Multiagent Value Alignment via Inverse Reinforcement Learning

Ritesh Noothigattu*
Machine Learning Department
Carnegie Mellon University

Tom Yan*
Machine Learning Department
Carnegie Mellon University

Ariel D. Procaccia
Computer Science Department
Carnegie Mellon University

Abstract

We study the problem of aligning the values of an AI system with those of multiple agents, represented by different reward functions in a Markov decision process. Motivated by Rawls' *linguistic analogy*, we assume that these reward functions are random perturbations of an underlying reward function, which embodies common moral principles. Under this assumption, we show that inverse reinforcement learning algorithms that satisfy a certain property — that of *matching feature expectations* — recover policies that are approximately optimal with respect to the underlying reward function, and that no algorithm can do better in the worst case. We support this conclusion with experiments involving a classic inverse reinforcement learning algorithm that matches feature expectations.

1 Introduction

A *Markov decision process (MDP)* is a formal specification of a sequential decision making environment, which consists of a set of states, a set of actions, a reward function, and a stochastic transition function. *Reinforcement learning (RL)* deals with learning a *policy* in an MDP — which specifies a possibly randomized action that is taken in each state — to maximize cumulative reward.

RL has long history in AI [24, 11], as well as in many other disciplines. But in recent years interest in the area has exploded, in part due to breakthroughs in game playing [15, 23] and fast-growing applications to robotics [12]. It is safe to say that, nowadays, RL is widely considered to be one of the basic building blocks in the construction of intelligent agents.

While most work in the area focuses on maximizing an exogenous reward function, recently researchers have been thinking about the problem of training an AI system to act in a way that is ethical, beneficial, or safe [22]. One of the prominent approaches is to align the values of the system with the values of a human through *inverse reinforcement learning (IRL)* [16, 1]. Specifically, the idea is to observe a human agent executing a policy in an MDP, where everything is known to the learner except the reward function, and extract a reward function that is most likely to be the one being optimized by the agent (and presumably takes ethical constraints into account). Using this reward function — and knowledge of the other components of the MDP — the agent can easily compute an optimal policy to follow.

Our point of departure is that we are interested in *multi-agent* value alignment rather than alignment with a single agent. Specifically we observe n different agents executing policies that are optimal for their individual reward functions, corresponding to different values. Our approach is to aggregate these observations into a single policy, by applying an inverse reinforcement learning algorithm to

*These authors contributed equally.

the set of all observations. Ideally, the resulting policy would represent moral principles better than any individual policy does.

Naturally, if individuals have wildly divergent values then the aggregate policy may not reflect coherent moral guidelines. For this reason, we assume that individual reward functions are nothing but noisy versions of an underlying reward function, which embodies a common set of moral propositions. This approach is couched in the *linguistic analogy*, originally introduced by Rawls [19]. It draws on the work of Chomsky [5], who argued that competent speakers have a set of grammatical principles in mind, but their linguistic behavior is hampered by "grammatically irrelevant conditions such as memory limitations, distractions, shifts of attention and interest, and errors." Analogously, Rawls claimed, humans have moral rules — a common 'moral grammar' — in our minds, but, due to various limitations, our moral behavior is only an approximation thereof. Interestingly, this theory lends itself to empirical experimentation, and, indeed, it has been validated through the work of Mikhail [14] in moral psychology.

In summary, our research challenge is this: *Given observations from policies that are optimal with respect to different reward functions, each of which is a perturbation of an underlying reward function, identify IRL algorithms that can recover a good policy with respect to the underlying reward function.* It is important to note that, while we are personally motivated by multiagent value alignment and ethical decision making, this technical challenge is both natural and general. Therefore, our work could be equally relevant to other applications.

1.1 Our Model and Results

We start from a common IRL setup: each reward function is associated with a weight vector $\mathbf{w}$, such that the reward for taking a given action in a given state is the dot product of the weight vector and the feature vector of that state-action pair. The twist is that there is an underlying reward function represented by a weight vector $\mathbf{w}^\star$, and each of the agents is associated with a weight vector $\mathbf{w}_i$, which induces an optimal policy π_i. We observe a trajectory from each π_i.

In Section 3, we focus on competing with a uniform mixture over the optimal policies of the agents, $\pi_1, \ldots, \pi_n$ (for reasons that we explicate momentarily). We can do this because the observed trajectories are "similar" to the uniform mixture, in the sense that their feature vectors — the discounted frequencies of the features associated with the observed state-action pairs — are close to that of the uniform mixture policy. Therefore, due to the linearity of the reward function, any policy whose feature expectations approximately match those of the observed trajectories must be close to the uniform mixture with respect to $\mathbf{w}^\star$. We formalize this idea in Theorem 3.2, which gives a lower bound on the number of agents and length of observed trajectories such that any policy that $\epsilon/4$-matches feature expectations is ϵ-close to the uniform mixture. Furthermore, we identify two well-known IRL algorithms, Apprenticeship Learning [1] and Max Entropy [29], which indeed output policies that match the feature expectations of the observed trajectories, and therefore enjoy the guarantees provided by this theorem.

Needless to say, competing with the uniform mixture is only useful insofar as this benchmark exhibits "good" performance. We show that this is indeed the case in Section 4, assuming (as stated earlier) that each weight vector $\mathbf{w}_i$ is a noisy perturbation of $\mathbf{w}^\star$. Specifically, we first establish that, under relatively weak assumptions on the noise, it is possible to bound the difference between the reward of the uniform mixture and that of the optimal policy (Theorem 4.2). More surprisingly, Theorem 4.4 asserts that in the worst case it is impossible to outperform the uniform mixture, by constructing an MDP where the optimal policy cannot be identified — even if we had an infinite number of agents and infinitely long trajectories!

Finally, in Section 5, we run the Apprenticeship Learning algorithm (which matches feature expectations) on trajectories from noisy agents in two different domains, and empirically examine the relation between the level of noise and the reward of the IRL policy.

Putting all of these results together, we conclude that directly running an IRL algorithm that matches feature expectations on the observed trajectories is a sensible approach to multiagent value alignment.

1.2 Related Work

The most closely related work deals with IRL when the observations come from an agent who acts according to multiple *intentions*, each associated with a different reward function [2, 4]. The main challenge stems from the need to cluster the observations — the observations in each cluster are

treated as originating from the same policy (or intention). By contrast, clustering is a nonissue in our framework, and our focus on value alignment with society justifies the assumption that the reward functions are correlated, which, in turn, allows us to provide theoretical guarantees.

Further afield, there is a body of work on robust RL and IRL under reward uncertainty [9, 20, 21], noisy rewards [27], and corrupted rewards [6]. Of these papers the closest to ours is that of Zheng et al. [27], who design robust IRL algorithms under *sparse* noise, in the sense that only a small fraction of the observations are anomalous; they do not provide theoretical guarantees. Our setting is quite different, as very few observations would typically be associated with a near-perfect policy.

From the viewpoint of *computational social choice* [3], our work is related to recent papers on learning and aggregating models of preferences in order to automate ethical decisions [17, 8]. Like Freedman et al. [8], aggregation does not take place explicitly; rather, in our case the IRL algorithm implicitly aggregates individual policies into a single policy.

2 MDP Terminology

We assume the environment is modeled as an MDP $\{S, A, T, \gamma, D\}$ with an unknown reward function. S is a finite set of states; A is a finite set of actions; $T(s, a, s')$ is the state transition probability of reaching state s' from state s when action a is taken; $\gamma \in [0, 1)$ is the discount factor; and D the initial-state distribution, from which the start state s_0 is drawn for every trajectory.

As is standard in the literature [1], we assume that there is a function $\phi : S \times A \to \mathbb{R}^d$ that maps state-action pairs to their real-valued features. We also overload notation, and say that the feature vector of a trajectory $\tau = \{(s_0, a_0), (s_1, a_1), \ldots, (s_L, a_L)\}$ is defined as $\phi(\tau) = \sum_{t=0}^{L} \gamma^t \phi(s_t, a_t)$.

We make the standard assumption that the immediate reward of executing action a from state s is linear in the features of the state-action pair, i.e. $r^{\mathbf{w}}(s, a) = \mathbf{w}^\intercal \phi(s, a)$. In our setting, this has a natural interpretation: ϕ represents the different ethical factors, and $\mathbf{w}$ weighs them in varying degrees.

Let μ denote the feature expectation of policy π, that is, $\mu(\pi) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t)|\pi]$, where π defines the action a_t taken from state s_t, and the expectation is taken over the transition probabilities $T(s_t, a_t, s_{t+1})$. Hence, the cumulative reward of a policy π under weight $\mathbf{w}$ can be rewritten as:

$$R^{\mathbf{w}}(\pi) = \mathbb{E}_{s_0 \sim D}[V^{\pi}(s_0)] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^{\mathbf{w}}(s_t, a_t) \middle| \pi\right] = \mathbf{w}^\intercal \cdot \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a) \middle| \pi\right] = \mathbf{w}^\intercal \mu(\pi)$$

Let $P_\pi(s, t)$ denote the probability of getting to state s at time t under policy π. Then, the cumulative reward $R^{\mathbf{w}}$ is $R^{\mathbf{w}}(\pi) = \sum_{t=0}^{\infty} \gamma^t \sum_{s \in S} P_\pi(s, t) r^{\mathbf{w}}(s, \pi(s))$.

3 Approximating the Uniform Mixture

We consider an environment with n agents $N = \{1, \ldots, n\}$. Furthermore, the reward function of each agent $i \in N$ is associated with a weight vector $\mathbf{w}_i$, and, therefore, with a reward function $r^{\mathbf{w}_i}$. This determines the optimal policy π_i executed by agent i, from which we observe the trajectory τ_i, which consists of L steps. We observe such a trajectory for each $i \in N$, giving us trajectories $\{\tau_1, ..., \tau_n\}$.

