

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 and requirements. 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.

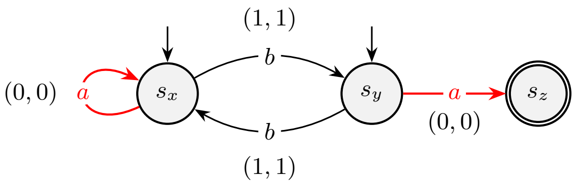

Figure 2: There are two agents, whose discount factors are $\gamma_1 = 0.9$ and $\gamma_2 = 0.5$, respectively. $S = \{s_x, s_y, s_z\}$, $A = \{a, b\}$, and all transitions are deterministic. Originally, the agents' reward functions are the same and their rewards are annotated as vectors on the edges,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

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 not following the target policy by a sufficiently large value. We construct an adjustment scheme $(\delta_i)_{i \in i}$ where

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

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 for the loss due to this penalty even if the highest rewards are attained at every subsequent step. □

Nevertheless, the reverse is not true. If we only allow non-negative schemes, the existence of a feasible solution cannot be taken for granted, and in general one cannot hope to teach a target policy by placing large bonuses on actions following the target policy. As we prove in Theorem 4.2, the example illustrated in Figure 2 does not admit any EF feasible solution (and hence neither an SEF one), even though it involves only two agents and the agents have the same reward function (but different discount factors).

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 their reward functions are the same.*

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:

$$x_1 \geq x_2, \quad \text{and} \quad y_2 \geq y_1. \tag{5}$$

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

$$0.5 \cdot V_i^{\pi^\star}(s_x \mid \delta_i) + 0.5 \cdot V_i^{\pi^\star}(s_y \mid \delta_i) \geq 0.5 \cdot V_i^{\pi^\star}(s_x \mid \delta_{-i}) + 0.5 \cdot V_i^{\pi^\star}(s_y \mid \delta_{-i}), \tag{6}$$

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

where 0.5 is the probability in the initial distribution. It is easy to derive the V-values of s_x and s_y under $\pi^\star$ as neither of them depends on the V-values of any other states. We have

$$V_i^{\pi^\star}(s_x \mid \delta_j) = Q_i^{\pi^\star}(s_x, a \mid \delta_j) = \tfrac{1}{1-\gamma_i} \cdot x_j, \tag{7}$$

$$\text{and} \quad V_i^{\pi^\star}(s_y \mid \delta_j) = Q_i^{\pi^\star}(s_y, a \mid \delta_j) = y_j. \tag{8}$$

Plugging these two equations back into (6) gives

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

Replacing γ_i with the corresponding values gives

$$10 \cdot x_1 + y_1 \geq 10 \cdot x_2 + y_2 \tag{9}$$

$$2 \cdot x_2 + y_2 \geq \ 2 \cdot x_1 + y_1 \tag{10}$$

Hence, (9)+(10) gives $x_1 \geq x_2$, and (9)+5×(10) gives $y_2 \geq y_1$.

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

$$Q_i^{\pi^\star}(s_x, a \mid \delta_i) \geq Q_i^{\pi^\star}(s_x, b \mid \delta_i) + \epsilon = 1 + \gamma_i \cdot V_i^{\pi^\star}(s_y \mid \delta_i) + \epsilon$$

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

Substituting (7) and (8) into the above two equations gives:

$$1 + \tfrac{\gamma_i}{1-\gamma_i} \cdot x_i < y_i < \tfrac{1}{\gamma_i(1-\gamma_i)} \cdot x_i - \tfrac{1}{\gamma_i}. \tag{11}$$

Using (5) and (11), we get that

$$9 \cdot x_1 + 1 < y_1 \leq y_2 < 4 \cdot x_2 - 2 \leq 4 \cdot x_1 - 2.$$

This means that $x_1 < 0$ and contradicts the assumption that δ is a non-negative scheme. □

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

Theorem 4.3. *When the agents have the same 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 $H = \frac{2}{1-\gamma} \cdot h + \epsilon$. We construct the following scheme $\delta = (\delta_i)_{i \in [n]}$:

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

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, so it remains to argue that it is feasible. Intuitively, δ_i results in the agent receiving a reward that is sufficiently large (and is roughly the same) at every step if the agent follows $\pi^\star$. Rewards are adjusted by a factor of $1/(1-\gamma)$ at the subsequent terminal states so that it is as if the process continues forever with the same reward H generated at every step (but the cumulative reward $\frac{1}{1-\gamma} \cdot H$ is paid off at once). Therefore, under δ_i, the process is equivalent to an infinite-horizon process where the agent gets a (roughly) constant positive reward H at every step if the agent follows $\pi^\star$. This loss due to not following $\pi^\star$ at some step is H, and it is sufficiently large so that the optimal choice for the agent in such a process is to always follow $\pi^\star$. □

5 Computing an Optimal Fair Solution

In terms of the computation of a cost-minimization fair solution, our main result is as follows. For each of the EF notions we defined above, the set of fair solutions lie in a convex polytope defined by polynomially many linear constraints. Hence, to find out a cost minimizing solution can be formulated as a convex optimization problem given that the cost function (2) is a convex function of the adjustment scheme. We show how the various types of constraints that need to be incorporated can be written as linear constraints next.

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

$$V_i(s) = Q_i(s, \pi^\star(s)) \qquad \text{for all } i, s \tag{12a}$$

$$Q_i(s,a) = R_i(s,a) + \delta_i(s,a) + \gamma_i \sum_{s' \in S} P(s,a,s') \cdot V_i(s') \qquad \text{for all } i, s, a \tag{12b}$$

$$Q_i\left(s, \pi^\star(s)\right) \geq Q_i(s,a) + \epsilon \qquad \text{for all } i, s, a \neq \pi^\star(s) \tag{12c}$$

Specifically, the first two lines follow from the Bellman equation and capture the values $V_i^{\pi^\star}(s \mid \delta_i)$ and $Q_i^{\pi^\star}(s, a \mid \delta_i)$; the last line is the incentive constraints and enforces δ to be feasible.

Next, we consider each of the fairness notions.

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.

$$\delta_i(s,a) = \delta_j(s,a) \qquad \text{for all } s, a \tag{13}$$

WEF Constraints To enforce WEF, we add variables $V_{i,j}$ and $Q_{i,j}$ to capture the values $V_i^{\pi^\star}(s \mid \delta_j)$ and $Q_i^{\pi^\star}(s, a \mid \delta_j)$, i.e., the values agent i would have got had they been offered the adjustment for agent j. Then we add the following constraints, which are similar to the Bellman equation, so that these additional variables acquire the desired values.

$$V_{i,j}(s) = Q_{i,j}(s, \pi^\star(s)) \qquad \text{for all } i, j, s \tag{14a}$$

$$Q_{i,j}(s,a) = R_i(s,a) + \delta_j(s,a) + \gamma_i \sum_{s' \in S} P(s,a,s') \cdot V_{i,j}(s') \qquad \text{for all } i, j, s, a \tag{14b}$$

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):

$$\textstyle\sum_{s \in S} z_s \cdot V_i(s) \geq \sum_{s \in S} z_s \cdot V_{i,j}(s) \qquad \text{for all } s, a \tag{15}$$

EF Constraints Similarly to the approach for handling the WEF constraints, we need additional variables to capture the values of each agent i had they been offered adjustment δ_j. Indeed, we also use the constraints in (14) and (15) but replace (14a) with the following one:

$$V_{i,j}(s) \geq Q_{i,j}(s,a) \qquad \text{for all } i, j, s, a \tag{16a}$$

which associates $V_{i,j}(s)$ to the maximum $Q_{i,j}(s,a)$, instead of $Q_{i,j}(s, \pi^\star(s))$. Note that under these constraints, the value of $V_{i,j}(s)$ in a solution is not necessarily equal to the V-value of s under the optimal policy; it is only an upper bound of them. This will not cause any issue to the approach since the solution is EF if and only if (15) holds for some upper bounds $V_{i,j}(s)$ of $V_i^{\pi^\star}(s \mid \delta_j)$.

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

6 Price of Fairness

We now consider the price of fairness (PoF). The PoF measures the increase of teaching cost due to consideration of fairness. In a similar spirit to the celebrated concept of the price of anarchy (PoA) in game theory, the PoF compares the ratio between the minimum costs with and without fairness constraints. 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

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

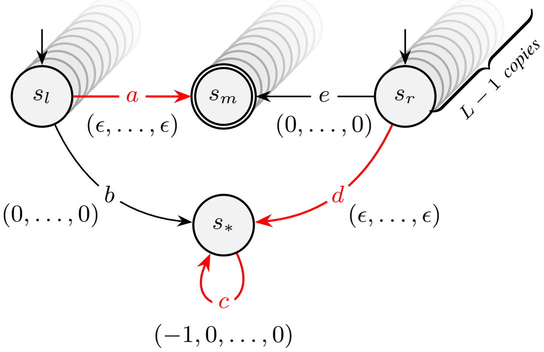

Figure 3: There are n agents with discount factors $\gamma_1 = \cdots = \gamma_n = \gamma$. $A = \{a, b, c, d, e\}$ and all transitions are deterministic. The initial rewards are annotated at the corresponding edges, and they are the same for agents $2, \dots, n$ (agent 1 has a different reward for action c). There are $L-1$ sets of additional copies of s_l, s_m, and s_r. Every copy of s_l and s_r is connected to the copy of s_m in the same set. In addition, copies of s_l and s_r are also connected to s_* (who has no copies). Each new connection has the same initial rewards as its original copy. 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 $\pi^\star(s_*) = c$ (and the same for the corresponding copies).

Namely, the value indicates how large the price can be for instances at the same scale. The PoWEF and PoSEF can be defined in the same way with the corresponding notions.

