Robust Reward Design for Markov Decision Processes

SHUO WU, University of Illinois Chicago, USA
HAOXIANG MA, University of Florida, USA
JIE FU, University of Florida, USA
SHUO HAN, University of Illinois Chicago, USA

The problem of reward design examines the interaction between a *leader* and a *follower*, where the leader aims to shape the follower's behavior to maximize the leader's payoff by modifying the follower's reward function. Current approaches to reward design rely on an accurate model of how the follower responds to reward modifications, which can be sensitive to modeling inaccuracies. To address this issue of sensitivity, we present a solution that offers robustness against uncertainties in modeling the follower, including 1) how the follower breaks ties in the presence of nonunique best responses, 2) inexact knowledge of how the follower perceives reward modifications, and 3) bounded rationality of the follower. Our robust solution is guaranteed to exist under mild conditions and can be obtained numerically by solving a mixed-integer linear program. Numerical experiments on multiple test cases demonstrate that our solution improves robustness compared to the standard approach without incurring significant additional computing costs.

JAIR Associate Editor: Mikko Koivisto

JAIR Reference Format:
Shuo Wu, Haoxiang Ma, Jie Fu, and Shuo Han. 2025. Robust Reward Design for Markov Decision Processes. *Journal of Artificial Intelligence Research* 84, Article 3 (September 2025), 46 pages. DOI: 10.1613/jair.1.19154

1 Introduction and Background

The problem of reward design is concerned with interactions between two types of players, a *leader* and a *follower*. The leader aims to induce the follower to behave in a desirable way by modifying the follower's reward function. The terms "leader" and "follower" derive from Stackelberg games (Hicks, 1935) to describe the asymmetric roles of the players: The leader always modifies the reward before the followers take action. As an example of reward design, imagine a class consisting of a teacher playing the role of the leader and a group of students playing the role of the follower. The teacher creates a grading policy that aims to motivate the students to achieve better learning outcomes. If the policy determines the final grade of a student solely based on the performance of the final exam, the student might neglect homework and class attendance, opting to cram for the final exam the night before. In contrast, a more effective grading policy is to include homework and attendance in the evaluation, which encourages students to learn on a regular basis. Another example arises in the training of autonomous agents via reinforcement learning, in which the agents are trained to maximize a predefined reward function. However, if the reward function is not carefully designed, it may be exploited by the agents to achieve high rewards in unintended ways, a phenomenon known as reward hacking (Pan et al., 2022). This can lead to behaviors that do not align with human values.

Authors' Contact Information: Shuo Wu, ORCID: 0009-0001-8561-062X, swu99@uic.edu, Department of Electrical and Computer Engineering, University of Illinois Chicago, Chicago, Illinois, USA; Haoxiang Ma, ORCID: 0000-0002-3823-385X, hmaalex21@gmail.com, Department of Electrical and Computer Engineering, University of Florida, Gainesville, Florida, USA; Jie Fu, ORCID: 0000-0002-4470-2827, fujie@ufl.edu, Department of Electrical and Computer Engineering, University of Florida, Gainesville, Florida, USA; Shuo Han, ORCID: 0000-0003-2204-6256, hanshuo@uic.edu, Department of Electrical and Computer Engineering, University of Illinois Chicago, Chicago, Illinois, USA.

DOI: 10.1613/jair.1.19154

We focus on the case where the sequential decision-making problem of the follower is modeled by a Markov decision process (MDP). Specifically, the follower faces a sequential decision-making problem and aims to maximize the cumulative reward. Such a setting is motivated by problems in cybersecurity. In a network, an attacker, who acts as the follower in our problem setting, seeks to compromise a set of target hosts to acquire confidential data or to disrupt the network operation. To achieve the objective, the attacker must breach a sequence of hosts within the network by exploiting causal and logical dependencies between vulnerabilities and carrying out sequential attacks. The planning of the sequential attacks is captured using deterministic or probabilistic attack graphs (Frigault et al., 2008; Hewett and Kijsanayothin, 2008; Jha et al., 2002; Nguyen et al., 2018), which can be viewed as a special instance of MDPs. A defender, who acts as the leader in this setting, aims to protect the network by allocating fake hosts or honey-patching vulnerabilities (Qin et al., 2023) and, consequently, mislead the attacker by manipulating the attacker's perceived rewards/payoffs.