As we discussed in Section 1, we assume that the reward function associated with each agent is a noisy version of an underlying reward function. Specifically, we assume that there exists a ground truth weight vector $\mathbf{w}^\star$, and for each agent $i \in N$ we let $\mathbf{w}_i = \mathbf{w}^\star + \boldsymbol{\eta}_i$, where $\boldsymbol{\eta}_i$ is the corresponding noise vector; we assume throughout that $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ are i.i.d. Following Abbeel and Ng [1], we also assume in some of our results (when stated explicitly) that $\|\mathbf{w}^\star\|_2 \leq 1$ and $\|\phi(s, a)\|_\infty \leq 1$.

Let us denote by π^u the *uniform mixture* over the policies $\pi_1, \ldots, \pi_n$, that is, the (randomized) policy that, in each trajectory, selects one of these policies uniformly at random and executes it throughout the trajectory.

Our goal in this section is to "approximate" the uniform mixture (and we will justify this choice in subsequent sections). To do so, we focus on IRL algorithms that "match feature expectations." Informally, the property of interest is that the feature expectations of the policy match the (discounted) feature vectors of observed trajectories. This idea is already present in the IRL literature, but it is

helpful to define it formally, as it allows us to identify specific IRL algorithms that work well in our setting.

Definition 3.1. Given n trajectories $\tau_1, ..., \tau_n$, a (possibly randomized) policy π *ϵ-matches their feature expectations* if and only if $\|\mu(\pi) - \frac{1}{n}\sum_{i=1}^{n}\phi(\tau_i)\|_2 \leq \epsilon$.

In a nutshell, due to the linearity of the reward function, two policies that have the same feature expectations have the same reward. Therefore, if the observed trajectories closely mimic the feature expectations of π_u, and a policy $\tilde{\pi}$ matches the feature expectations of the observed trajectories, then the reward of $\tilde{\pi}$ would be almost identical to that of π^u. This is formalized in the following theorem, whose proof is given in Appendix B.

Theorem 3.2. *Assume that $\|\phi(s,a)\|_\infty \leq 1$ for all $s \in S, a \in A$. Let $\mathbf{w}^\star$ such that $\|\mathbf{w}^\star\|_2 \leq 1$, fix any $\mathbf{w}_1, \ldots, \mathbf{w}_n$, and, for all $i \in N$, let τ_i be a trajectory of length L sampled by executing π_i. Let $\tilde{\pi}$ be a policy that $\epsilon/3-$matches the feature expectation of these trajectories. If*

$$n \geq \frac{72\ln\left(\frac{2}{\delta}\right)d}{\epsilon^2(1-\gamma)^2} \quad \textit{and} \quad L \geq \log_{1/\gamma}\frac{3\sqrt{d}}{(1-\gamma)\epsilon}$$

then, with probability at least $1-\delta$, $\left|R^{\mathbf{w}^\star}(\tilde{\pi}) - R^{\mathbf{w}^\star}(\pi^u)\right| \leq \epsilon$.

Note that the required number of agents n may be significant; fortunately, we can expect access to data from many agents in applications of interest. For example, Noothigattu et al. [17] built a system that decides ethical dilemmas based on data collected from 1.3 million people.

To apply Theorem 3.2, we need to use IRL algorithms that match feature expectations. We have identified two algorithms that satisfy this property: the *Apprenticeship Learning* algorithm of Abbeel and Ng [1], and the *Max Entropy* algorithm of Ziebart et al. [29]. For completeness we present these algorithms, and formally state their feature-matching guarantees, in Appendix A.

4 How Good is the Uniform Mixture?

In Section 3 we showed that it is possible to (essentially) match the performance of the uniform mixture with respect to the ground truth reward function. In this section we justify the idea of competing with the uniform mixture in two ways: first, we show that the uniform mixture approximates the optimal policy under certain assumptions on the noise, and, second, we prove that in the worst case it is actually impossible to outperform the uniform mixture.

4.1 The Uniform Mixture Approximates the Optimal Policy

Recall that for all $i \in n$, $\mathbf{w}_i = \mathbf{w}^\star + \boldsymbol{\eta}_i$. It is clear that without imposing some structure on the noise vectors $\boldsymbol{\eta}_i$, no algorithm would be able to recover a policy that does well with respect to $\mathbf{w}^\star$.

Let us assume, then, that the noise vectors $\boldsymbol{\eta}_i$ are such that the η_{ik} are independent and each η_{ik}^2 is sub-exponential. Formally, a random variable X with mean $u = \mathbb{E}[X]$ is *sub-exponential* if there are non-negative parameters (ν, b) such that $\mathbb{E}\left[\exp\left(\lambda(X-u)\right)\right] \leq \exp\left(\nu^2\lambda^2/2\right)$ for all $|\lambda| < 1/b$. This flexible definition simply means that the moment generating function of the random variable X is bounded by that of a Gaussian in a neighborhood of 0. Note that if a random variable is sub-Gaussian, then its square is sub-exponential. Hence, our assumption is strictly weaker than assuming that each η_{ik} is sub-Gaussian. For our purposes, the key property of sub-exponential random variables is captured by the following well known tail inequality; its proof can be found, for example, in Chapter 2 of [25].

Lemma 4.1. *Let $X_1, \ldots, X_m$ be independent sub-exponential random variables with parameters (ν, b). Then*

$$\Pr\left[\frac{1}{m}\sum_{j=1}^{m}(X_j - u_j) \geq t\right] \leq \begin{cases}\exp\left(-\frac{mt^2}{2\nu^2}\right) & \textit{for } 0 \leq t \leq \frac{\nu^2}{b} \\ \exp\left(-\frac{mt}{2b}\right) & \textit{for } t > \frac{\nu^2}{b}\end{cases},$$

where $u_j = \mathbb{E}[X_j]$.

Despite our assumption about the noise, it is *a priori* unclear that the uniform mixture would do well. The challenge is that the noise operates on the coordinates of the individual weight vectors,

which in turn determine individual rewards, but, at first glance, it seems plausible that relatively small perturbations of rewards would lead to severely suboptimal policies. Our result shows that this is not the case: π^u is approximately optimal with respect to $R^{\mathbf{w}^\star}$, in expectation.

Theorem 4.2. *Assume that $\|\phi(s,a)\|_\infty \leq 1$ for all $s \in S, a \in A$. Let $\mathbf{w}^\star$ such that $\|\mathbf{w}^\star\|_2 \leq 1$, and suppose that $\mathbf{w}_1, ..., \mathbf{w}_n$ are drawn from i.i.d. noise around $\mathbf{w}^\star$, i.e., $\mathbf{w}_i = \mathbf{w}^\star + \boldsymbol{\eta}_i$, where each of its coordinates is such that η_{ik}^2 is an independent sub-exponential random variable with parameters (ν, b). Then*

$$\mathbb{E}[R^{\mathbf{w}^\star}(\pi^u)] \geq R^{\mathbf{w}^\star}(\pi^\star) - O\left(d\sqrt{u} + \nu\sqrt{\frac{d}{u}} + \frac{b}{\sqrt{u}}\right),$$

where $u = \frac{1}{d}\sum_{k=1}^{d} \mathbb{E}\left[\eta_{ik}^2\right]$, and the expectation is taken over the noise.

The exact expression defining the gap between $\mathbb{E}[R^{\mathbf{w}^\star}(\pi^u)]$ and $R^{\mathbf{w}^\star}(\pi^\star)$ can be found in the proof of Theorem 4.2, which appears in Appendix C; we give the asymptotic expression in the theorem's statement because it is easier to interpret. As one might expect, this gap increases as ν or b is increased (and, in a linear fashion). This is intuitive because a smaller ν or b imposes a strictly stronger assumption on the sub-exponential random variable (and its tails).

To gain more insight, we analyze the upper bound on the gap when η_{ik} follows a Gaussian distribution, that is, $\eta_{ik} \sim \mathcal{N}(0, \sigma^2)$. Note that this implies that η_{ik}^2 follows a χ_1^2 distribution scaled by σ^2; a χ_1^2 distributed random variable is known to be sub-exponential with parameters $(2, 4)$, and hence this implies that η_{ik}^2 is sub-exponential with parameters $(2\sigma^2, 4\sigma^2)$. Further, in this case, $u = \mathbb{E}\left[\eta_{ik}^2\right] = \sigma^2$. Plugging these quantities into the upper bound of Theorem 4.2 shows that the gap is bounded by $O(d\sigma)$.

Theorem 4.2 shows that the the gap depends linearly on the number of features d. We informally establish through an example that this upper bound is tight. Assume $\eta_{ik} \sim \mathcal{N}(0, \sigma^2)$ with $\sigma \leq 2/d$ (to avoid violating the constraint $\|\phi(s,a)\|_\infty \leq 1$). Suppose the MDP has just one state and $2^{d-1}+1$ actions. One action has feature vector $(d\sigma/2, 0, \ldots, 0)$, and for each subset $S \subseteq \{2, \ldots, d\}$, there is an action a_S with a binary feature vector such that it is 1 for coordinates in S and 0 everywhere else. Let $w^\star = (1, 0, ..., 0)$. The optimal policy is to pick the first action which has cumulative reward of $\frac{d\sigma}{2(1-\gamma)}$. As $\eta_{ik} \sim \mathcal{N}(0, \sigma^2)$ for each k, with constant probability, roughly $d/2$ of the coordinates of the noised vector reward $\mathbf{w}_i$ will deviate by roughly $+\sigma$ and the first coordinate will not increase too much. In this case, the action corresponding to the coordinates with positive deviations will have reward on the order of $d\sigma/2$, beating action 1 to become optimal. Hence, this would lead to π_i picking this action and having 0 reward under $\mathbf{w}^\star$. As this occurs with constant probability for a policy in the data, and π^u is simply a mean of their rewards, its expected value would deviate from the optimum by at least a constant fraction of $d\sigma/2$.