We analyze the asymptotic growth of the PoF as functions of n, m, and λ. The results are presented in Theorem 6.1 and all the bounds are tight. The PoF increases linearly with λ and sublinearly with the size of the MDP in all the cases, and the PoEF and PoSEF also grows linearly with the number of agents involved.

Theorem 6.1. $\mathrm{PoWEF}(n, m, \lambda) = \Theta(\lambda \cdot \sqrt{m})$*,* $\mathrm{PoEF}(n, m, \lambda) = \Theta(\lambda \cdot n \cdot \sqrt{m})$*, and* $\mathrm{PoSEF}(n, m, \lambda) = \Theta(\lambda \cdot n \cdot \sqrt{m})$*.*

Due to space limit, we leave the detailed proofs of the PoF bounds to the appendix and only provide some intuition about the bounds here. The lower bounds are obtained with the hard instances 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. Hence, it suffices to give agent 1 a bonus of 1 for taking action c, and the overall cost is 1. Now consider the fairness constraints and suppose that we still provide a bonus $\delta_1(s_*, c) = 1$. The consequence is that agents $2, \dots, n$ 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. However, a bonus on c will also incentivize 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 (and only the reward for agent 1 when EF and SEF are considered). Modifying the other rewards is however much more costly since each one of them has $L-1$ copies of themselves, which requires the same modification by symmetry.

To derive the upper bounds, for the PoWEF we construct the following adjustment scheme $\delta = (\delta_i)_{i\in[n]}$ in a similar approach to proving the existence of a fair solution in Theorem 4.1:

$$\delta_i(s, x) = \begin{cases} 0, & \text{if } x = \pi^\star(s) \\ -\frac{2}{1-\gamma_i} \cdot C_i, & \text{otherwise} \end{cases}$$

where C_i denotes the minimum cost for teaching agent i when fairness is not considered. Let $\widehat{\delta}_i$ be the adjustment achieving the minimum cost for each i. It can be easily verified that $\frac{\|\delta_i\|}{\|\widehat{\delta}_i\|} \le 2\lambda \cdot \sqrt{m}$ for all $i \in [n]$, and hence $\mathrm{PoWEF}(n, m, \lambda) \le \frac{\sum_{i\in[n]} \|\delta_i\|}{\sum_{i\in[n]} \|\widehat{\delta}_i\|} = O\left(\lambda \cdot \sqrt{m}\right)$. Moreover, δ is WEF since the scheme only penalizes actions that do not follow the target policy. The argument for showing that δ is feasible is more involved and we leave it to the appendix. A similar approach can be used to derive the upper bounds of the PoEF and the PoSEF, where we penalize actions that do not follow the policy even more, by $-\max_{j\in[n]} \frac{3}{1-\gamma_j} \cdot C_j$. This also leads to a dependency on n in the bound.

	PoWEF	PoEF	PoSEF
No restriction	$\Theta(\lambda \cdot \sqrt{m})$	$\Theta(\lambda \cdot n \cdot \sqrt{m})$	$\Theta(\lambda \cdot n \cdot \sqrt{m})$
Non-neg & identical γ	$\Theta(\lambda \cdot n \cdot \sqrt{m})$	$\Theta(\lambda^2 \cdot n \cdot \sqrt{m})$	$\Theta(\lambda^2 \cdot n \cdot \sqrt{m})$

Table 1: Summary of the PoF.

6.1 PoF with Non-negative Adjustments

We also investigate PoF with non-negative adjustments and compare the costs of the best non-negative adjustment schemes with and without the fairness constraints. Since a feasible and fair solution may not exist with non-negative adjustments, we analyze the case where the agents have the same discount factor. The existence of a feasible fair solution is guaranteed in this case according to Theorem 4.3. The PoF bounds are presented in Theorem 6.2, where the PoWEF now also depend on the number of agents, and the bounds of the PoEF and PoSEF depend quadratically on λ.

Theorem 6.2. *When the scheme is required to be non-negative and all the agents have the same discount factor, it holds that* $\mathrm{PoWEF}(n, m, \lambda) = \Theta(\lambda \cdot n \cdot \sqrt{m})$*,* $\mathrm{PoEF}(n, m, \lambda) = \Theta(\lambda^2 \cdot n \cdot \sqrt{m})$*, and* $\mathrm{PoSEF}(n, m, \lambda) = \Theta(\lambda^2 \cdot n \cdot \sqrt{m})$*.*

The reason why the lower bound of the PoWEF now increases with n can be seen intuitively from the instances in Figure 3, too. Now that the adjustments must be non-negative, to incentivize agent 1 to choose action c, a bonus of at least 1 has to be offered. Accordingly, in order for agent 2 (or any agent $i \geq 2$) to not envy this bonus, additional bonuses must be offered to them as well, resulting in a growth with the number of agents. (Without non-negativity, we can penalize agent 1 instead of offering agents $2, \dots, n$ bonuses to avoid a dependency on n in the lower bound of the PoWEF.) The proofs of the bounds can be found in the appendix and all the PoF bounds are summarized in Table 1.

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 or cost measures can be considered. For example, one can use the cumulative payment of the teacher as the cost measure. Indeed, since the cumulative payment is a linear function of the adjustments, the same computation approach we presented applies by replacing the objective function, whereby we obtain a linear program. In terms of the PoF bounds, in a previous version of this work we conjectured that similar bounds can be derived with the cumulative payment cost measure, but it turns out the PoF might also depend on other factors such as the initial state distribution. A detailed analysis of the bounds is an interesting direction for future work.

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 considers fairness when launching a poisoning attack. 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.

Acknowledgments and Disclosure of Funding

The authors thank Warut Suksompong for useful suggestions on the related work and the anonymous reviewers for their insightful comments.

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 *International Conference on Learning Representations (ICLR '18)*, 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] Francis Edward Su. Rental harmony: Sperner's lemma in fair division. *The American mathematical monthly*, 106(10):930–942, 1999.

[5] Jiarui Gan, Warut Suksompong, and Alexandros A Voudouris. Envy-freeness in house allocation problems. *Mathematical Social Sciences*, 101:104–106, 2019.

[6] Elias Koutsoupias and Christos Papadimitriou. Worst-case equilibria. In *Annual symposium on theoretical aspects of computer science (STACS'99)*, pages 404–413. Springer, 1999.

[7] Haoqi Zhang and David C. Parkes. Value-based policy teaching with active indirect elicitation. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI '08)*, pages 208–214, 2008.

[8] Haoqi Zhang, David C Parkes, and Yiling Chen. Policy teaching through reward function learning. In *Proceedings 10th ACM Conference on Electronic Commerce (EC '09)*, pages 295–304, 2009.

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

[10] Yuzhe Ma, Xuezhou Zhang, Wen Sun, and Jerry Zhu. Policy poisoning in batch reinforcement learning and control. In *Advances in Neural Information Processing Systems (NeurIPS '19)*, pages 14543–14553, 2019.

[11] Yunhan Huang and Quanyan Zhu. Deceptive reinforcement learning under adversarial manipulations on cost signals. In *International Conference on Decision and Game Theory for Security (GameSec '19)*, pages 217–237, 2019.

[12] 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.

[13] Yanchao Sun, Da Huo, and Furong Huang. Vulnerability-aware poisoning mechanism for online rl with unknown dynamics. In *International Conference on Learning Representations (ICLR '21)*, 2021.

[14] Xuezhou Zhang, Yuzhe Ma, Adish Singla, and Xiaojin Zhu. Adaptive reward-poisoning attacks against reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML '20)*, pages 11225–11234, 2020.

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

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

[17] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning (ICML '99)*, pages 278–287, 1999.

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

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

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

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

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

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

[24] 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.

[25] 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.

[26] 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.

[27] Ritesh Noothigattu, Tom Yan, and Ariel D. Procaccia. Inverse reinforcement learning from like-minded teachers. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI '21)*, pages 9197–9204, 2021.

[28] 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]

A Existence of Fair Solutions

Theorem 4.1. *For any robustness guarantee $\epsilon > 0$, an SEF and feasible adjustment scheme always exists.*[2]

Proof. 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

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

for all $s \in S$ and $i \in [n]$. The scheme is SEF as δ_i is the same for all the agents.

To see that it is also feasible, observe that by following the target policy $\pi^\star$, an agent obtains reward at least $-h$ in every step. Hence, for all $s \in S$ and all $a \neq \pi^\star(s)$, we have

$$Q_i^{\pi^\star}(s, \pi^\star(s) \mid \delta_i) \geq -\frac{h}{1-\gamma_i} \geq -\max_{i' \in [n]} \frac{h}{1-\gamma_{i'}}.$$

It then follows that

$$\begin{aligned} Q_i^{\pi^\star}(s, \pi^\star(s) \mid \delta_i) &\geq \delta_i(s,a) + \frac{h}{1-\gamma_i} + \epsilon \\ &\geq \delta_i(s,a) + \gamma_i \cdot \sum_{s' \in S} P(s,a,s') \cdot V_i^{\pi^\star}(s' \mid \delta_i) + \epsilon \\ &= Q_i^{\pi^\star}(s, a \mid \delta_i) + \epsilon, \end{aligned}$$

where we used the fact that $V_i^{\pi^\star}(s' \mid \delta_i) \leq \frac{h}{1-\gamma_i}$ for all s', which is due to the fact that the reward obtained at every step is at most h. □

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

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

$$\delta_i(s,a) = \begin{cases} H + \frac{\gamma}{1-\gamma} \cdot H \cdot \sum_{s' \in S^{\mathrm{T}}} P(s,a,s'), & \text{if } a = \pi^\star(s) \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

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. We show that it is also feasible.