The reward design problem can be formulated as a Stackelberg game (Ben-Porat et al., 2024; Carroll and Grosu, 2009; Chakraborty et al., 2024; Chen et al., 2022; Ma et al., 2023; Zhang and Parkes, 2008), for which algorithmic solutions are available. Nevertheless, existing solutions to reward design suffer from three issues of sensitivity to uncertainties in the model of the follower. First, the designed reward may be sensitive to tie-breaking of the follower. When the best response of the follower is not unique under the designed reward, the leader's payoff may depend on the follower's choice, as illustrated in Example 2 and Example 3 in this paper. Second, when the leader does not know the exact reward function of the follower, the leader's payoff may drop significantly upon small errors in modeling the follower's reward function. This is because the follower's response does not vary continuously as his reward changes. Lastly, when the follower is irrational and does not play a best response, the leader's payoff may be sensitive to the level of suboptimality of the follower's response (Table 5 and Table 6).

Solutions to these three issues of sensitivity remain under-explored. For achieving robustness to nonunique best responses, one notable folklore solution is to slightly perturb the optimal strategy of the leader (von Stengel and Zamir, 2004). However, it remains unclear how such a perturbation can be determined algorithmically. The method of perturbation may also fail in the presence of a budget constraint on reward allocation, as illustrated in Example 2. The closest work to ours is the study of robust Stackelberg equilibria in Gan et al. (2023), which addresses the issue of sensitivity to irrational followers. In their setting, the follower may play any δ-optimal response, i.e., a response that yields a follower's payoff within δ from optimality, whereas the leaders aims to find a robust strategy that maximizes the worst-case payoff against all possible δ-optimal responses. They discussed the existence of such a robust strategy and showed how the robust strategy can be computed in theory. In comparison, our work uses the quantal response model (Luce, 1959), which is a probabilistic model for characterizing the irrationality of the follower. Interestingly, our proposed robust solution is found to relate to the robust solution studied in Gan et al. (2023). Proposition 25 shows that our robust solution is nearly robust optimal against δ-optimal responses of the follower.

Main results. The main results of this paper are summarized below:

- We formulate the reward design problem as a Stackelberg game and identify the issue of sensitivity to illustrate the need for obtaining a robust solution. When the follower's best response is nonunique, we give examples showing that the leader's payoff can be sensitive to how the follower breaks ties (Example 2). While a folklore solution in the case of normal-form Stackelberg games is to induce a unique best response of the follower by perturbing the leader's strategy, we show that a unique best response may not be inducible in reward design for MDPs, even for very common cases (Example 3).
- We propose a robust solution, named an *optimal interior-point allocation*, for the reward design problem. It is an optimal solution whose neighboring points are also optimal, i.e., an interior point within the region of optimal solutions. An optimal interior-point allocation is provably robust in three aspects under a mild assumption (Assumption 5), without which the robustness can be inherently unachievable. First,

the solution is robust to the follower's tie-breaking choices (Proposition 8). Second, the solution offers robustness when the leader does not know the follower's reward function precisely (Proposition 9). Third, the solution is robust when the follower's decision-making deviates from rationality, a phenomenon often referred to as *bounded rationality*; the robustness guarantee holds under three different models of bounded rationality, including two based on quantal response models (Propositions 10 and 24), and one based on δ-optimal responses (Proposition 25).