4.2 It is Impossible to Outperform the Uniform Mixture in the Worst Case

An ostensible weakness of Theorem 4.2 is that even as the number of agents n goes to infinity, the reward of the uniform mixture may not approach that of the optimal policy, that is, there is a persistent gap. The example given in Section 4.1 shows the gap is not just an artifact of our analysis. This is expected, because the data contains some agents with suboptimal policies π_i, and a uniform mixture over these suboptimal policies must itself be suboptimal.

It is natural to ask, therefore, whether it is generally possible to achieve performance arbitrarily close to $\pi^\star$ (at least in the limit that n goes to infinity). The answer is negative. In fact, we show that — in the spirit of *minimax optimality* [10, 18] — one cannot hope to perform better than π^u itself in the worst case. Intuitively, there exist scenarios where it is impossible to tell good and bad policies apart by looking at the data, which means that the algorithm's performance depends on what can be gleaned from the "average data".

This follows from a surprising[1] result that we think of as "non-identifiability" of the optimal policy. To describe this property, we introduce some more notation. The distribution over the weight vector of each agent i, $\mathbf{w}_i = \mathbf{w}^\star + \boldsymbol{\eta}_i$, in turn induces a distribution over the optimal policy π_i executed by each agent. Denote this distribution by $\mathcal{P}(\mathbf{w}^\star)$.[2] Hence, each agent's optimal policy π_i is just

[1] At least it was surprising for us — we spent significant effort trying to prove the opposite result!

[2] Note that this distribution does not depend on i itself since the noise $\boldsymbol{\eta}_i$ is i.i.d. across the different agents.

a sample from this distribution $\mathcal{P}(\mathbf{w}^\star)$. In particular, as the number of agents goes to infinity, the empirical distribution of their optimal policies would exactly converge to $\mathcal{P}(\mathbf{w}^\star)$.

For the rest of this section, we make minimal assumptions on the noise vector $\boldsymbol{\eta}_i$. In particular, we merely assume that $\boldsymbol{\eta}_i$ follows a continuous distribution and that each of its coordinates is i.i.d. We are now ready to state our non-identifiability lemma.

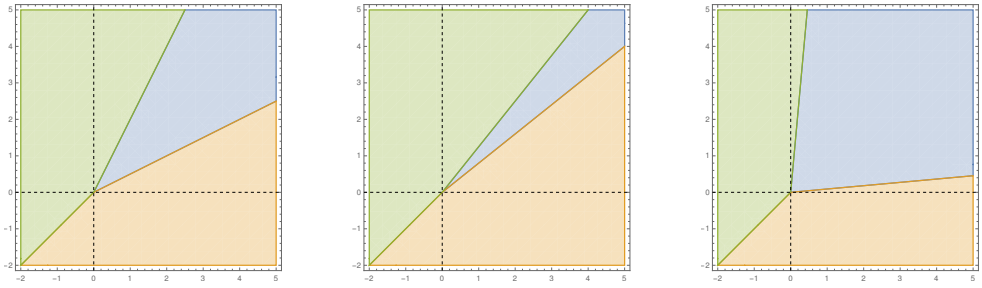
Lemma 4.3 (non-identifiability)**.** *For every continuous distribution $\mathcal{D}$ over $\mathbb{R}$, if η_{ik} is independently sampled from $\mathcal{D}$ for all $i \in N$ and $k \in [d]$, then there exists an MDP and weight vectors $\mathbf{w}_a^\star$, $\mathbf{w}_b^\star$ with optimal policies $\pi_a^\star$, $\pi_b^\star$, respectively, such that $\pi_a^\star \neq \pi_b^\star$ but $\mathcal{P}(\mathbf{w}_a^\star) = \mathcal{P}(\mathbf{w}_b^\star)$.*

Even if we had an infinite number of trajectories in our data, and even if we knew the exact optimal policy played by each player i, this information would amount to knowing $\mathcal{P}(\mathbf{w}^\star)$. Hence, if there exist two weight vectors $\mathbf{w}_a^\star$, $\mathbf{w}_b^\star$ with optimal policies $\pi_a^\star$, $\pi_b^\star$ such that $\pi_a^\star \neq \pi_b^\star$ and $\mathcal{P}(\mathbf{w}_a^\star) = \mathcal{P}(\mathbf{w}_b^\star)$, then we would not be able to identify whether the optimal policy is $\pi_a^\star$ or $\pi_b^\star$ regardless of how much data we had.

The proof of Lemma 4.3 is relegated to Appendix D. Here we provide a proof sketch.

Proof sketch of Lemma 4.3. The intuition for the lemma comes from the construction of an MDP with three possible policies, all of which have probability $1/3$ under $\mathcal{P}(\mathbf{w}^\star)$, even though one is better than the others. This MDP has a single state s, and three actions $\{a, b, c\}$ that lead back to s. Denote the corresponding policies by π_a, π_b, π_c. Let the feature expectations be $\phi(s, a) = [0.5, 0.5], \phi(s, b) = [1, -\delta/2], \phi(s, c) = [-\delta/2, 1]$, where $\delta > 0$ is a parameter. Let the ground truth weight vector be $\mathbf{w}^\star = (v_o, v_o)$, where v_o is such that the noised weight vector $\mathbf{w} = \mathbf{w}^\star + \boldsymbol{\eta}$ has probability strictly more than $1/3$ of lying in the first quadrant; such a value always exists for any noise distribution that is continuous and i.i.d. across coordinates.

Let us look at weight vectors $\mathbf{w}$ for which each of the three policies π_a, π_b and π_c are optimal. π_a is the optimal policy when $\mathbf{w}^\intercal \mu_a > \mathbf{w}^\intercal \mu_b$ and $\mathbf{w}^\intercal \mu_a > \mathbf{w}^\intercal \mu_c$, which is the intersection of the half-spaces $\mathbf{w}^\intercal(-1, 1+\delta) > 0$ and $\mathbf{w}^\intercal(1+\delta, -1) > 0$. Similarly, we can reason about the regions where π_b and π_c are optimal. These regions are illustrated in Figure 1 for different values of δ. Informally, as δ is decreased, the lines separating (π_a, π_c) and (π_a, π_b) move closer to each other (as shown for $\delta = 0.25$), while as δ is increased, these lines move away from each other (as shown for $\delta = 10$). By continuity and symmetry, there exists δ such that the probability of each of the regions (with respect to the random noise) is exactly $1/3$, showing that the MDP has the desired property.



$\delta = 1$ $\delta = 0.25$ $\delta = 10$

Figure 1: Regions of each optimal policy for different values of δ. Blue depicts the region where π_a is optimal, orange is where π_b is optimal, and green is where π_c is optimal.

To complete the proof of the lemma, we create a similar MDP with four features instead of three. We show that the two weight vectors $\mathbf{w}_a^\star = (v_o, v_o, 0, 0)$ and $\mathbf{w}_b^\star = (0, 0, v_o, v_o)$ lead to $\mathcal{P}(\mathbf{w}_a^\star) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) = \mathcal{P}(\mathbf{w}_b^\star)$, even though their corresponding optimal policies are π_a and π_b, respectively. □

For the next theorem, therefore, we can afford to be "generous:" we will give the algorithm (which is trying to compete with π^u) access to $\mathcal{P}(\mathbf{w}^\star)$, instead of restricting it to sampled trajectories. Formally, the theorem holds for any algorithm that takes a distribution over policies as input, and returns a randomized policy.

Theorem 4.4. *For every continuous distribution $\mathcal{D}$ over $\mathbb{R}$, if η_{ik} is independently sampled from $\mathcal{D}$ for all $i \in N$ and $k \in [d]$, then there exists an MDP such that for any algorithm $\mathcal{A}$ from*

distributions over policies to randomized policies, there exists a ground truth weight vector $\mathbf{w}^\star$ *such that* $R^{\mathbf{w}^\star}(\mathcal{A}(\mathcal{P}(\mathbf{w}^\star)) \leq R^{\mathbf{w}^\star}(\pi^u) < R^{\mathbf{w}^\star}(\pi^\star)$.

In words, the constructed instance is such that, even given infinite data, no algorithm can outperform the uniform mixture, and, moreover, the reward of the uniform mixture is bounded away from the optimum. The theorem's proof is give in Appendix E.

5 Empirical Results

As we have seen in Section 4.1, the gap between $R^{\mathbf{w}^\star}(\pi^\star)$ and $R^{\mathbf{w}^\star}(\pi^u)$ is upper bounded by $O(d\sqrt{u} + \nu\sqrt{d/u} + b/\sqrt{u})$ when η_{ik}^2 is sub-exponential, or $O(d\sigma)$ when η_{ik} is Gaussian. Further, Section 3 shows that a policy $\tilde{\pi}$ that matches feature expectations of the observed trajectories is very close to π^u in terms of cumulative reward $R^{\mathbf{w}^\star}$. In this section, we empirically examine the gaps between $\tilde{\pi}$ (obtained by a "feature matching" IRL algorithm), π^u and $\pi^\star$.

5.1 Methodology

As our IRL algorithm we use Apprenticeship Learning, which guarantees the feature-matching property (see Section 3 and Appendix A). By Theorem 3.2 we may safely assume that any IRL algorithm that matches feature expectations would have essentially identical rewards, and therefore would show very similar behavior in our experiments.

We perform our experiments in the following two domains.