Consider an arbitrary agent i. We first argue that $V_i^{\pi^\star}(s \mid \delta_i) \in \left[\frac{H-h}{1-\gamma}, \frac{H+h}{1-\gamma}\right]$ for all $s \in S \setminus S^{\mathrm{T}}$. Indeed, if the original reward function R_i was a zero function ($R_i(s,a) = 0$), it can be easily verified that the solution to the Bellman equation would be: $V_i^{\pi^\star}(s \mid \delta_i) = \frac{H}{1-\gamma}$ for all $s \in S \setminus S^{\mathrm{T}}$ and $V_i^{\pi^\star}(s \mid \delta_i) = 0$ for all $s \in S^{\mathrm{T}}$. Now the original reward $R_i(s,a)$ is bounded in $[-h, h]$, which means an additional reward in this range in each step and, hence, an additional cumulative reward in the interval $\left[\frac{-h}{1-\gamma}, \frac{h}{1-\gamma}\right]$. Adding this to $\frac{H}{1-\gamma}$ gives the desired range $\left[\frac{H-h}{1-\gamma}, \frac{H+h}{1-\gamma}\right]$.

Hence, $V_i^{\pi^\star}(s \mid \delta_i) \in \left[\frac{H-h}{1-\gamma}, \frac{H+h}{1-\gamma}\right]$ for all $s \in S$. This further implies that, for any actions $a, b \in A$, it holds that

$$\sum_{s' \in S} P(s,a,s') \cdot V_i^{\pi^\star}(s' \mid \delta_i) \geq \sum_{s' \in S} P(s,b,s') \cdot V_i^{\pi^\star}(s' \mid \delta_i) - \frac{2h}{1-\gamma}. \tag{18}$$

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

We have

$$\begin{aligned} Q_i^{\pi^\star}(s,\pi^\star(s)\mid\delta_i) &= R_i(s,\pi^\star(s)) + \delta_i(s,\pi^\star(s)) + \gamma\cdot\sum_{s'\in S} P(s,\pi^\star(s),s')\cdot V_i^{\pi^\star}(s'\mid\delta_i) \\ &\geq -h + H + \gamma\cdot\sum_{s'\in S} P(s,\pi^\star(s),s')\cdot V_i^{\pi^\star}(s'\mid\delta_i) \\ &\geq h + \epsilon + \gamma\cdot\sum_{s'\in S} P(s,a,s')\cdot V_i^{\pi^\star}(s'\mid\delta_i) \end{aligned}$$

for any $a \in A$, where the last line follows by (18) and the fact that $H = \frac{2\gamma}{1-\gamma}\cdot h + 2h + \epsilon$. By definition, we have $\delta_i(s,a) = 0$ for all $a \neq \pi^\star(s)$. It follows that

$$\begin{aligned} Q_i^{\pi^\star}(s,\pi^\star(s)\mid\delta_i) &\geq R_i(s,a) + \delta_i(s,a) + \gamma\cdot\sum_{s'\in S} P(s,a,s')\cdot V_i^{\pi^\star}(s'\mid\delta_i) + \epsilon \\ &= Q_i^{\pi^\star}(s,a\mid\delta_i) + \epsilon. \end{aligned}$$

Therefore, δ is a feasible scheme. □

B PoF Bounds

We analyze PoWEF first, and then PoEF and PoSEF.

B.1 PoWEF

To analyze the PoWEF, we first derive its lower bound.

Lemma B.1. $\mathrm{PoWEF}(n,m,\lambda) = \Omega(\lambda\cdot\sqrt{m})$.

Proof. Consider the family of instances illustrated in Figure 3, and we consider the two-agent version of this example ($n = 2$) that consists of only agents 1 and 2. We show that the PoWEF of this particular family of instances is $\Omega(\lambda\cdot\sqrt{|S|\cdot|A|})$ to establish the lower bound of PoWEF.

First, the cost of teaching $\pi^\star$ without fairness constraints is at most 1. Indeed, without fairness constraints, $\pi^\star$ is already the optimal policy of agent 2 up to a robustness of ϵ. As for agent 1, it suffices to set $\delta_1(c) = 1$. Hence, the total cost is 1.

Now consider the case with fairness constraints and suppose that $\delta = (\delta_1,\delta_2)$ is a WEF and feasible adjustment scheme. We argue that $\|\delta_1\| + \|\delta_2\| = \Omega(\lambda\cdot\sqrt{|S|\cdot|A|})$.

By symmetry, we can assume without loss of generality that each δ_i assigns the same reward for a state-action pair and its copies in the instance. Hence, in our analysis, it suffices to consider only the values associated with the original state-action pairs, which are representative of the values associated with their copies. Given this, we omit the state in the notation and write, e.g., $\delta_i(a) = \delta_i(s_l,a)$, as each action is associated with a unique state.

Consider the following two cases:

Case 1: $\delta_1(c) \leq 1/2$. Since δ_1 incentivizes agent 1 to use the target policy $\pi^\star$, we have $Q_1^{\pi^\star}(s_r,d) \geq Q_1^{\pi^\star}(s_r,e) + \epsilon$, or equivalently,

$$\delta_1(d) + \epsilon + \frac{\gamma}{1-\gamma}\cdot(\delta_1(c) - 1) \geq \delta_1(e) + \epsilon.$$

Rearranging the terms gives

$$\delta_1(e) - \delta_1(d) \leq \frac{\gamma}{1-\gamma}\cdot(\delta_1(c) - 1) \leq -\frac{1}{2}\cdot\frac{\gamma}{1-\gamma}.$$

Note that for any two real numbers x and y, we have $x^2 + y^2 \geq \frac{(x-y)^2}{2}$. Hence,

$$\begin{aligned} \|\delta_1\| \geq \sqrt{L}\cdot\sqrt{\delta_1^2(e) + \delta_1^2(d)} &\geq \sqrt{L}\cdot\sqrt{\frac{(\delta_1(e) - \delta_1(d))^2}{2}} \\ &\geq \sqrt{L}\cdot\frac{1}{\sqrt{8}}\cdot\frac{\gamma}{1-\gamma} = \Omega(\lambda\cdot\sqrt{|S|\cdot|A|}). \end{aligned}$$

Case 2: $\delta_1(c) \geq 1/2$. By WEF, we have $\rho_1^{\pi^\star}(\delta_1) \geq \rho_1^{\pi^\star}(\delta_2)$ and $\rho_2^{\pi^\star}(\delta_2) \geq \rho_2^{\pi^\star}(\delta_1)$. Let $\varrho_i^{\pi^\star}(\delta_j) = \rho_i^{\pi^\star}(\delta_j) - \rho_i^{\pi^\star}(0)$, where $\rho_i^{\pi^\star}(0)$ denotes the agent's cumulative reward without any adjustment. Since now both agents 1 and 2 have the same discount factor γ, we have

$$\varrho_1^{\pi^\star}(\delta_j) = \varrho_2^{\pi^\star}(\delta_j)$$

for any j. Hence,

$$\begin{aligned} &\rho_1^{\pi^\star}(\delta_1) \geq \rho_1^{\pi^\star}(\delta_2) \quad \Longrightarrow \quad \varrho_1^{\pi^\star}(\delta_1) \geq \varrho_1^{\pi^\star}(\delta_2) = \varrho_2^{\pi^\star}(\delta_2), \\ \text{and} \quad &\rho_2^{\pi^\star}(\delta_2) \geq \rho_2^{\pi^\star}(\delta_1) \quad \Longrightarrow \quad \varrho_2^{\pi^\star}(\delta_2) \geq \varrho_2^{\pi^\star}(\delta_1) = \varrho_1^{\pi^\star}(\delta_1), \end{aligned}$$

which means that $\varrho_1^{\pi^\star}(\delta_1) = \varrho_2^{\pi^\star}(\delta_2)$. Expanding this gives

$$\delta_1(a) + \left(\delta_1(d) + \frac{\gamma}{1-\gamma} \cdot \delta_1(c)\right) = \delta_2(a) + \left(\delta_2(d) + \frac{\gamma}{1-\gamma} \cdot \delta_2(c)\right). \tag{19}$$

Moreover, δ_2 incentivizes agent 2 to use the target policy $\pi^\star$, so we have $Q_2^{\pi^\star}(s_l, a) \geq Q_2^{\pi^\star}(s_l, b) + \epsilon$, expanding which gives

$$\delta_2(a) + \epsilon \geq \delta_2(b) + \frac{\gamma}{1-\gamma} \cdot \delta_2(c) + \epsilon.$$

Combining (19) with the above equation gives

$$2 \cdot \delta_2(a) - \delta_2(b) + \delta_2(d) - \delta_1(a) - \delta_1(d) \geq \frac{\gamma}{1-\gamma} \cdot \delta_1(c) \geq \frac{1}{2} \cdot \frac{\gamma}{1-\gamma}.$$

Note that for any real numbers $x_1, \ldots, x_k$ and nonzero coefficients $a_1, \ldots, a_k$, we have $\sum_{i=1}^k x_i^2 \geq \left(\sum_{i=1}^k a_i \cdot x_i\right)^2 / \sum_{i=1}^k a_i^2$. It follows that

$$\begin{aligned} \|\delta_1\| + \|\delta_2\| &\geq \sqrt{L} \cdot \sqrt{\delta_2^2(a) + \delta_2^2(b) + \delta_2^2(d) + \delta_1^2(a) + \delta_1^2(d)} \\ &\geq \sqrt{L} \cdot \frac{1}{\sqrt{32}} \cdot \frac{\gamma}{1-\gamma} \\ &= \Omega(\lambda \cdot \sqrt{|S| \cdot |A|}). \end{aligned}$$

Therefore, in both cases, we have $\|\delta_1\| + \|\delta_2\| = \Omega(\lambda \cdot \sqrt{|S| \cdot |A|})$, which completes the proof. □

Lemma B.2. $\mathrm{PoWEF}(n, m, \lambda) = O(\lambda \cdot \sqrt{m})$.