- We prove that the existence of an optimal interior-point allocation only depends on the reward allocation budget of the leader (Theorem 16). Specifically, an optimal interior-point allocation is guaranteed to exist when the leader can achieve the optimal payoff without exhausting the reward allocation budget in the optimistic case, i.e., when the follower breaks ties in favor of the leader.
- Moreover, we prove that the existence of an optimal interior-point allocation is not only sufficient but also necessary for achieving robustness to tie-breaking of the follower (Theorem 17). This establishes the pivotal role of interior-point allocations in achieving robustness. It further motivates the development of algorithms for finding an optimal interior-point allocation.
- We show that an optimal interior-point allocation can be computed from a mixed-integer linear program (MILP). Numerical experiments were conducted to validate the robustness of the solution and to evaluate the computational cost on problems of practical interest (e.g., defending against cyberattacks). Compared to the standard solution to reward design given by Ma et al. (2023), our solution was found to improve robustness in all the test cases while maintaining a reasonable computational cost.

2 Related Work

Reward design. The challenge of selecting an appropriate reward function to achieve a desired outcome is commonly referred to as *reward design* or *reward shaping*. For reinforcement learning, providing rewards in strategically chosen states has been shown to facilitate the learning process by mitigating issues associated with sparse rewards (Ng et al., 1999). Early work on reward design in MDPs focuses on policy invariance, which examines how reward modifications can be made without altering the set of optimal policies. Ng et al. (1999) introduced the notion of potential-based reward shaping and showed guaranteed policy invariance in MDPs. Potential-based reward shaping was later extended to multi-agent systems for preserving the set of Nash equilibria (Devlin and Kudenko, 2011).

Reward design has also been studied in a leader-follower setup (Ben-Porat et al., 2024), where the goal is to incentivize the follower to act in the leader's interest. The setting is also known as *policy teaching* (Banihashem et al., 2022; Zhang and Parkes, 2008; Zhang et al., 2009b) and can be viewed as a special case of *model design* (Chen et al., 2022; Thoma et al., 2024) and *environment design* (Yu and Ho, 2022; Zhang et al., 2009a), in which elements (e.g., the transition kernel) other than the reward function of the MDP can also be modified. For solving the reward design problem in a leader-follower setup, Zhang and Parkes (2008) established the NP-hardness of the problem and provided a solution based on mixed-integer programming. More recently, Ben-Porat et al. (2024) gave a polynomial-time approximation algorithm. A special case is when the leader aims to induce a specific policy rather than a policy that maximizes the leader's payoff. For this case, Zhang et al. (2009b) showed that the desired reward function can be obtained by finding a feasible solution to a linear program.

Principal-agent problem. In the reward design problem, the leader and the follower are sometimes called the *principal* and the *agent*, respectively, which are terms originated from economics (Grossman and Hart, 1983; Ross, 1973). In the principal-agent problem, an agent takes actions on behalf of the principal through a contract designed by the principal (Bolton and Dewatripont, 2005; Gan et al., 2024; Hart and Holmström, 1987; Salanié, 2005). Two key aspects of the principal-agent problem are *adverse selection* and *moral hazard* (Myerson, 1982). The former refers to a situation in which the agents have private information that the principal cannot readily

access; the latter refers to the situation in which the agents take private actions that the principal cannot directly control or observe. The problem of reward design resembles the principal-agent problem; for instance, the policy used by the follower in the reward design problem cannot be enforced by the leader. However, there are some key technical differences in their mathematical formulations. In the principal-agent problem, the private information of the agent is captured by the type of the agent, which is assumed to follow a known probability distribution. In addition, the set of strategies played by the agent is assumed to be finite. Neither of these assumptions hold in the robust reward design problem considered in this paper. The unknown reward function is not drawn from a distribution, and the strategy space of the follower consists of all state-dependent policies and is thus uncountably infinite.

The principal-agent problem has also been studied in control theory, where it is commonly known as *incentive design* (Ho et al., 1982); see recent survey papers such as Ratliff et al. (2019) and Başar (2024). In incentive design, a decision-maker needs to determine a reward-based incentive strategy to encourage a desired behavior of another agent. The incentive strategy plays a similar role as the contract in the principal-agent problem. The form of optimal incentive strategies was studied in Başar (1982); Zheng and Başar (1982) for deterministic decision problems and in Başar (1984) for stochastic decision problems. Similar to moral hazard in the principal-agent problem, some settings of incentive design assume that the agent may influence the state of the environment through actions and that the decision-maker only has access to partial information of the state.