Grab a Milk. We adapt the "Grab a Milk" MDP, a route planning RL domain [26] involving ethical considerations, to our setting. The MDP is defined by a 10 by 10 grid room, where the agent starts at $(0, 0)$ and has to reach a bottle of milk positioned at $(9, 9)$. There are also 16 babies in the room, 5 of which are crying for attention. When the agent crosses a crying baby, they can help soothe the baby, but on crossing a non-crying baby, the agent disturbs the baby. Hence, the goal of this task is to minimize the number of steps to the milk, while at the same time soothing as many crying babies as possible along the way and avoiding crossing non-crying babies. This MDP is adapted to our setting, by defining each state (or grid square) to have three features $\phi(s)$.[3] The first feature captures the reward of taking a step, and is set to -1 if the state is non-terminal, whereas it is set to 5 for the terminal state $(9, 9)$. The second is a boolean feature depicting whether there is a crying baby in the particular grid square, and similarly the third is a boolean feature depicting whether there is a non-crying baby in the particular grid square. The rewards in the MDP are then defined as $r^{\mathbf{w}^\star}(s) = (\mathbf{w}^\star)^\intercal\phi(s)$ where the ground truth weight vector is given by $\mathbf{w}^\star = [1, 0.5, -0.5]$. Intuitively, this weight vector $\mathbf{w}^\star$ can be interpreted as the weights for different ethical factors, and each member of society has a noised version of this weight.

Sailing. The other domain we use is a modified version of the "Sailing" MDP [13]. The Sailing MDP is also a gridworld domain (we use the same size of 10 by 10), where there is a sailboat starting at $(0, 0)$ and navigating the grid under fluctuating wind conditions. The goal of the MDP is to reach a specified grid square as quickly as possible. We adapt this domain to our setting by removing the terminal state, and instead adding features for each grid square.[4] Now, the goal of the agent is not to reach a certain point as quickly as possible, but to navigate this grid while maximizing (or minimizing) the weighted sum of these features. We use 10 features for each grid square, and these are independently sampled from a uniform distribution over $(-1, 1)$. The ground truth weight vector $\mathbf{w}^\star$, which defines the weights of these features for the net reward, is also randomly sampled from independent Unif$(-1, 1)$ for each coordinate. As before, this weight vector $\mathbf{w}^\star$ can be interpreted as the weights for different bounties, and each member has a noised version of this weight.

Being gridworld domains, in both the MDPs, the agent has four actions to choose from at each state (one for each direction). The transition dynamics are as follows: On taking a particular action from a given state, the agent moves in that direction with probability 0.95, but with a probability of 0.05 it moves in a different direction uniformly at random. We use a discount factor of 0.95 in both domains.

[3] For these MDPs, the rewards depend only on the states and not state-action pairs, and hence the reward function can be defined as $r^{\mathbf{w}}(s, a) = r^{\mathbf{w}}(s) = \mathbf{w}^\intercal\phi(s)$.

[4] Intuitively, these features could represent aspects like "abundance of fish" in that grid square for fishing, "amount of trash" in that square that could be cleaned up, "possible treasure" for treasure hunting, etc.

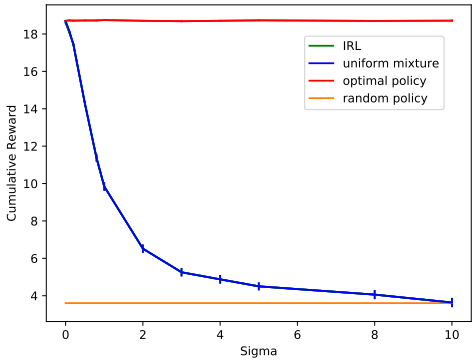


Figure 2: Performance on the Sailing MDP. Error bars show 95% confidence intervals.

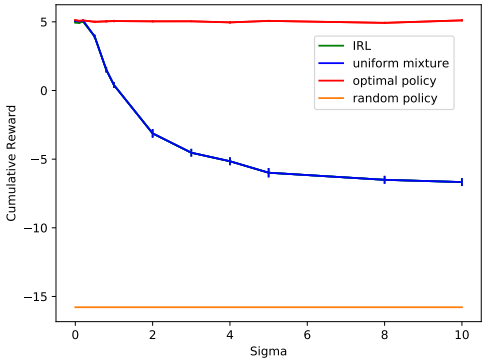


Figure 3: Performance on the Grab a Milk MDP. Error bars show 95% confidence intervals.

We generate the trajectories $\{\tau_1, \ldots, \tau_n\}$ as described in Section 3, and use a Gaussian distribution for the noise. That is, $\boldsymbol{\eta}_i \sim \mathcal{N}(0, \sigma^2 I_d)$. We generate a total of $n = 50$ trajectories, each of length $L = 30$. IRL is then performed on this data and we analyze its reward as σ is varied. A learning rate of 0.001 is used for the Apprenticeship Learning algorithm.

5.2 Results

Figures 2 and 3 show the performance of π^u and the IRL algorithm as σ is varied. We also include the performance of $\pi^\star$ and a purely random policy π^r (which picks a uniformly random action at each step), as references. Each point in these graphs is averaged over 50 runs (of data generation).

For both domains, the first thing to note is that the uniform mixture π^u and the IRL algorithm have nearly identical rewards, which is why the green IRL curve is almost invisible. This confirms that matching feature expectations leads to performance approximating the uniform mixture.

Next, as expected, one can observe that as σ increases, the gap between $R^\star(\pi^\star)$ and $R^\star(\pi^u)$ also increases. Further, for both domains, this gap saturates around $\sigma = 10$ and the $R^\star(\pi^u)$ curve flattens from there (hence, we do not include larger values of σ in either graph). Note that, in both domains, the ground truth weight vector $\mathbf{w}^\star$ is generated such that $\|\mathbf{w}^\star\|_\infty \leq 1$. Hence, a standard deviation of 10 in the noise overshadows the true weight vector $\mathbf{w}^\star$, leading to the large gap shown in both graphs. Looking at more reasonable levels of noise (with respect to the norm of the weights), like $\sigma \in [0, 1]$, we can see that $R^\star(\pi^u)$ drops approximately linearly, as suggested by Theorem 4.2. In particular, it is 14.27 at $\sigma = 0.5$ and 9.84 at $\sigma = 1.0$ for Sailing, and it is 3.93 at $\sigma = 0.5$ and 0.39 at $\sigma = 1.0$ for Grab a Milk.

Finally, we compare the performance of π^u with that of the purely random policy π^r. As σ becomes very large, each $\mathbf{w}_i$ is distributed almost identically across the coordinates. Nevertheless, because of the structure of the Grab a Milk MDP, $R^\star(\pi^u)$ still does significantly better than $R^\star(\pi^r)$. By contrast, Sailing has features that are sampled i.i.d. from $\text{Unif}(-1, 1)$ for each state, which leads the two policies, π^u and π^r, to perform similarly for large values of σ.

6 Discussion

We have shown that the uniform mixture π^u over the policies of individual agents provides strong performance guarantees. However, one can plausibly argue that π^u is not "better" than the individual policies *ex post*, only in expectation. By contrast, IRL algorithms pool the feature expectations of the trajectories $\tau_1, \ldots, \tau_n$ together, and try to recover a policy that approximately matches them. Therefore, we believe that IRL algorithms do a much better job of aggregating the individual policies than π^u does, while giving almost the same optimality guarantees.

Apropos aggregation, one could make it more explicit. Specifically, suppose that we have learned (via IRL) a reward function and an optimal policy for each agent, which reflect the agent's values and beliefs. Note that, in contrast to our setting, this would require a significant amount of data for each agent. Still, how should these policies be aggregated into a single policy? We can cast this as a problem of allocating public goods. A naïve approach would compute each agent's reward for each

possible policy, and choose the policy that, say, maximizes the Nash social welfare [7]; but this is a pipe dream, due to seemingly insurmountable computational barriers. The discovery of tractable methods for this policy aggregation problem may provide attractive alternatives to the approach presented in this paper.

References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 1–8, 2004.

[2] M. Babeş-Vroman, V. Marivate, K. Subramanian, and M. L. Littman. Apprenticeship learning about multiple intentions. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 897–904, 2011.

[3] F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia, editors. *Handbook of Computational Social Choice*. Cambridge University Press, 2016.

[4] J. Choi and K.-E. Kim. Nonparametric Bayesian inverse reinforcement learning for multiple reward functions. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 314–322, 2012.

[5] N. Chomsky. *Aspects of the Theory of Syntax*. MIT Press, 1965.

[6] T. Everitt, V. Krakovna, L. Orseau, and S. Legg. Reinforcement learning with a corrupted reward channel. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4705–4713, 2017.

[7] B. Fain, K. Munagala, and N. Shah. Fair allocation of indivisible public goods. In *Proceedings of the 19th ACM Conference on Economics and Computation (EC)*, pages 575–592, 2018.

[8] R. Freedman, J. Schaich Borg, W. Sinnott-Armstrong, J. P. Dickerson, and V. Conitzer. Adapting a kidney exchange algorithm to align with human values. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 1636–1643, 2018.

[9] R. Givan, S. Leach, and T. Dean. Bounded-parameter Markov decision processes. *Artificial Intelligence*, 122(1–2):71–109, 2000.

[10] J. L. Hodges Jr and E. L. Lehmann. Some problems in minimax point estimation. *The Annals of Mathematical Statistics*, pages 182–197, 1950.

[11] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[12] J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[13] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.

[14] J. Mikhail. *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge University Press, 2011.

[15] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[16] A. Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pages 663–670, 2000.

[17] R. Noothigattu, S. S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, and A. D. Procaccia. A voting-based system for ethical decision making. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 1587–1594, 2018.