Proof. Suppose that without the fairness constraints the minimum costs for teaching $\pi^\star$ is C_i for each agent $i \in [n]$; let $\widehat{\delta}_i$ be the adjustment achieving this minimum cost for each $i \in [n]$, and let $\widehat{\delta} = \left(\widehat{\delta}_i\right)_{i \in [n]}$. Hence, $\left|\widehat{\delta}_i(s,a)\right| \leq \left\|\widehat{\delta}_i\right\| = C_i$ for all i, s, and a.

We construct the following adjustment scheme $\delta = (\delta_i)_{i \in [n]}$ in an approach similar to that in the proof of Theorem 4.1. We let

$$\delta_i(s,a) = \begin{cases} 0, & \text{if } a = \pi^\star(s) \\ -\frac{2}{1-\gamma_i} \cdot C_i, & \text{otherwise} \end{cases} \tag{20}$$

for all $s \in S$ and $i \in [n]$. With this δ, we have

$$\frac{\|\delta_i\|}{\left\|\widehat{\delta}_i\right\|} = \frac{\sqrt{\sum_{s \in S, a \in A} (\delta_i(s,a))^2}}{C_i} \leq \frac{\sqrt{|S| \cdot |A|} \cdot \frac{2}{1-\gamma_i} \cdot C_i}{C_i} = 2\lambda \cdot \sqrt{|S| \cdot |A|}. \tag{21}$$

Hence, the price of using δ is

$$\frac{\sum_{i \in [n]} \|\delta_i\|}{\sum_{i \in [n]} \left\|\widehat{\delta}_i\right\|} \leq 2\lambda \cdot \sqrt{|S| \cdot |A|} = O\left(\lambda \cdot \sqrt{|S| \cdot |A|}\right).$$

Therefore, it remains to argue that δ is feasible and WEF.

Feasibility Compare the differences in the V-values when $\widehat{\delta}$ and δ are applied. Since $V_i^{\pi^\star}$ only depends on the rewards of state-action pairs chosen by $\pi^\star$, we have

$$\begin{aligned}
\left|V_i^{\pi^\star}(s \mid \delta_i) - V_i^{\pi^\star}\left(s \,\middle|\, \widehat{\delta}_i\right)\right| &= \left|\mathbb{E}\left[\sum_{t=0}^{\infty}(\gamma_i)^t \cdot \left(\delta_i(s_t, \pi^\star(s_t)) - \widehat{\delta}_i(s_t, \pi^\star(s_t))\right) \,\middle|\, s_0 \sim \mathbf{z}, \pi^\star\right]\right| \\
&= \left|\mathbb{E}\left[\sum_{t=0}^{\infty}(\gamma_i)^t \cdot \widehat{\delta}_i(s_t, \pi^\star(s_t)) \,\middle|\, s_0 \sim \mathbf{z}, \pi^\star\right]\right| \\
&\le \left|\sum_{t=0}^{\infty}(\gamma_i)^t \cdot C_i\right| \\
&= \frac{1}{1-\gamma_i} \cdot C_i. \qquad (22)
\end{aligned}$$

Now compare the Q-values. We have

$$\begin{aligned}
&Q_i^{\pi^\star}(s, \pi^\star(s) \,|\, \delta_i) - Q_i^{\pi^\star}\left(s, \pi^\star(s) \,\middle|\, \widehat{\delta}_i\right) \\
&= \delta_i(s, \pi^\star(s)) - \widehat{\delta}_i(s, \pi^\star(s)) + \gamma_i \cdot \mathbb{E}_{x \sim P(s, \pi^\star(s), \cdot)}\left(V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right)\right) \\
&\ge -C_i - \frac{\gamma_i}{1-\gamma_i} \cdot C_i \qquad \text{(by (22))} \\
&= -\frac{1}{1-\gamma_i} \cdot C_i.
\end{aligned}$$

Whereas for any $a \ne \pi^\star(s)$,

$$\begin{aligned}
Q_i^{\pi^\star}(s, a \,|\, \delta_i) - Q_i^{\pi^\star}\left(s, a \,\middle|\, \widehat{\delta}_i\right) &= \delta_i(s, a) - \widehat{\delta}_i(s, a) + \gamma_i \cdot \mathbb{E}_{x \sim P(s, a, \cdot)}\left(V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right)\right) \\
&\le \delta_i(s, a) - \widehat{\delta}_i(s, a) + \frac{\gamma_i}{1-\gamma_i} \cdot C_i \qquad \text{(by (22))} \\
&\le -\frac{2}{1-\gamma_i} \cdot C_i + C_i + \frac{\gamma_i}{1-\gamma_i} \cdot C_i \qquad \text{(by (20) and } \left\|\widehat{\delta}_i\right\| = C_i\text{)} \\
&= -\frac{1}{1-\gamma_i} \cdot C_i
\end{aligned}$$

Combining the above two equations gives

$$Q_i^{\pi^\star}(s, \pi^\star(s) \,|\, \delta_i) - Q_i^{\pi^\star}(s, a \,|\, \delta_i) \ge Q_i^{\pi^\star}\left(s, \pi^\star(s) \,\middle|\, \widehat{\delta}_i\right) - Q_i^{\pi^\star}\left(s, a \,\middle|\, \widehat{\delta}_i\right)$$

for any $s \in S$ and $a \ne \pi^\star(s)$. Indeed, since $\widehat{\delta}$ is feasible, by definition we have

$$Q_i^{\pi^\star}\left(s, \pi^\star(s) \,\middle|\, \widehat{\delta}_i\right) \ge Q_i^{\pi^\star}\left(s, a \,\middle|\, \widehat{\delta}_i\right) + \epsilon$$

if $a \ne \pi^\star(s)$. It then follows that

$$Q_i^{\pi^\star}(s, \pi^\star(s) \,|\, \delta_i) - Q_i^{\pi^\star}(s, a \,|\, \delta_i) \ge \epsilon$$

for all $a \ne \pi^\star(s)$. Since the choice of i is arbitrary, by definition δ is feasible.

Fairness Indeed, since δ offers no additional reward for state-action pairs specified by the target policy $\pi^\star$, we have $\rho_i^{\pi^\star}(\delta_i) = \rho_i^{\pi^\star}(0) = \rho_i^{\pi^\star}(\delta_j)$ for all $i, j \in [n]$. Hence, δ is WEF. □

B.2 PoEF and PoSEF

Next we turn to PoEF and PoSEF.

Lemma B.3. $\mathrm{PoEF}(n, m, \lambda) = \Omega(\lambda \cdot n \cdot \sqrt{m})$.

Proof. We use the class of instances illustrated in Figure 3. Similarly to the two-agent version of the instances we used in the proof of Lemma B.1, the cost of teaching $\pi^\star$ without fairness constraints is at most 1. It suffices to set $\delta_1(s_*, c) = 1$ for agent 1, and keep the reward functions of all other agents as is since $\pi^\star$ is already optimal for agents $2, \dots, n$ up to robustness ϵ.

Now consider the case with fairness constraints. Suppose that $\delta = (\delta_1, \dots, \delta_n)$ is an EF and feasible adjustment scheme, and without loss of generality $\delta_2 = \dots = \delta_n$. We argue that $\sum_{i \in [n]} \|\delta_i\| = \Omega(\lambda \cdot n \cdot \sqrt{|S| \cdot |A|})$ to complete the proof.

Similarly to the argument in the proof of Lemma B.1, by symmetry we can assume without loss of generality that each δ_i assigns the same reward for a state-action pair and its copy, so we omit the state in the notation of δ_i and write, e.g., $\delta_i(a) = \delta_i(s_l, a)$, as each action is associated with a unique state that is not a copy.

Consider the following two cases:[3]

Case 1: $\delta_2(c) \geq 1/2$**.** Since δ_2 incentivizes agent 2 to use the target policy $\pi^\star$, we have $Q_2^{\pi^\star}(s_l, a) \geq Q_2^{\pi^\star}(s_l, b) + \epsilon$, or equivalently,

$$\delta_2(a) + \epsilon \geq \delta_2(b) + \frac{\gamma}{1-\gamma} \cdot \delta_2(c) + \epsilon.$$

Rearranging the terms gives

$$\delta_2(a) - \delta_2(b) \geq \frac{\gamma}{1-\gamma} \cdot \delta_2(c) \geq \frac{1}{2} \cdot \frac{\gamma}{1-\gamma}.$$

For any real numbers x and y, we have $x^2 + y^2 \geq \frac{(x-y)^2}{2}$. Hence,

$$\begin{aligned} \|\delta_2\| \geq \sqrt{L} \cdot \sqrt{\delta_2^2(a) + \delta_2^2(b)} &\geq \sqrt{L} \cdot \sqrt{\frac{(\delta_2(a) - \delta_2(b))^2}{2}} \\ &\geq \sqrt{L} \cdot \frac{1}{\sqrt{8}} \cdot \frac{\gamma}{1-\gamma} = \Omega(\lambda \cdot \sqrt{|S| \cdot |A|}). \end{aligned}$$

Case 2: $\delta_2(c) \leq 1/2$**.** By EF, we have $\rho_1^{\pi^\star}(\delta_1) \geq \rho_1^{\pi^\star}(\delta_2)$ and $\rho_2^{\pi^\star}(\delta_2) \geq \rho_2^{\pi^\star}(\delta_1)$. The same as the proof of Lemma B.1, since the agents have the same discount factor, we have $\rho_1^{\pi^\star}(\delta_1) - \rho_1^{\pi^\star}(0) = \rho_2^{\pi^\star}(\delta_2) - \rho_1^{\pi^\star}(0)$, expanding which gives the following equation (the same as (19)).