Robust solution to Stackelberg games. The reward design problem can be viewed as a Stackelberg game or, mathematically, a bilevel optimization problem (Ben-Porat et al., 2024; Chakraborty et al., 2024; Chen et al., 2022; Ma et al., 2023). Stackelberg games were first introduced by Hicks (1935) to demonstrate the benefits of leadership in games involving two parties. von Stengel and Zamir (2004) extended the strategy space in the model from pure strategies to mixed strategies. When the number of pure strategies is finite, Conitzer and Sandholm (2006) showed that an optimal strategy of the leader can be computed in polynomial time by solving multiple linear programs. Our results are inspired by the alternative solution presented by Paruchuri et al. (2008), who showed that an optimal strategy for the leader can be computed from a single MILP.

In a Stackelberg game, the influence of nonunique best responses of the follower on the leader's payoff is captured by the notions of strong and weak Stackelberg equilibria (Breton et al., 1988). In a strong Stackelberg equilibrium, the leader plays optimally assuming that the follower breaks ties in favor of the leader. In contrast, in a weak Stackelberg equilibrium, the leader assumes that the follower breaks ties adversarially. Consequently, the leader will play a strategy that is robust to tie-breaking. Nevertheless, a weak Stackelberg equilibrium may not exist in general (Dempe and Zemkoho, 2020, Section 4.3.2).

Robustness to the unknown follower's reward has been studied by Cansever and Başar (1982); Cansever and Başar (1985). The results are based on analyzing the local sensitivity of the leader's payoff, which is characterized by derivatives of the payoff function with respect to unknown parameters in the follower's reward function. In comparison, our work uses the notion of *robustness margin* and can guarantee non-local robustness against bounded modeling errors of the follower's reward function.

Robustness to irrational followers has been studied by Gan et al. (2023), where the objective is to find an optimal leader strategy against a follower who deterministically selects any near-optimal response. This is one of the three models of irrational followers studied in our work. Specifically, the model of irrationality used by Gan et al. (2023) is discussed in Appendix C.3, where we show that any optimal interior-point allocation is robust to arbitrary near-optimal responses from the follower. Theorem 2 of Gan et al. (2023) shows that even approximating a robust leader strategy under their near-optimal response model is computationally hard. This is consistent with our finding that computing an optimal interior-point allocation can be formulated as an MILP, which is known to be NP-hard to solve in general.

Robust control of MDPs. Robust control of MDPs aims to find reliable policies when precise knowledge of the environment model, including the transition probabilities and the reward function, is unavailable. As one of the earliest results, Nilim and El Ghaoui (2005) studied robust control of tabular MDPs with uncertain transitions. They introduced a minimax framework that leads to the *robust dynamic programming* algorithm for computing worst-case optimal policies under both time-invariant and time-varying unknown transitions. When the set of unknown transition matrices takes certain special forms, the computational complexity of their algorithm is similar to that of the standard dynamic programming algorithm for solving MDPs. Lim et al. (2013) extended the results to reinforcement learning, enabling agents to learn robust policies through interaction with the environment without fully knowing the uncertainty a priori. By bridging robust control and reinforcement learning, their work reduces the conservatism of robust control through learning while ensuring policy reliability in uncertain environments. The main differences between robust control and the robust reward design problem studied in this paper are in what the decision-making agent can control and the kind of uncertainty faced by the agent: In robust control, the agent chooses the decision-making policy, and the uncertainty is in the transition probabilities or the reward function. In comparison, in the robust reward design problem, the agent (i.e., the leader) chooses how to modify the follower's reward function, and the uncertainty is in how the follower responds to the modified reward.