[18] F. Perron and E. Marchand. On the minimax estimator of a bounded normal mean. *Statistics and Probability Letters*, 58:327–333, 2002.

[19] J. Rawls. *A Theory of Justice*. Harvard University Press, 1971.

[20] K. Regan and C. Boutilier. Regret-based reward elicitation for Markov decision processes. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 444–451, 2009.

[21] K. Regan and C. Boutilier. Robust policy computation in reward-uncertain MDPs using nondominated policies. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI)*, pages 1127–1133, 2010.

[22] S. Russell, D. Dewey, and M. Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4):105–114, 2015.

[23] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

[24] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[25] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[26] Yueh-Hua Wu and Shou-De Lin. A low-cost ethics shaping approach for designing reinforcement learning agents. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 1687–1694, 2018.

[27] J. Zheng, S. Liu, and L. M. Ni. Robust Bayesian inverse reinforcement learning with sparse behavior noise. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2198–2205, 2014.

[28] B. D. Ziebart. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. Ph.D. thesis, Carnegie Mellon University, 2010.

[29] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*, pages 1433–1438, 2008.

A IRL Algorithms

In this appendix we identify two well-known algorithms that match feature expectations.

A.1 Apprenticeship Learning

Under the classic Apprenticeship Learning algorithm, designed by Abbeel and Ng [1], a policy $\pi^{(0)}$ is selected to begin with. Its feature expectation $\mu(\pi^{(0)})$ is computed and added to the bag of feature expectations. At each step,

$$t^{(i)} = \max_{\mathbf{w}:\|\mathbf{w}\|_2\leq 1} \min_{j\in\{0,..,i-1\}} \mathbf{w}^\intercal \left(\frac{1}{n}\sum_{i=1}^{n}\phi(\tau_i) - \mu\left(\pi^{(j)}\right)\right)$$

is computed along with the weight $\mathbf{w}^{(i)}$ that achieved this. When $t^{(i)} \leq \epsilon$ the algorithm terminates, otherwise the associated optimal policy $\pi^{(i)}$ is computed, and its corresponding feature expectation vector $\mu(\pi^{(i)})$ is added to the bag of feature expectations. The algorithm provides the following guarantee.

Theorem A.1 (adapted from Abbeel and Ng 2004)**.** *For any $\epsilon > 0$, the Apprenticeship Learning algorithm terminates with $t^{(i)} \leq \epsilon$ after a number of iterations bounded by*

$$T = O\left(\frac{d}{(1-\gamma)^2\epsilon^2}\ln\frac{d}{(1-\gamma)\epsilon}\right),$$

and outputs a mixture over $\pi^{(1)}, ..., \pi^{(T)}$ that ϵ-matches the feature expectations of the observed trajectories.

Note that it is necessary for us to use a randomized policy, in contrast to the case where a single deterministic policy generated all the trajectory samples, as, in our case, typically there is no single deterministic policy that matches the feature expectations of the observed trajectories.

A.2 Max Entropy

We next discuss the Max Entropy algorithm of Ziebart et al. [29], which optimizes the max entropy of the probability distribution over trajectories subject to the distribution satisfying approximate feature matching. This is done to resolve the potential ambiguity of there being multiple stochastic policies that satisfy feature matching. Optimizing entropy is equivalent to maximizing the regularized likelihood $L(\mathbf{w})$ of the observed trajectories. Specifically, the objective is

$$L(\mathbf{w}) = \max_{\mathbf{w}} \sum_{i=1}^{n} \log \Pr[\tau_i|\mathbf{w}, T] - \sum_{i=1}^{d} \rho_i \|\mathbf{w}_i\|_1,$$

with

$$\Pr[\tau_i|\mathbf{w}, T] = \frac{e^{\mathbf{w}^\intercal \phi(\tau_i)}}{Z(\mathbf{w}, T)} \prod_{s_t, a_t, s_{t+1} \in \tau_i} T(s_t, a_t, s_{t+1}).$$

The regularization term is introduced to allow for approximate feature matching since the observed empirical feature expectation may differ from the true expectation. Let $\boldsymbol{\rho}$ be an upper bound on this difference, i.e., for all $k = 1, \ldots, d$,

$$\rho_k \geq \left| \frac{1}{n} \sum_{i=1}^{n} \phi(\tau_i)_k - \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \phi(\tau_i)_k \right] \right|.$$

One may then derive that the gradient of $L(\mathbf{w})$ is the difference between the feature expectation induced $\mathbf{w}$ and the observed feature expectation.

Theorem A.2 (adapted from Ziebart et al. 2008)**.** *Let $\epsilon > 0$, and assume that the Max Entropy algorithm finds $\mathbf{w}$ such that $|\nabla L(\mathbf{w})| < \epsilon$, then this $\mathbf{w}$ corresponds to a randomized policy that $(\epsilon + \|\boldsymbol{\rho}\|_1)$-matches the feature expectations of the observed trajectories.*

The assumption on the gradient is needed because the above optimization objective is derived only with the approximate feature matching constraint. MDP dynamics is not explicitly encoded into the optimization. Instead, heuristically, the likelihood of each trajectory $\Pr[\tau_i|\mathbf{w}, T]$ is weighted by the product of the transition probabilities of its steps. The follow-up work of Ziebart [28] addresses this by explicitly introducing MDP constraints into the optimization, and optimizing for the causal entropy, thereby achieving unconditional feature matching.

B Proof of Theorem 3.2

We need to bound the difference between $R^{\mathbf{w}^\star}(\tilde{\pi})$ and $R^{\mathbf{w}^\star}(\pi^u)$. First, recall that $\tilde{\pi}$ $\epsilon/3-$matches the feature expectations of $\tau_1, \ldots, \tau_n$. It holds that

$$\begin{aligned} \left| R^{\mathbf{w}^\star}(\tilde{\pi}) - (\mathbf{w}^\star)^\intercal \left(\frac{1}{n} \sum_{i=1}^{n} \phi(\tau_i) \right) \right| &= \left| (\mathbf{w}^\star)^\intercal \left(\mu(\tilde{\pi}) - \frac{1}{n} \sum_{i=1}^{n} \phi(\tau_i) \right) \right| \\ &\leq \|\mathbf{w}^\star\|_2 \left\| \mu(\tilde{\pi}) - \frac{1}{n} \sum_{i=1}^{n} \phi(\tau_i) \right\|_2 \\ &\leq \frac{\epsilon}{3}, \end{aligned} \tag{1}$$

where the second transition follows from the Cauchy–Schwarz inequality, and the last from the assumption that $\|\mathbf{w}^\star\|_2 \leq 1$. Hence, it is sufficient to demonstrate that, with probability at least $1 - \delta$,

$$\left| (\mathbf{w}^\star)^\intercal \left(\frac{1}{n} \sum_{i=1}^{n} \phi(\tau_i) \right) - R^{\mathbf{w}^\star}(\pi^u) \right| \leq \frac{2\epsilon}{3}, \tag{2}$$

as the theorem would then follow from Equations (1), and (2) by the triangle inequality.

We note that the difference on the left hand side of Equation (2) is due to two sources of noise.

1. The finite number of samples of trajectories which, in our setting, originates from multiple policies.

2. The truncated trajectories τ_i which are limited to L steps.

Formally, let τ_i' denote the infinite trajectory for each i, then the difference can be written as

$$\left|(\mathbf{w}^\star)^\intercal \left(\frac{1}{n}\sum_{i=1}^{n}\phi(\tau_i)\right) - R^{\mathbf{w}^\star}(\pi^u)\right|$$
$$\leq \left|(\mathbf{w}^\star)^\intercal \left(\frac{1}{n}\sum_{i=1}^{n}\phi(\tau_i)\right) - (\mathbf{w}^\star)^\intercal \left(\frac{1}{n}\sum_{i=1}^{n}\phi(\tau_i')\right)\right| + \left|(\mathbf{w}^\star)^\intercal \left(\frac{1}{n}\sum_{i=1}^{n}\phi(\tau_i')\right) - R^{\mathbf{w}^\star}(\pi^u)\right|$$

Bounding finite sample noise. We wish to bound:

$$\left|(\mathbf{w}^\star)^\intercal \left(\frac{1}{n}\sum_{i=1}^{n}\phi(\tau_i')\right) - R^{\mathbf{w}^\star}(\pi^u)\right| = \left|\frac{1}{n}\left(\sum_{i=1}^{n}(\mathbf{w}^\star)^\intercal(\phi(\tau_i') - \mu(\pi_i))\right)\right|. \tag{3}$$

Define random variable $Z_i = (\mathbf{w}^\star)^\intercal(\phi(\tau_i') - \mu(\pi_i))$. Then the right-hand side of Equation (3) may be expressed as $|\frac{1}{n}\sum_{i=1}^{n} Z_i|$. Furthermore, Z_i is such that $\mathbb{E}[\phi(\tau_i')_k] = \mu(\pi_i)_k$ for all $k = 1, \ldots, d$. This is because a policy π_i defines a distribution over trajectories, and τ_i' is a draw from this distribution. Using the linearity of expectation, it follows that

$$\mathbb{E}[Z_i] = (\mathbf{w}^\star)^\intercal \mathbb{E}[\phi(\tau_i') - \mu(\pi_i)] = 0.$$

Moreover,

$$|Z_i| \leq \|\mathbf{w}^\star\|_2 \|\phi(\tau_i')\|_2 + \|\mathbf{w}^\star\|_2 \|\mu(\pi_i)\|_2 \leq \frac{2\sqrt{d}}{1-\gamma},$$

since $\|\phi(s,\cdot)\|_\infty = 1$. Thus, using Hoeffding's inequality, we conclude that

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| > \frac{\epsilon}{3}\right] \leq 2\exp\left(-\frac{2n\left(\frac{\epsilon}{3}\right)^2}{(\frac{4\sqrt{d}}{1-\gamma})^2}\right) \leq \delta,$$

where the last transition holds by our choice of n.