$$\delta_1(a) + \left(\delta_1(d) + \frac{\gamma}{1-\gamma} \cdot \delta_1(c)\right) = \delta_2(a) + \left(\delta_2(d) + \frac{\gamma}{1-\gamma} \cdot \delta_2(c)\right). \tag{23}$$

Now by EF, agent 1 would not be better off if they were given δ_2 and deviated to a policy π with $\pi(s_l) = a$ and $\pi(s_r) = e$. Namely, $\rho_1^{\pi^*}(\delta_1) \geq \rho_1^{\pi}(\delta_2)$, or equivalently

$$\delta_1(a) + \delta_1(d) + \frac{\gamma}{1-\gamma} \cdot (\delta_1(c) - 1) \geq \delta_2(a) + \delta_2(e).$$

Combining (23) with the above equation gives

$$\delta_2(d) - \delta_2(e) \geq \frac{\gamma}{1-\gamma} \cdot (1 - \delta_2(c)) \geq \frac{1}{2} \cdot \frac{\gamma}{1-\gamma}.$$

For any real numbers x and y, we have $x^2 + y^2 \geq \frac{(x-y)^2}{2}$. It follows that

$$\begin{aligned} \|\delta_2\| &\geq \sqrt{L} \cdot \sqrt{\delta_2^2(d) + \delta_2^2(e)} \\ &\geq \sqrt{L} \cdot \frac{1}{\sqrt{8}} \cdot \frac{\gamma}{1-\gamma} = \Omega(\lambda \cdot \sqrt{|S| \cdot |A|}). \end{aligned}$$

[3] The analysis of these two cases are similar to the analysis in the proof of Lemma B.1, but with a few differences. In particular, we focus on the adjustment for agent 2 in this proof and aim to show that $\|\delta_2\| = \Omega(\lambda \cdot \sqrt{|S| \cdot |A|})$ for both cases, whereas when WEF is considered we can only bound $\|\delta_1\|$ or $\|\delta_1\| + \|\delta_2\|$ in the proof of Lemma B.1.

Therefore, in both cases, we have $\|\delta_2\| = \Omega(\lambda \cdot \sqrt{|S| \cdot |A|})$. Since $\delta_2 = \delta_3 = \cdots = \delta_n$, we have

$$\mathrm{cost}(\delta) \geq \sum_{i=2}^{n} \|\delta_i\| = \Omega(\lambda \cdot n \cdot \sqrt{|S| \cdot |A|}),$$

which completes the proof. □

Lemma B.4. $\mathrm{PoSEF}(n, m, \lambda) = O(\lambda \cdot n \cdot \sqrt{m})$.

Proof. The proof is similar to the proof of Lemma B.2. We penalize actions off the policy and let

$$\delta_i(s, a) = \begin{cases} 0, & \text{if } a = \pi^\star(s) \\ -\max_{j \in [n]} \frac{3}{1-\gamma_j} \cdot C_j, & \text{otherwise} \end{cases}$$

for all $s \in S$ and $i \in [n]$. Hence, δ is SEF as all δ_i's are the same.

Similarly to (21), with this adjustment scheme δ, we now have

$$\frac{\|\delta_i\|}{\max_{j \in [n]} \left\|\widehat{\delta}_j\right\|} = \frac{\sqrt{\sum_{s \in S, a \in A} (\delta_i(s, a))^2}}{\max_{j \in [n]} C_j} \leq 3\lambda \cdot \sqrt{|S| \cdot |A|}.$$

Hence, the price of using δ is

$$\frac{\sum_{i \in [n]} \|\delta_i\|}{\sum_{i \in [n]} \left\|\widehat{\delta}_i\right\|} \leq \frac{\sum_{i \in [n]} \|\delta_i\|}{\max_{i \in [n]} \left\|\widehat{\delta}_i\right\|} \leq n \cdot 3\lambda \cdot \sqrt{|S| \cdot |A|} = O\left(\lambda \cdot n \cdot \sqrt{m}\right).$$

The feasibility of δ follows by the same argument in the proof of Lemma B.2. □

Summarizing the above lemmas, we get the following main theorem.

Theorem 6.1. $\mathrm{PoWEF}(n, m, \lambda) = \Theta(\lambda \cdot \sqrt{m})$, $\mathrm{PoEF}(n, m, \lambda) = \Theta(\lambda \cdot n \cdot \sqrt{m})$, *and* $\mathrm{PoSEF}(n, m, \lambda) = \Theta(\lambda \cdot n \cdot \sqrt{m})$.

Proof. The bound of the PoWEF follows by the lower and upper bounds established in Lemmas B.1 and B.2.

Since SEF is a stronger requirement than EF, the bounds of the PoEF and PoSEF follow by Lemmas B.3 and B.4. □

C PoF Bounds with Non-negativity

Since a feasible and fair solution may not exist with non-negative adjustments, we analyze the case where the agents have the same discount factor. The existence of a feasible fair solution is guaranteed in this case according to Theorem 4.3.

C.1 PoWEF

Lemma C.1. $\mathrm{PoWEF}(n, m, \lambda) = \Omega(\lambda \cdot n \cdot \sqrt{m})$ *when the scheme is required to be non-negative and all the agents have the same discount factor.*

Proof. Consider the family of instances illustrated in Figure 4. We show that the PoWEF of this particular family of instances is $\Omega(\lambda \cdot n \cdot \sqrt{m})$ to establish the lower bound.

First, the cost of teaching $\pi^\star$ without fairness constraints is at most 1: the target policy $\pi^\star$ is already optimal for agent 2, and it suffices to set $\delta_1(s_r, c) = 1$ to incentivize agent 1.

Now consider the case with fairness constraints and suppose that $\delta = (\delta_1, \ldots, \delta_n)$ is a WEF and feasible adjustment scheme. Without loss of generality, we can assume that $\delta_2 = \delta_3 = \cdots = \delta_n$, and we argue that $\|\delta_2\| = \Omega(\lambda \cdot \sqrt{m})$ to finish the proof.

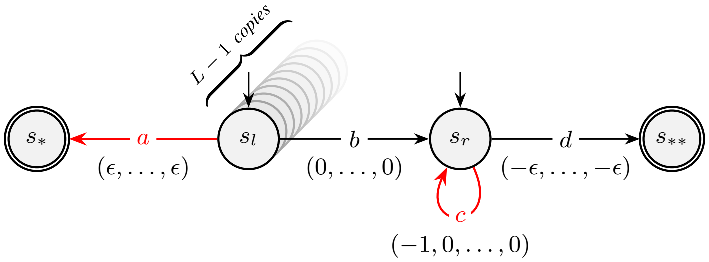

Figure 4: There are n agents, all with discount factor γ. $A = \{a, b, c, d\}$ and all transitions are deterministic. The initial rewards are annotated on the corresponding edges, and they are identical for agents $2, \ldots, n$. There are $L-1$ copies of s_l, each connected to s_* and s_r the same way s_l is connected to these two states (and with the same initial rewards). The initial state distribution has probability $0.5/L$ on s_l as well as each of its copies, and 0.5 on s_r. The target policy is highlighted in red: $\pi^\star(s) = a$ for $s = s_l$ and its copies, and $\pi^\star(s_r) = c$.

By symmetry, we can assume without loss of generality that each δ_i assigns the same reward for a state-action pair and its copies in the instance. Hence, it suffices to consider only the values associated with the original state-action pairs, and we omit the state in the notation and write, e.g., $\delta_i(a) = \delta_i(s_l, a)$, as each action is associated with a unique state.

Consider the following two cases.

Case 1: $\delta_2(c) \geq 1/2$. Since δ_2 incentivizes agent 2 to use the target policy $\pi^\star$, we have $Q_2^{\pi^\star}(s_l, a) \geq Q_2^{\pi^\star}(s_l, b) + \epsilon$, or equivalently,

$$\delta_2(a) + \epsilon \geq \delta_2(b) + \frac{\gamma}{1-\gamma} \cdot \delta_2(c) + \epsilon.$$

Since δ_2 is non-negative and by assumption $\delta_2(c) \geq 1/2$ in this case, we get that $\delta_2(a) \geq \frac{1}{2} \cdot \frac{\gamma}{1-\gamma}$. By symmetry this also holds for all copies of action a. It follows that

$$\|\delta_2\| \geq \frac{\sqrt{L}}{2} \cdot \frac{\gamma}{1-\gamma} = \Omega(\lambda\sqrt{m}).$$

Case 2: $\delta_2(c) \leq 1/2$. Note that since δ_1 is non-negative and it incentivizes agent 1 to select action c, it must be that $\delta_1(c) \geq 1$. By WEF, we have $\rho_2^{\pi^\star}(\delta_2) \geq \rho_2^{\pi^\star}(\delta_1)$, which means

$$0.5 \cdot (\epsilon + \delta_2(a)) + 0.5 \cdot \frac{1}{1-\gamma} \cdot \delta_2(c) \geq 0.5 \cdot (\epsilon + \delta_1(a)) + 0.5 \cdot \frac{1}{1-\gamma} \cdot \delta_1(c).$$

Rearranging the terms and using the facts that $\delta_1(c) \geq 1$ and all adjustments are non-negative, we get that $\delta(a) \geq \frac{1}{2} \cdot \frac{1}{1-\gamma}$ and

$$\|\delta_2\| \geq \frac{\sqrt{L}}{2} \cdot \frac{1}{1-\gamma} = \Omega(\lambda\sqrt{m}).$$

Therefore, in both cases, $\|\delta_2\| = \Omega(\lambda \cdot \sqrt{m})$. Since $\delta_2 = \delta_3 = \cdots = \delta_n$, we have $\mathrm{cost}(\delta) \geq \sum_{i=2}^{n} \|\delta_i\| = \Omega(\lambda \cdot n \cdot \sqrt{m})$, which completes the proof. □

Lemma C.2. $\mathrm{PoWEF}(n, m, \lambda) = O(\lambda \cdot n \cdot \sqrt{m})$ *when the scheme is required to be non-negative and all the agents have the same discount factor.*

Proof. Suppose that without the fairness constraints the minimum costs for teaching $\pi^\star$ is C_i for each agent $i \in [n]$; let $\widehat{\delta}_i$ be the adjustment achieving this minimum cost for each $i \in [n]$, and let $\widehat{\delta} = \left(\widehat{\delta}_i\right)_{i \in [n]}$. Hence, $\left|\widehat{\delta}_i(s, x)\right| \leq \left\|\widehat{\delta}_i\right\| = C_i$ for all i, s, and x.

Note that since the agents have the same discount factor, the improvement $\varrho^{\pi^\star}$ of the cumulative reward is the same for all $i \in [n]$:

$$\varrho^{\pi^\star}\left(\widehat{\delta}_j\right) := \rho_i^{\pi^\star}\left(\widehat{\delta}_j\right) - \rho_i^{\pi^\star}(0).$$

For each $i \in [n]$, we let

$$H_i = (1-\gamma) \cdot \left(\max_{j \in [n]} \varrho^{\pi^\star}\left(\widehat{\delta}_j\right) - \varrho^{\pi^\star}\left(\widehat{\delta}_i\right) \right).$$

Then we construct the following adjustment scheme $\delta = (\delta_i)_{i \in [n]}$:

$$\delta_i(s,a) = \begin{cases} \widehat{\delta}_i(s,a) + H_i + \frac{\gamma}{1-\gamma} \cdot H_i \cdot \sum_{s' \in S^{\mathrm{T}}} P(s,a,s'), & \text{if } a = \pi^\star(s) \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

For any s and a, we have

$$\begin{aligned} \delta_i(s,a) &\le \widehat{\delta}_i(s,a) + \frac{1}{1-\gamma} \cdot H_i \\ &\le \widehat{\delta}_i(s,a) + \max_{j \in [n]} \varrho_i^{\pi^\star}\left(\widehat{\delta}_j\right) \le \frac{2}{1-\gamma} \cdot \max_{j \in [n]} C_j, \end{aligned}$$

where we use $\widehat{\delta}_i(s,a) \le \max_{j \in [n]} C_j$ and $\varrho_i^{\pi^\star}\left(\widehat{\delta}_j\right) \le \frac{1}{1-\gamma} \cdot C_j$, and the latter is due to the fact that the agent gets an additional reward of at most C_j at each time step when $\widehat{\delta}_j$ is applied. It follows that the price of using δ is

$$\frac{\operatorname{cost}(\delta)}{\operatorname{cost}\left(\widehat{\delta}\right)} \le \frac{\sum_{i \in [n]} \|\delta_i\|}{\max_{i \in [n]} \left\|\widehat{\delta}_i\right\|} \le \frac{n \cdot 2\lambda \cdot \max_{i \in [n]} C_i \cdot \sqrt{|S| \cdot |A|}}{\max_{i \in [n]} C_i} = O\left(\lambda \cdot n \cdot \sqrt{m}\right).$$

Therefore, it remains to argue that δ is feasible and WEF.

Now that non-negativity is imposed, we can assume without loss of generality that $\widehat{\delta}_i(s,a) = 0$ for all $s \in S$ and $a \neq \pi^\star(s)$. Therefore, the way δ is defined in (24) is equivalent to adding an additional reward H_i to agent i on top of what is already offered by $\widehat{\delta}_i$. The term $\frac{\gamma}{1-\gamma} \cdot H_i \cdot \sum_{s' \in S^{\mathrm{T}}} P(s,a,s')$ adjusts the reward in consideration of subsequent terminal states, so that it is as if the process continues forever with an additional H_i offered at every subsequent step. Consequently, this improves the V-value of every non-terminal state by $\frac{1}{1-\gamma} \cdot H_i$, i.e., for every $s \in S \setminus S^{\mathrm{T}}$ and every pair $i, j \in [n]$ we have

$$V_i^{\pi^\star}(s \mid \delta_j) = V_i^{\pi^\star}\left(s \,\middle|\, \widehat{\delta}_j\right) + \frac{1}{1-\gamma} \cdot H_i. \tag{25}$$

Feasibility Since the V-values of all non-terminal states increase by the same amount, δ remains feasible. Specifically, since $\widehat{\delta}$ is feasible, we have

$$Q_i^{\pi^\star}\left(s, \pi^\star(s) \,\middle|\, \widehat{\delta}_i\right) \ge Q_i^{\pi^\star}\left(s, a \,\middle|\, \widehat{\delta}_i\right) + \epsilon$$

for all s and $a \neq \pi^\star(s)$. Now compare δ and $\widehat{\delta}$. We have

$$\begin{aligned} & Q_i^{\pi^\star}\left(s, \pi^\star(s) \,|\, \delta_i\right) - Q_i^{\pi^\star}\left(s, \pi^\star(s) \,\middle|\, \widehat{\delta}_i\right) \\ &= \delta_i(s, \pi^\star(s)) - \widehat{\delta}_i(s, \pi^\star(s)) + \gamma \cdot \mathbb{E}_{x \sim P(s, \pi^\star(s), \cdot)} \left(V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right) \right) \\ &= \delta_i(s, \pi^\star(s)) - \widehat{\delta}_i(s, \pi^\star(s)) + \gamma \cdot \sum_{x \in S \setminus S^{\mathrm{T}}} P(s, \pi^\star(s), x) \cdot \left(V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right) \right) \\ &\qquad + \gamma \cdot \sum_{x \in S^{\mathrm{T}}} P(s, \pi^\star(s), x) \cdot \left(V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right) \right) \\ &= H_i + \gamma \cdot \sum_{x \in S \setminus S^{\mathrm{T}}} P(s, \pi^\star(s), x) \cdot \left(V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right) \right) \\ &\qquad + \gamma \cdot \sum_{x \in S^{\mathrm{T}}} P(s, \pi^\star(s), x) \cdot \left(\frac{1}{1-\gamma} \cdot H_i + V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right) \right), \end{aligned}$$

Using (25) and the fact that the V-values of all the terminal states are zero, we further get that

$$\begin{aligned}
&Q_i^{\pi^\star}(s,\pi^\star(s)\mid\delta_i) - Q_i^{\pi^\star}\left(s,\pi^\star(s)\,\middle|\,\widehat{\delta}_i\right)\\
&= H_i + \gamma\cdot\sum_{x\in S\setminus S^{\mathrm{T}}} P(s,\pi^\star(s),x)\cdot\frac{1}{1-\gamma}\cdot H_i + \gamma\cdot\sum_{x\in S^{\mathrm{T}}} P(s,\pi^\star(s),x)\cdot\frac{1}{1-\gamma}\cdot H_i\\
&= \frac{1}{1-\gamma}\cdot H_i.
\end{aligned}$$

Next, consider actions $a \neq \pi^\star(s)$. We have

$$\begin{aligned}
Q_i^{\pi^\star}(s,a\mid\delta_i) - Q_i^{\pi^\star}\left(s,a\,\middle|\,\widehat{\delta}_i\right) &= \delta_i(s,a) - \widehat{\delta}_i(s,a) + \gamma\cdot\mathbb{E}_{x\sim P(s,a,\cdot)}\left(V_i^{\pi^\star}(x\mid\delta_i) - V_i^{\pi^\star}\left(x\,\middle|\,\widehat{\delta}_i\right)\right)\\
&\le \gamma\cdot\mathbb{E}_{x\sim P(s,a,\cdot)}\left(V_i^{\pi^\star}(x\mid\delta_i) - V_i^{\pi^\star}\left(x\,\middle|\,\widehat{\delta}_i\right)\right)\\
&\le \frac{\gamma}{1-\gamma}\cdot H_i.
\end{aligned}$$

It follows that

$$Q_i^{\pi^\star}(s,\pi^\star(s)\mid\delta_i) - Q_i^{\pi^\star}(s,a\mid\delta_i) \ge Q_i^{\pi^\star}\left(s,\pi^\star(s)\,\middle|\,\widehat{\delta}_i\right) - Q_i^{\pi^\star}\left(s,a\,\middle|\,\widehat{\delta}_i\right) \ge \epsilon$$

for any $s \in S$ and $a \neq \pi^\star(s)$. Since the choice of i is arbitrary, δ is feasible.

Fairness By definition $\rho_i^{\pi^\star}(\delta_j) = V_i^{\pi^\star}(\mathbf{z}\mid\delta_j)$, where $\mathbf{z}$ is the initial state distribution. Using (25), we then get that

$$\begin{aligned}
\rho_i^{\pi^\star}(\delta_j) &= \rho_i^{\pi^\star}\left(\widehat{\delta}_j\right) + \frac{1}{1-\gamma}\cdot H_i\\
&= \rho_i^{\pi^\star}\left(\widehat{\delta}_j\right) + \max_{i'\in[n]}\varrho^{\pi^\star}\left(\widehat{\delta}_{i'}\right) - \varrho^{\pi^\star}\left(\widehat{\delta}_i\right)\\
&\le \rho_i^{\pi^\star}\left(\widehat{\delta}_i\right) + \max_{i'\in[n]}\varrho^{\pi^\star}\left(\widehat{\delta}_{i'}\right) - \varrho^{\pi^\star}\left(\widehat{\delta}_i\right) && \text{(as } \widehat{\delta} \text{ is WEF)}\\
&= \rho_i^{\pi^\star}(0) + \max_{i'\in[n]}\varrho^{\pi^\star}\left(\widehat{\delta}_{i'}\right)
\end{aligned}$$

for all $i,j \in [n]$. The right side does not depend on j, which means $\rho_i^{\pi^\star}(\delta_i) = \rho_i^{\pi^\star}(\delta_j)$, for all j, so δ is WEF. □

C.2 PoEF and PoSEF

Lemma C.3. $\mathrm{PoEF}(n,m,\lambda) = \Omega(\lambda^2\cdot n\cdot\sqrt{m})$ *when the scheme is required to be non-negative and all the agents have the same discount factor.*

Proof. Consider the family of instances illustrated in Figure 5. We show that the PoEF of this particular family of instances is $\Omega(\lambda^2\cdot n\cdot\sqrt{m})$ to establish the lower bound.