Bounding bias due to truncated trajectories. We wish to bound:

$$\left|(\mathbf{w}^\star)^\intercal \left(\frac{1}{n}\sum_{i=1}^{n}\phi(\tau_i)\right) - (\mathbf{w}^\star)^\intercal \left(\frac{1}{n}\sum_{i=1}^{n}\phi(\tau_i')\right)\right|.$$

For each trajectory τ_i, truncating after L steps incurs a reward difference of:

$$\begin{aligned}
|(\mathbf{w}^\star)^\intercal\phi(\tau_i') - (\mathbf{w}^\star)^\intercal\phi(\tau_i)| &= \left|(\mathbf{w}^\star)^\intercal \sum_{t=L}^{\infty}\gamma^t \phi(\tau_i'(s_t), \tau_i'(a_t))\right| \\
&\leq \sum_{t=L}^{\infty}\gamma^t \|\mathbf{w}^\star\|_2 \|\phi(\tau_i'(s_t), \tau_i'(a_t))\|_2 \\
&\leq \gamma^L \frac{\sqrt{d}}{1-\gamma} \\
&\leq \frac{\epsilon}{3},
\end{aligned}$$

where the third transition holds because $\|\phi(\tau_i(s_t), \tau_i(a_t))\|_2 \leq \sqrt{d}$, and the last transition follows from our choice of L. Hence, we obtain

$$\left|(\mathbf{w}^\star)^\intercal \left(\frac{1}{n}\sum_{i=1}^{n}\phi(\tau_i)\right) - (\mathbf{w}^\star)^\intercal \left(\frac{1}{n}\sum_{i=1}^{n}\phi(\tau_i')\right)\right| \leq \frac{1}{n}\sum_{i=1}^{n}|(\mathbf{w}^\star)^\intercal\phi(\tau_i) - (\mathbf{w}^\star)^\intercal\phi(\tau_i')| \leq \frac{\epsilon}{3}.$$

□

C Proof of Theorem 4.2

As π^u is a uniform distribution over the policies $\pi_1, \ldots, \pi_n$, its expected reward is given by

$$R^{\mathbf{w}^\star}(\pi^u) = \frac{1}{n}\sum_{i=1}^{n} R^{\mathbf{w}^\star}(\pi_i). \tag{4}$$

Observe that $R^{\mathbf{w}^\star}(\pi_i)$ is a random variable which is i.i.d. across i, as the corresponding noise $\boldsymbol{\eta}_i$ is i.i.d. as well. We analyze the expectation of the difference with respect to $R^{\mathbf{w}^\star}(\pi^\star)$.

First, note that for a weight vector $\mathbf{w}$ and policy π,

$$R^{\mathbf{w}}(\pi) = \sum_{t=0}^{\infty}\gamma^t \sum_{s\in S} P_\pi(s,t)\mathbf{w}^\intercal\phi(s,\pi(s)), \tag{5}$$

where $P_\pi(s,t)$ denotes the probability of being in state s on executing policy π from the start. Hence, for each $i \in N$, we have

$$\begin{aligned}
&R^{\mathbf{w}^\star}(\pi^\star) - R^{\mathbf{w}^\star}(\pi_i)\\
&\quad= \sum_{t=0}^{\infty}\gamma^t \sum_{s\in S}\Big[P_{\pi^\star}(s,t)(\mathbf{w}^\star)^\intercal\phi(s,\pi^\star(s)) - P_{\pi_i}(s,t)(\mathbf{w}^\star)^\intercal\phi(s,\pi_i(s))\Big]\\
&\quad= \sum_{t=0}^{\infty}\gamma^t \sum_{s\in S}\Big[P_{\pi^\star}(s,t)(\mathbf{w}_i-\boldsymbol{\eta}_i)^\intercal\phi(s,\pi^\star(s)) - P_{\pi_i}(s,t)(\mathbf{w}_i-\boldsymbol{\eta}_i)^\intercal\phi(s,\pi_i(s))\Big]\\
&\quad= R^{\mathbf{w}_i}(\pi^\star) - R^{\mathbf{w}_i}(\pi_i) + \sum_{t=0}^{\infty}\gamma^t \sum_{s\in S}\Big[-P_{\pi^\star}(s,t)\boldsymbol{\eta}_i^\intercal\phi(s,\pi^\star(s)) + P_{\pi_i}(s,t)\boldsymbol{\eta}_i^\intercal\phi(s,\pi_i(s))\Big]\\
&\quad\le \sum_{t=0}^{\infty}\gamma^t \sum_{s\in S}\Big[-P_{\pi^\star}(s,t)\boldsymbol{\eta}_i^\intercal\phi(s,\pi^\star(s)) + P_{\pi_i}(s,t)\boldsymbol{\eta}_i^\intercal\phi(s,\pi_i(s))\Big]\\
&\quad= \sum_{k=1}^{d}\eta_{ik}\left[\sum_{t=0}^{\infty}\gamma^t \sum_{s\in S}\Big[-P_{\pi^\star}(s,t)\phi(s,\pi^\star(s))_k + P_{\pi_i}(s,t)\phi(s,\pi_i(s))_k\Big]\right]\\
&\quad:= \sum_{k=1}^{d}\eta_{ik}\alpha_{ik},
\end{aligned} \tag{6}$$

where the inequality holds since $R^{\mathbf{w}_i}(\pi_i) \ge R^{\mathbf{w}_i}(\pi^\star)$, which, in turn, holds because π_i is optimal under $\mathbf{w}_i$.

Using the assumption that $\|\phi(s,a)\|_\infty \le 1$, it holds that $\left|\sum_{s\in S} P_\pi(s,t)\phi(s,a)_k\right| \le 1$ for any policy π. We can therefore bound $|\alpha_{ik}|$ as follows.

$$\begin{aligned}
|\alpha_{ik}| &\le \sum_{t=0}^{\infty}\gamma^t\left|\sum_{s\in S}[-P_{\pi^\star}(s,t)\phi(s,\pi^\star(s))_k + P_{\pi_i}(s,t)\phi(s,\pi_i(s))_k]\right|\\
&\le \sum_{t=0}^{\infty}\gamma^t\left[\left|\sum_{s\in S}P_{\pi^\star}(s,t)\phi(s,\pi^\star(s))_k\right| + \left|\sum_{s\in S}P_{\pi_i}(s,t)\phi(s,\pi_i(s))_k\right|\right]\\
&\le \frac{2}{1-\gamma}.
\end{aligned}$$

Therefore, it holds that

$$\|\boldsymbol{\alpha}_i\|_2 = \sqrt{\sum_{k=1}^{d}\alpha_{ik}^2} \le \sqrt{\sum_{k=1}^{d}\left(\frac{2}{1-\gamma}\right)^2} = \frac{2\sqrt{d}}{(1-\gamma)}.$$

Using this bound along with Equation (6), we obtain

$$R^{\mathbf{w}^\star}(\pi^\star) - R^{\mathbf{w}^\star}(\pi_i) \le \sum_{k=1}^{d}\eta_{ik}\alpha_{ik} \le \|\boldsymbol{\eta}_i\|_2\|\boldsymbol{\alpha}_i\|_2 \le \frac{2\sqrt{d}}{(1-\gamma)}\sqrt{\sum_{k=1}^{d}\eta_{ik}^2}$$

$$= \frac{2d}{(1-\gamma)} \sqrt{\frac{1}{d} \sum_{k=1}^{d} \eta_{ik}^2}. \tag{7}$$

Denote $u = \mathbb{E}[\frac{1}{d} \sum_{k=1}^{d} \eta_{ik}^2]$. To compute the expected value of the previous expression (with respect to the randomness of the noise $\boldsymbol{\eta}_i$), we analyze

$$\begin{aligned}
\mathbb{E}\left[\sqrt{\frac{1}{d} \sum_{k=1}^{d} \eta_{ik}^2}\right] &= \int_0^\infty \Pr\left[\sqrt{\frac{1}{d} \sum_{k=1}^{d} \eta_{ik}^2} \geq x\right] dx = \int_0^\infty \Pr\left[\frac{1}{d} \sum_{k=1}^{d} \eta_{ik}^2 \geq x^2\right] dx \\
&= \int_0^{\sqrt{u}} \Pr\left[\frac{1}{d} \sum_{k=1}^{d} \eta_{ik}^2 \geq x^2\right] dx + \int_{\sqrt{u}}^\infty \Pr\left[\frac{1}{d} \sum_{k=1}^{d} \eta_{ik}^2 \geq x^2\right] dx \\
&\leq \int_0^{\sqrt{u}} 1 \, dx + \int_{\sqrt{u}}^\infty \Pr\left[\frac{1}{d} \sum_{k=1}^{d} \eta_{ik}^2 \geq x^2\right] dx \\
&= \sqrt{u} + \int_0^\infty \Pr\left[\frac{1}{d} \sum_{k=1}^{d} \eta_{ik}^2 \geq u + t\right] \frac{1}{2\sqrt{u+t}} dt \\
&\leq \sqrt{u} + \frac{1}{2\sqrt{u}} \int_0^\infty \Pr\left[\frac{1}{d} \sum_{k=1}^{d} \eta_{ik}^2 \geq u + t\right] dt,
\end{aligned}$$

where the fourth transition is obtained by changing the variable using $x = \sqrt{u+t}$. But since each η_{ik}^2 is sub-exponential with parameters (ν, b), from Lemma 4.1 we have

$$\Pr\left[\frac{1}{d} \sum_{k=1}^{d} \eta_{ik}^2 \geq u + t\right] \leq \begin{cases} \exp\left(-\frac{dt^2}{2\nu^2}\right) & \text{for } 0 \leq t \leq \frac{\nu^2}{b} \\ \exp\left(-\frac{dt}{2b}\right) & \text{for } t > \frac{\nu^2}{b} \end{cases}.$$