First, the cost of teaching $\pi^\star$ without fairness constraints is at most 2: the target policy $\pi^\star$ is already optimal for agents $3,\dots,n$, and it suffices to set $\delta_1(s_l,c) = 1$ to incentivize agent 1.

Now consider the case with fairness constraints and suppose that $\delta = (\delta_1,\dots,\delta_n)$ is EF and feasible. Without loss of generality, we can assume that $\delta_3 = \cdots = \delta_n$, and we argue that $\|\delta_2\| = \Omega(\lambda^2\cdot n\cdot\sqrt{m})$ to finish the proof.

By symmetry, we can assume without loss of generality that each δ_i assigns the same reward for a state-action pair and its copies in the instance. Hence, it suffices to consider only the values associated with the original state-action pairs, and we omit the state in the notation and write, e.g., $\delta_i(a) = \delta_i(s_l,a)$, as each action is associated with a unique state.

Observe that the structure of the MDP is symmetric with respect to agents 1 and 2. Hence, without loss of generality, we can also assume the same symmetry in δ:

$$\delta_1(a) = \delta_2(h), \quad \delta_1(h) = \delta_2(a), \quad \delta_1(c) = \delta_2(f), \quad \text{and } \delta_1(f) = \delta_2(c). \tag{26}$$

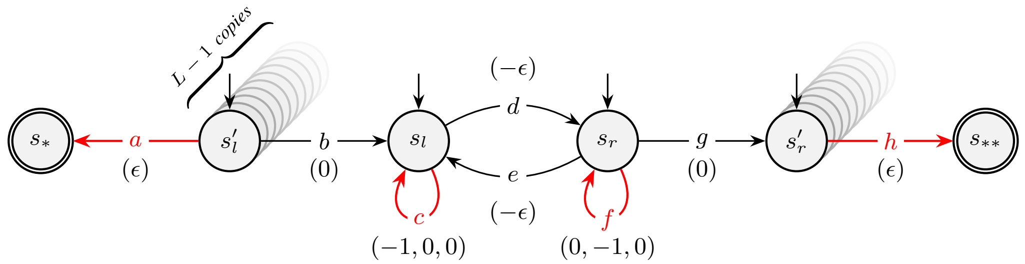

Figure 5: There are n agents, all with discount factor γ. $A = \{a, b, c, \dots, h\}$ and all transitions are deterministic. The initial rewards of agents 1, 2, and 3 are annotated on the corresponding edges (if there is only one number, then all the agents have the same reward). Agents $4, \dots, n$ have the same reward function as agent 3. There are $L-1$ copies of s'_l and s'_r, each connected to the other states the same way s_l and s_r are connected (and with the same initial rewards). The initial state distribution has probability $0.25/L$ on each of s'_l and s'_r as well as each of their copies, and 0.25 on each of s_l and s_r. The target policy is highlighted in red: $\pi^\star(s'_l) = a$, $\pi^\star(s_l) = c$, $\pi^\star(s_r) = f$, and $\pi^\star(s'_r) = h$ (and the same for the corresponding copies).

Next, we first show that $\delta_1(c) \geq \frac{1}{1-\gamma} - \epsilon$ and $\delta_1(f) \geq \frac{1}{1-\gamma} - \epsilon$. Since δ incentivizes agent 1 to take action c instead of d, we have $Q_1^{\pi^\star}(s_l, c \mid \delta_1) \geq Q_1^{\pi^\star}(s_l, d \mid \delta_1) + \epsilon$, expanding which gives

$$\frac{1}{1-\gamma} \cdot (\delta_1(c) - 1) \geq -\epsilon + \frac{\gamma}{1-\gamma} \cdot \delta_1(f) + \epsilon,$$

or

$$\delta_1(c) \geq \gamma \cdot \delta_1(f) + 1. \tag{27}$$

Since δ is EF, agent 1 cannot be better off with the following policy π and δ_2: $\pi(s_l) = d$ and $\pi(s) = \pi^\star(s)$ for all other s. Namely, $\rho_1^\pi(\delta_2) \leq \rho_1^{\pi^\star}(\delta_1)$, or

$$\begin{aligned} &(\delta_2(a) + \epsilon) \quad \overbrace{-\epsilon + \frac{\gamma}{1-\gamma} \cdot \delta_2(f)}^{V_1^\pi(s_l|\delta_2)} \quad + \quad \frac{1}{1-\gamma} \cdot \delta_2(f) \quad + \quad (\delta_2(h) + \epsilon) \\ &\leq (\delta_1(a) + \epsilon) \quad + \quad \frac{1}{1-\gamma} \cdot (\delta_1(c) - 1) \quad + \quad \frac{1}{1-\gamma} \cdot \delta_1(f) \quad + \quad (\delta_1(h) + \epsilon), \end{aligned}$$

where we omit the initial probability 0.25 as the coefficients on both sides of the equation. Applying (26), we can reduce the above equation to

$$1 + \gamma \cdot \delta_1(c) - (1-\gamma) \cdot \epsilon \leq \delta_1(f).$$

Combining (27) with the above equation gives

$$\begin{aligned} & \delta_1(f) \geq \gamma^2 \cdot \delta_1(f) + \gamma + 1 - (1-\gamma) \cdot \epsilon, \\ \Longrightarrow \quad & \delta_1(f) \geq \frac{1}{1-\gamma} - \frac{\epsilon}{1+\gamma} \geq \frac{1}{1-\gamma} - \epsilon; \\ \text{and} \quad & \delta_1(c) \geq \gamma \cdot \delta_1(f) + 1 \geq \frac{1}{1-\gamma} - \epsilon. \end{aligned}$$

The remainder of the proof is then similar to the proof of Lemma C.1 (where we had $\delta_1(c) \geq 1$ but now $\delta_1(c) \geq \frac{1}{1-\gamma} - \epsilon$). We analyze the following three cases.

Case 1: $\delta_3(c) \geq \lambda/2$**.** Since δ_3 incentivizes agent 3 to use the target policy $\pi^\star$, we have $Q_3^{\pi^\star}(s'_l, a) \geq Q_3^{\pi^\star}(s'_l, b) + \epsilon$, or equivalently,

$$\delta_3(a) + \epsilon \geq \delta_3(b) + \frac{\gamma}{1-\gamma} \cdot \delta_3(c) + \epsilon.$$

Since δ_3 is non-negative and by assumption $\delta_3(c) \geq \lambda/2$ in this case, we get that $\delta_3(a) \geq \frac{\lambda}{2} \cdot \frac{\gamma}{1-\gamma}$. By symmetry this also holds for all copies of action a. It follows that

$$\|\delta_3\| \geq \sqrt{L} \cdot \frac{\lambda}{2} \cdot \frac{\gamma}{1-\gamma} = \Omega(\lambda^2 \sqrt{m}).$$

Case 2: $\delta_3(f) \geq \lambda/2$. Applying the same arguments for Case 1 gives $\|\delta_3\| = \Omega(\lambda^2 \sqrt{m})$ in this case.

Case 3: $\delta_3(c) \leq \lambda/2$ and $\delta_3(f) \leq \lambda/2$. We have shown that $\delta_1(c) \geq \frac{1}{1-\gamma} - \epsilon$ and $\delta_1(f) \geq \frac{1}{1-\gamma} - \epsilon$. By WEF, we have $\rho_3^{\pi^\star}(\delta_3) \geq \rho_3^{\pi^\star}(\delta_1)$, which means

$$\begin{aligned}
&(\delta_3(a) + \epsilon) \;+\; \frac{1}{1-\gamma} \cdot \delta_3(c) \;+\; \frac{1}{1-\gamma} \cdot \delta_3(f) \;+\; (\delta_3(h) + \epsilon) \\
\geq\; &(\delta_1(a) + \epsilon) \;+\; \frac{1}{1-\gamma} \cdot \delta_1(c) \;+\; \frac{1}{1-\gamma} \cdot \delta_1(f) \;+\; (\delta_1(h) + \epsilon)\,.
\end{aligned}$$

Rearranging the terms and using non-negativity and the facts that $\delta_1(c) \geq \frac{1}{1-\gamma} - \epsilon$ and $\delta_1(f) \geq \frac{1}{1-\gamma} - \epsilon$, as well as the assumption that $\delta_3(c) \leq \lambda/2$ and $\delta_3(f) \leq \lambda/2$ in this case, we get that

$$\delta_3(a) + \delta_3(h) \geq \left(\frac{1}{1-\gamma}\right)^2 - \frac{2\epsilon}{1-\gamma} = \lambda^2 - 2\epsilon \cdot \lambda.$$

It follows that

$$\|\delta_3\| \geq \sqrt{\frac{L \cdot (\delta_3(a) + \delta_3(h))^2}{2}} = \Omega(\lambda^2 \sqrt{m}).$$

Therefore, in all cases, $\|\delta_3\| = \Omega(\lambda^2 \cdot \sqrt{m})$. Since $\delta_3 = \cdots = \delta_n$, we have $\operatorname{cost}(\delta) \geq \sum_{i=3}^n \|\delta_i\| = \Omega(\lambda^2 \cdot n \cdot \sqrt{m})$, which completes the proof. □

Lemma C.4. $\operatorname{PoSEF}(n, m, \lambda) = O(\lambda^2 \cdot n \cdot \sqrt{m})$ *when the scheme is required to be non-negative and all the agents have the same discount factor.*

Proof. Let $\gamma_1 = \cdots = \gamma_n = \gamma$. Suppose that without the fairness constraints, the minimum costs for teaching $\pi^\star$ is C_i for each agent $i \in [n]$; let $\widehat{\delta}_i$ be the adjustment achieving this minimum cost for each $i \in [n]$, and let $\widehat{\delta} = \left(\widehat{\delta}_i\right)_{i \in [n]}$. Since the schemes are non-negative, we have $0 \leq \widehat{\delta}_i(s, a) \leq C_i$ for all i, s, and a.

Now consider SEF and the following adjustment scheme (similar to (17)), where we let $H = \frac{1}{1-\gamma} \max_{i \in [n]} C_i$ and S^{T} be the set of terminal states.

$$\delta_i(s, a) = \begin{cases} H + \frac{\gamma}{1-\gamma} \cdot H \cdot \sum_{s' \in S^{\mathrm{T}}} P(s, a, s'), & \text{if } a = \pi^\star(s) \\ 0, & \text{otherwise} \end{cases} \tag{28}$$

As defined above, δ is non-negative, and δ_i is identical for all $i \in [n]$, so δ is SEF. Moreover, we have $0 \leq \delta_i(s, a) \leq \frac{1}{1-\gamma} \cdot H$ for all i, s, and a. Hence,

$$\frac{\operatorname{cost}(\delta)}{\operatorname{cost}\left(\widehat{\delta}\right)} \leq \frac{\sum_{i \in [n]} \|\delta_i\|}{\max_{i \in [n]} \left\|\widehat{\delta}_i\right\|} \leq \frac{n \cdot \lambda \cdot H \cdot \sqrt{|S| \cdot |A|}}{\max_{i \in [n]} C_i} = O\left(\lambda^2 \cdot n \cdot \sqrt{m}\right).$$

It remains to argue that δ is also feasible.

Consider an arbitrary agent i. We first argue that

$$V_i^{\pi^\star}(s \mid \delta_i) = V_i^{\pi^\star}(s \mid 0) + \frac{1}{1-\gamma} \cdot H \tag{29}$$

for all $s \in S \setminus S^{\mathrm{T}}$, where $V_i^{\pi^\star}(s \mid 0)$ denotes the original value function when no adjustment is provided. Indeed, since the V-function is additive for two reward functions, it suffices to argue that in a process where the δ_i is the reward function, the corresponding V-values are $\frac{1}{1-\gamma} \cdot H$ for every $s \in S \setminus S^{\mathrm{T}}$. This can be verified via the Bellman equation: The V-values are 0 for all the terminal states, whereas for the non-terminal states, the term $\frac{\gamma}{1-\gamma} \cdot H \cdot \sum_{s' \in S^{\mathrm{T}}} P(s, a, s')$ makes it as if the process continues forever with a reward H generated in every subsequent step, whereby the V-values are exactly $\frac{1}{1-\gamma} \cdot H$. Hence, (29) then follows.

Next consider $\widehat{\delta}$, we have

$$V_i^{\pi^\star}\left(s \,\middle|\, \widehat{\delta}_i\right) = V_i^{\pi^\star}(s) + \mathbb{E}\left[\sum_{t=0}^{\infty} (\gamma_i)^t \cdot \widehat{\delta}_i(s_t, \pi^\star(s_t)) \,\middle|\, s_0 \sim \mathbf{z}, \pi^\star\right].$$

Hence,

$$V_i^{\pi^\star}(s \mid 0) \le V_i^{\pi^\star}\left(s \,\middle|\, \widehat{\delta}_i\right) \le V_i^{\pi^\star}(s \mid 0) + \frac{1}{1-\gamma} \cdot C, \tag{30}$$

where we let $C = \max_{i \in [n]} C_i$. The first inequality follows by the non-negativity of $\widehat{\delta}$, and the second follows by the fact that $\widehat{\delta}_i(s, a) \le C_i \le C$ for all i, s, and a.

Compare the differences in the Q-values when $\widehat{\delta}$ and δ are applied. We have

$$\begin{aligned}
& Q_i^{\pi^\star}(s, \pi^\star(s) \,|\, \delta_i) - Q_i^{\pi^\star}\left(s, \pi^\star(s) \,\middle|\, \widehat{\delta}_i\right) \\
&= \delta_i(s, \pi^\star(s)) - \widehat{\delta}_i(s, \pi^\star(s)) + \gamma \cdot \mathbb{E}_{x \sim P(s, \pi^\star(s), \cdot)}\left(V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right)\right) \\
&= \delta_i(s, \pi^\star(s)) - \widehat{\delta}_i(s, \pi^\star(s)) + \gamma \cdot \sum_{x \in S \setminus S^{\mathrm{T}}} P(s, \pi^\star(s), x) \cdot \left(V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right)\right) \\
&\qquad\qquad + \gamma \cdot \sum_{x \in S^{\mathrm{T}}} P(s, \pi^\star(s), x) \cdot \left(V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right)\right) \\
&= H - \widehat{\delta}_i(s, \pi^\star(s)) + \gamma \cdot \sum_{x \in S \setminus S^{\mathrm{T}}} P(s, \pi^\star(s), x) \cdot \left(V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right)\right) \\
&\qquad\qquad + \gamma \cdot \sum_{x \in S^{\mathrm{T}}} P(s, \pi^\star(s), x) \cdot \left(\frac{1}{1-\gamma} \cdot H + V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right)\right),
\end{aligned}$$

where the last equality follows by replacing $\delta_i(s, \pi^\star(s))$ according to (28). Note that for all terminal states $x \in S^{\mathrm{T}}$, we have $V_i^{\pi^\star}(x \,|\, \delta_i) = V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right) = 0$. Moreover, using (29) and (30), we have $V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right) \ge \frac{1}{1-\gamma} \cdot (H - C)$. Hence, the above equation continues as:

$$\begin{aligned}
& Q_i^{\pi^\star}(s, \pi^\star(s) \,|\, \delta_i) - Q_i^{\pi^\star}\left(s, \pi^\star(s) \,\middle|\, \widehat{\delta}_i\right) \\
&\ge H - \widehat{\delta}_i(s, \pi^\star(s)) + \gamma \sum_{x \in S \setminus S^{\mathrm{T}}} P(s, \pi^\star(s), x) \cdot \frac{1}{1-\gamma} \cdot (H - C) + \gamma \sum_{x \in S^{\mathrm{T}}} P(s, \pi^\star(s), x) \cdot \frac{1}{1-\gamma} \cdot H \\
&\ge H - C + \frac{\gamma \cdot H}{1-\gamma} - \frac{\gamma \cdot C}{1-\gamma} \\
&\ge \frac{\gamma}{1-\gamma} \cdot H.
\end{aligned}$$

Next, we consider actions $a \neq \pi^\star(s)$.

$$\begin{aligned}
Q_i^{\pi^\star}(s, a \,|\, \delta_i) - Q_i^{\pi^\star}\left(s, a \,\middle|\, \widehat{\delta}_i\right) &= \delta_i(s, a) - \widehat{\delta}_i(s, a) + \gamma \cdot \mathbb{E}_{x \sim P(s, a, \cdot)}\left(V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right)\right) \\
&\le \gamma \cdot \mathbb{E}_{x \sim P(s, a, \cdot)}\left(V_i^{\pi^\star}(x \,|\, \delta_i) - V_i^{\pi^\star}\left(x \,\middle|\, \widehat{\delta}_i\right)\right) \\
&\le \frac{\gamma}{1-\gamma} \cdot H,
\end{aligned}$$

where the last transition follows by (28) and (30).

Combining the above two equations gives

$$Q_i^{\pi^\star}\left(s, \pi^\star(s) \,|\, \delta_i\right) - Q_i^{\pi^\star}\left(s, a \,|\, \delta_i\right) \geq Q_i^{\pi^\star}\left(s, \pi^\star(s) \,\middle|\, \widehat{\delta}_i\right) - Q_i^{\pi^\star}\left(s, a \,\middle|\, \widehat{\delta}_i\right)$$

for any $s \in S$ and $a \neq \pi^\star(s)$. Indeed, since $\widehat{\delta}$ is feasible, by definition we have

$$Q_i^{\pi^\star}\left(s, \pi^\star(s) \,\middle|\, \widehat{\delta}_i\right) \geq Q_i^{\pi^\star}\left(s, a \,\middle|\, \widehat{\delta}_i\right) + \epsilon.$$

It then follows that

$$Q_i^{\pi^\star}\left(s, \pi^\star(s) \,|\, \delta_i\right) - Q_i^{\pi^\star}\left(s, a \,|\, \delta_i\right) \geq \epsilon$$

for all $a \neq \pi^\star(s)$. Since the choice of i is arbitrary, δ is feasible. □

Summarizing the above two lemmas, we get the following result.

Theorem 6.2. *When the scheme is required to be non-negative and all the agents have the same discount factor, it holds that* $\mathrm{PoWEF}(n, m, \lambda) = \Theta(\lambda \cdot n \cdot \sqrt{m})$*,* $\mathrm{PoEF}(n, m, \lambda) = \Theta(\lambda^2 \cdot n \cdot \sqrt{m})$*, and* $\mathrm{PoSEF}(n, m, \lambda) = \Theta(\lambda^2 \cdot n \cdot \sqrt{m})$*.*

Proof. Lemmas C.1 and C.2 establish the bound of the PoWEF.

Since SEF is a stronger requirement than EF, Lemmas C.3 and C.4 establish the bounds of the PoEF and PoSEF. □