Plugging this into the upper bound for the expected value gives us

$$\begin{aligned}
\mathbb{E}\left[\sqrt{\frac{1}{d} \sum_{k=1}^{d} \eta_{ik}^2}\right] &\leq \sqrt{u} + \frac{1}{2\sqrt{u}} \int_0^\infty \Pr\left[\frac{1}{d} \sum_{k=1}^{d} \eta_{ik}^2 \geq u + t\right] dt \\
&\leq \sqrt{u} + \frac{1}{2\sqrt{u}} \left[\int_0^{\frac{\nu^2}{b}} \exp\left(-\frac{dt^2}{2\nu^2}\right) dt + \int_{\frac{\nu^2}{b}}^\infty \exp\left(-\frac{dt}{2b}\right) dt\right] \\
&= \sqrt{u} + \frac{1}{2\sqrt{u}} \left[\int_0^{\frac{\nu\sqrt{d}}{b}} \exp\left(-\frac{z^2}{2}\right) \frac{\nu}{\sqrt{d}} dz + \left(-\frac{2b}{d}\right) \exp\left(-\frac{dt}{2b}\right) \Big|_{\frac{\nu^2}{b}}^\infty\right] \\
&= \sqrt{u} + \frac{1}{2\sqrt{u}} \left[\sqrt{\frac{2\pi}{d}} \nu \int_0^{\frac{\nu\sqrt{d}}{b}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz + \frac{2b}{d} \exp\left(-\frac{d\nu^2}{2b^2}\right)\right] \\
&= \sqrt{u} + \frac{1}{2\sqrt{u}} \left[\sqrt{\frac{2\pi}{d}} \nu \left(\Phi\left(\frac{\nu\sqrt{d}}{b}\right) - \frac{1}{2}\right) + \frac{2b}{d} \exp\left(-\frac{d\nu^2}{2b^2}\right)\right] \\
&= \sqrt{u} + \sqrt{\frac{\pi}{2ud}} \nu \left(\Phi\left(\frac{\nu\sqrt{d}}{b}\right) - \frac{1}{2}\right) + \frac{b}{d\sqrt{u}} \exp\left(-\frac{d\nu^2}{2b^2}\right),
\end{aligned} \tag{8}$$

where the transition in the third line is obtained by changing the variable using $t = \frac{v}{\sqrt{d}} z$, and Φ denotes the CDF of a standard normal distribution. Hence, taking an expected value for Equation (7) and plugging in Equation (8), we obtain

$$\mathbb{E}\left[R^{\mathbf{w}^\star}(\pi^\star) - R^{\mathbf{w}^\star}(\pi_i)\right] \leq \frac{2d}{(1-\gamma)} \left[\sqrt{u} + \sqrt{\frac{\pi}{2ud}} \nu \left(\Phi\left(\frac{\nu\sqrt{d}}{b}\right) - \frac{1}{2}\right) + \frac{b}{d\sqrt{u}} \exp\left(-\frac{d\nu^2}{2b^2}\right)\right].$$

Rearranging this equation, we have

$$\mathbb{E}\left[R^{\mathbf{w}^\star}(\pi_i)\right] \geq R^{\mathbf{w}^\star}(\pi^\star) - \frac{2d}{(1-\gamma)}\left[\sqrt{u} + \sqrt{\frac{\pi}{2ud}}\nu\left(\Phi\left(\frac{\nu\sqrt{d}}{b}\right) - \frac{1}{2}\right) + \frac{b}{d\sqrt{u}}\exp\left(-\frac{d\nu^2}{2b^2}\right)\right].$$

Taking an expectation over Equation (4) gives us $\mathbb{E}\left[R^{\mathbf{w}^\star}(\pi^u)\right] = \mathbb{E}\left[R^{\mathbf{w}^\star}(\pi_i)\right]$, and the theorem directly follows. □

We remark that Theorem 4.2 can easily be strengthened to obtain a high probability result (at the cost of complicating its statement). Indeed, the reward of the uniform mixture $R^{\mathbf{w}^\star}(\pi^u)$ is the average of the individual policy rewards $R^{\mathbf{w}^\star}(\pi_i)$, which are i.i.d. Further, each of these rewards is bounded, because of the constraints on $\mathbf{w}^\star$ and ϕ. Hence, Hoeffding's inequality would show that $R^{\mathbf{w}^\star}(\pi^u)$ strongly concentrates around its mean.

D Proof of Lemma 4.3

Before proving the lemma, we look at a relatively simple example that we will use later to complete the proof.

D.1 Simpler Example

Consider an MDP with a single state s, and three actions $\{a, b, c\}$. Since s is the only state, $T(s, a, s) = T(s, b, s) = T(s, c, s) = 1$, and D is degenerate at s. This implies that there are only three possible policies, denoted by π_a, π_b, π_c (which take actions a, b, c respectively from s). Let the feature expectations be

$$\begin{aligned}\phi(s, a) &= [0.5, 0.5],\\ \phi(s, b) &= [1, -\delta/2],\\ \phi(s, c) &= [-\delta/2, 1],\end{aligned}$$

where $\delta > 0$ is a parameter. Hence, the feature expectations of the policies $\{\pi_a, \pi_b, \pi_c\}$ are respectively

$$\begin{aligned}\mu_a &= \frac{1}{2(1-\gamma)}[1, 1],\\ \mu_b &= \frac{1}{2(1-\gamma)}[2, -\delta],\\ \mu_c &= \frac{1}{2(1-\gamma)}[-\delta, 2].\end{aligned}$$

Let the ground truth weight vector be $\mathbf{w}^\star = (v_o, v_o)$, where v_o is a "large enough" positive constant. In particular, v_o is such that the noised weight vector $\mathbf{w} = \mathbf{w}^\star + \boldsymbol{\eta}$ has probability strictly more than $1/3$ of lying in the first quadrant. For concreteness, set v_o to be such that $\Pr(\mathbf{w} > 0) = 1/2$. Such a point always exists for any noise distribution (that is continuous and i.i.d. across coordinates). Specifically, it is attained at $v_o = -F^{-1}(1 - \frac{1}{\sqrt{2}})$, where F^{-1} is the inverse CDF of each coordinate of the noise distribution. This is because at this value of v_o,

$$\begin{aligned}\Pr(\mathbf{w} > 0) &= \Pr((v_o, v_o) + (\eta_1, \eta_2) > 0) = \Pr(v_o + \eta_1 > 0)^2\\ &= \Pr(\eta_1 > -v_o)^2 = (1 - F(-v_o))^2 = \left(\frac{1}{\sqrt{2}}\right)^2 = \frac{1}{2}.\end{aligned}$$

Let us look at weight vectors $\mathbf{w}$ for which each of the three policies π_a, π_b and π_c are optimal. π_a is the optimal policy when $\mathbf{w}^\intercal\mu_a > \mathbf{w}^\intercal\mu_b$ and $\mathbf{w}^\intercal\mu_a > \mathbf{w}^\intercal\mu_c$, which is the intersection of the half-spaces $\mathbf{w}^\intercal(-1, 1+\delta) > 0$ and $\mathbf{w}^\intercal(1+\delta, -1) > 0$. On the other hand, π_b is optimal when $\mathbf{w}^\intercal\mu_b > \mathbf{w}^\intercal\mu_a$ and $\mathbf{w}^\intercal\mu_b > \mathbf{w}^\intercal\mu_c$, which is the intersection of the half-spaces $\mathbf{w}^\intercal(-1, 1+\delta) < 0$ and $\mathbf{w}^\intercal(1, -1) > 0$. Finally, π_c is optimal when $\mathbf{w}^\intercal\mu_c > \mathbf{w}^\intercal\mu_a$ and $\mathbf{w}^\intercal\mu_c > \mathbf{w}^\intercal\mu_b$, which is the intersection of the half-spaces $\mathbf{w}^\intercal(1+\delta, -1) < 0$ and $\mathbf{w}^\intercal(1, -1) < 0$. These regions are illustrated in Figure 1 for different values of δ. Informally, as δ is decreased, the lines separating (π_a, π_c) and

(π_a, π_b) move closer to each other (as shown for $\delta = 0.25$), while as δ is increased, these lines move away from each other (as shown for $\delta = 10$).

Formally, let R_δ denote the region of $\mathbf{w}$ for which π_a is optimal (i.e. the blue region in the figures), that is,

$$R_\delta = \left\{ \mathbf{w} : \frac{w_1}{1+\delta} < w_2 < w_1(1+\delta) \right\}.$$

This is bounded below by the line $w_1 = (1+\delta)w_2$, which makes an angle of $\theta_\delta = \text{Tan}^{-1}(\frac{1}{1+\delta})$ with the x-axis, and bounded above by the line $w_2 = (1+\delta)w_1$, which makes an angle of θ_δ with the y-axis. We first show that for any value of δ, the regions of π_b and π_c have the exact same probability. The probability that π_b is optimal is the probability of the orange region which is

$$\begin{aligned}
\Pr(\pi_b \text{ is optimal}) &= \int_{-\infty}^{0} \int_{-\infty}^{w_1} \Pr(\mathbf{w}) dw_2 dw_1 + \int_{0}^{\infty} \int_{-\infty}^{\frac{w_1}{(1+\delta)}} \Pr(\mathbf{w}) dw_2 dw_1 \\
&= \int_{-\infty}^{0} \int_{-\infty}^{t_2} \Pr(t_2, t_1) dt_1 dt_2 + \int_{0}^{\infty} \int_{-\infty}^{\frac{t_2}{(1+\delta)}} \Pr(t_2, t_1) dt_1 dt_2 \\
&= \int_{-\infty}^{0} \int_{-\infty}^{t_2} \Pr(t_1, t_2) dt_1 dt_2 + \int_{0}^{\infty} \int_{-\infty}^{\frac{t_2}{(1+\delta)}} \Pr(t_1, t_2) dt_1 dt_2 \\
&= \Pr(\pi_c \text{ is optimal}),
\end{aligned}$$

where the second equality holds by changing the variables as $t_1 = w_2$ and $t_2 = w_1$, and the third one holds because the noise distribution is i.i.d. across the coordinates. Hence, we have

$$\Pr(\pi_b \text{ is optimal}) = \Pr(\pi_c \text{ is optimal}) = \frac{1 - \Pr(R_\delta)}{2},$$

as R_δ denotes the region where π_a is optimal.

Finally, we show that there exists a value of δ such that $\Pr(R_\delta) = 1/3$. Observe that as $\delta \to 0$, the lines bounding the region R_δ make angles that approach $Tan^{-1}(1) = \pi/4$ and the two lines touch, causing the region to have zero probability. On the other hand, as $\delta \to \infty$, the angles these lines make approach $Tan^{-1}(0) = 0$, so the region coincides with the first quadrant in the limit. Based on our selection of v_o, the probability of this region is exactly $1/2$. Hence, as δ varies from 0 to ∞, the probability of the region R_δ changes from 0 to $1/2$. Next, note that as $\theta_\delta = \text{Tan}^{-1}(\frac{1}{1+\delta})$, this angle changes continuously as δ changes, and hence does the region R_δ. Finally, as the noise distribution is continuous, the probability of this region R_δ also changes continuously as δ is varied. That is, $\lim_{\epsilon \to 0} \Pr(R_{\delta+\epsilon}) = \Pr(R_\delta)$. Coupling this with the fact that $\Pr(R_\delta)$ changes from 0 to $1/2$ as δ changes from 0 to ∞, it follows that there exists a value of δ in between such that $\Pr(R_\delta)$ is exactly $1/3$. Denote this value of δ by δ_o.

We conclude that for $\mathbf{w}^\star = (v_o, v_o)$ and our MDP construction with $\delta = \delta_o$, $\mathcal{P}(\mathbf{w}^\star) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

D.2 Completing the Proof

Consider the same MDP as in Section D.1. However, for this example, let the feature expectations be

$$\begin{aligned}
\phi(s,a) &= [0.5, 0.5 \ , \ -\delta_o/2, 1], \\
\phi(s,b) &= [1, -\delta_o/2, \ \ 0.5, 0.5], \\
\phi(s,c) &= [-\delta_o/2, 1, \ \ 1, -\delta_o/2],
\end{aligned}$$

where δ_o is as defined in Section D.1. Hence, the feature expectations of the policies $\{\pi_a, \pi_b, \pi_c\}$ are respectively

$$\begin{aligned}
\mu_a &= \frac{1}{2(1-\gamma)}[1, 1 \quad , \ -\delta_o, 2], \\
\mu_b &= \frac{1}{2(1-\gamma)}[2, -\delta_o, \ \ 1, 1], \\
\mu_c &= \frac{1}{2(1-\gamma)}[-\delta_o, 2, \ \ 2, -\delta_o].
\end{aligned}$$

Consider two weight vectors $\mathbf{w}_a^\star = (v_o, v_o, 0, 0)$ and $\mathbf{w}_b^\star = (0, 0, v_o, v_o)$, where v_o is as defined in Section D.1. Since $\mathbf{w}_a^\star$ completely discards the last two coordinates, it immediately follows from the example of Section D.1 that $\mathcal{P}(\mathbf{w}_a^\star) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Similarly, the same analysis on the last two coordinates shows that $\mathcal{P}(\mathbf{w}_b^\star) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ as well. On the other hand, the optimal policy according to $\mathbf{w}_a^\star$ is π_a while the optimal policy according to $\mathbf{w}_b^\star$ is π_b. Hence, $\pi_a^\star \neq \pi_b^\star$, but we still have $\mathcal{P}(\mathbf{w}_a^\star) = \mathcal{P}(\mathbf{w}_b^\star)$, leading to non-identifiability. □

E Proof of Theorem 4.4

The proof of this theorem strongly relies on Lemma 4.3 and the example used to prove it. Consider the MDP as in Section D.2, but now with 6 features instead of just 4. In particular, let the feature expectations of the three policies be

$$\begin{aligned}
\phi(s,a) &= [0.5, 0.5 \ \ , \ -\delta_o/2, 1, \ \ 1, -\delta_o/2], \\
\phi(s,b) &= [1, -\delta_o/2, \ \ 0.5, 0.5 \ \ , \ -\delta_o/2, 1], \\
\phi(s,c) &= [-\delta_o/2, 1, \ \ 1, -\delta_o/2, \ \ 0.5, 0.5 \ ].
\end{aligned}$$

Hence, the feature expectations of the policies $\{\pi_a, \pi_b, \pi_c\}$ are respectively

$$\begin{aligned}
\mu_a &= \frac{1}{2(1-\gamma)}[1, 1 \ \ \ , \ -\delta_o, 2, \ \ 2, -\delta_o], \\
\mu_b &= \frac{1}{2(1-\gamma)}[2, -\delta_o, \ \ 1, 1 \ \ \ , -\delta_o, 2], \\
\mu_c &= \frac{1}{2(1-\gamma)}[-\delta_o, 2, \ \ 2, -\delta_o, \ 1, 1 \ \ ].
\end{aligned}$$

Consider three weight vectors

$$\begin{aligned}
\mathbf{w}_a^\star &= (v_o, v_o, 0, 0, 0, 0), \\
\mathbf{w}_b^\star &= (0, 0, v_o, v_o, 0, 0), \\
\mathbf{w}_c^\star &= (0, 0, 0, 0, v_o, v_o).
\end{aligned}$$

Since $\mathbf{w}_a^\star$ completely discards the last four coordinates, the example of Section D.1 shows that $\mathcal{P}(\mathbf{w}_a^\star) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Similarly, the same analysis on the middle two and last two coordinates shows that $\mathcal{P}(\mathbf{w}_b^\star) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $\mathcal{P}(\mathbf{w}_c^\star) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, respectively. However, the optimal policy according to $\mathbf{w}_a^\star$ is π_a, according to $\mathbf{w}_b^\star$ it is π_b, and according to $\mathbf{w}_c^\star$ it is π_c.

Now, consider an arbitrary algorithm $\mathcal{A}$, which takes as input a distribution over policies and outputs a (possibly randomized) policy. Look at the randomized policy $\mathcal{A}(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ returned by $\mathcal{A}$ when the input is $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, and let p_a, p_b, p_c be the probabilities it assigns to playing π_a, π_b and π_c. Let p_i (where $i \in \{a, b, c\}$) denote the smallest probability among the three. Then, $p_i \leq 1/3$. Pick the ground truth weight vector to be $\mathbf{w}_i^\star$. As $\mathcal{P}(\mathbf{w}_a^\star) = \mathcal{P}(\mathbf{w}_b^\star) = \mathcal{P}(\mathbf{w}_c^\star)$, the data generated by $\mathbf{w}_i^\star$ follows the distribution $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, and the policy distribution chosen by $\mathcal{A}$ is simply (p_a, p_b, p_c).

Now, with probability $p_i \leq 1/3$, the policy played is π_i leading to a reward of $\mathbf{w}_i^{\star\intercal}\mu_i = \frac{v_o}{(1-\gamma)}$, and with probability $(1 - p_i)$, the policy played is some π_j (where $j \neq i$) leading to a reward of $\mathbf{w}_i^{\star\intercal}\mu_j = \frac{(2-\delta_o)}{2}\frac{v_o}{(1-\gamma)}$ (which is independent of the value of j).[5] Hence, the expected reward of algorithm $\mathcal{A}$ in this case is

$$\begin{aligned}
p_i \cdot \frac{v_o}{(1-\gamma)} + (1-p_i) \cdot \frac{(2-\delta_o)}{2}\frac{v_o}{(1-\gamma)} &= \frac{(2-\delta_o)}{2}\frac{v_o}{(1-\gamma)} + p_i \cdot \frac{\delta_o}{2}\frac{v_o}{(1-\gamma)} \\
&\leq \frac{(2-\delta_o)v_o}{2(1-\gamma)} + \frac{\delta_o v_o}{6(1-\gamma)}.
\end{aligned}$$

Observe that the uniform mixture π^u in this case is just the input distribution $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Whatever be the chosen $\mathbf{w}_i^\star$, the expected reward of this distribution is exactly

$$\frac{1}{3} \cdot \frac{v_o}{(1-\gamma)} + \frac{2}{3} \cdot \frac{(2-\delta_o)}{2}\frac{v_o}{(1-\gamma)} = \frac{(2-\delta_o)v_o}{2(1-\gamma)} + \frac{\delta_o v_o}{6(1-\gamma)},$$

[5] An interesting point to note is that by carefully selecting v_o, one could get the corresponding δ_o to be arbitrarily large, thereby causing the optimal and suboptimal policies to have a much larger gap (equally affecting the uniform mixture π^u as well).

which is nothing but the upper bound on the expected reward of $\mathcal{A}$. Hence, for any algorithm $\mathcal{A}$ there exists a ground truth weight vector $\mathbf{w}_i^\star$ such that $\mathcal{A}$ has an expected reward at most that of π^u (which in turn is strictly suboptimal). $\square$