Bayesian persuasion. When the follower has incomplete information, besides changing the reward function of the follower, another way of changing the follower's behavior is to design a *signaling scheme*, which determines how information is disclosed to the follower. Bayesian persuasion provides a game-theoretic framework for understanding how strategic information disclosure shapes the follower's beliefs and consequently influences the follower's decision-making (Kamenica and Gentzkow, 2011). Gan et al. (2022) introduced Bayesian persuasion to sequential decision-making and considered both myopic and farsighted agents in MDPs. Wu et al. (2022) further introduced the concept of Markov Persuasion Process and designed an online algorithm that learns an optimal signaling policy from interactions. Bernasconi et al. (2023) demonstrated that Markovian signaling schemes are suboptimal for influencing farsighted followers and argued that history-dependent signaling schemes should be considered. To avoid the exponential growth in the complexity of representing general history-dependent signaling schemes, they introduced a subclass of history-dependent signaling schemes, named *promise-form*, which is both optimal and efficiently representable. Furthermore, they showed that an optimal promise-form signaling scheme can be computed in polynomial time.

Models of irrationality. Several models have been used in the literature to characterize irrational behaviors of a decision-making agent. Of particular interest is modeling irrational behaviors due to limited reasoning capability of the agent, commonly known as *bounded rationality* (Simon, 1955). Our work adopts the *quantal response* model, which assumes that the agent may play suboptimal strategies with a probability that decreases exponentially as the corresponding payoff decreases. The quantal response model in economics intends to model how the selection probabilities of an individual depend on explanatory variables (McFadden, 1976). In game-theoretic settings, quantal response was used to model the behaviors of players in normal-form games (McKelvey and Palfrey, 1995) and in Stackelberg games (Yang et al., 2012). The resulting equilibrium is known as the *quantal response equilibrium*. Besides quantal response, another popular model of bounded rationality is the *cognitive hierarchy* model (Camerer et al., 2004). The model uses iterative decision rules to represent strategic players with different levels of sophistication and can be used to explain nonequilibrium behaviors of the players. Lastly, one may assume that the agent plays any suboptimal strategy that leads to a payoff close to optimality (Gan et al., 2023).

Comparison with our work. For clarity, Table 1 summarizes the comparison between our work and related studies.

Table 1. Comparison of related work with our proposed framework.

Related Work	Comparison with Our Work
Reward Design	Existing studies do not consider the uncertainty in the follower's behavior, which is addressed in our work.
Principal-Agent Problem	The leader-follower setup in our model shares similarities with that in the principal-agent problem. However, in our model, the uncertainty in the follower's reward is non-probabilistic, and the strategy space of the follower is uncountably infinite. Neither of these conditions holds for the principal-agent problem in the existing literature.
Robust Stackelberg Game	Our proposed notion of robustness is inspired by similar notions in Stackelberg Games. However, reward design in MDPs introduces additional complexities due to the state transition dynamics experienced by the follower.
Robust Control of MDPs	Robust control of MDPs considers a sequential decision-making agent who faces uncertainties in the transition probabilities or their own reward. In contrast, robust reward design for MDPs considers modifying the reward of a sequential decision-making agent whose behavior is uncertain due to arbitrary tie-breaking or unknown reward.
Bayesian Persuasion	In Bayesian persuasion, an agent's decision is influenced by belief shaping through controlled information disclosure. In reward design, the influence is achieved by modifying the agent's reward function.
Models of Irrationality	The quantal response model and the near-optimal response model used in our work to characterize bounded rationality can be found in the existing literature.

3 Preliminaries

This section introduces the basic setup of the problem, including the reward design problem for MDPs, its reformulation via occupancy measures, and an optimization problem for computing optimal reward allocations. We also summarize the key notational conventions used throughout the paper.

3.1 Reward Design for MDPs

The decision-making process of the follower is modeled as an MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma, \rho)$. The set $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and $\rho \in \Delta(\mathcal{S})$ is the initial distribution, where $\Delta(\mathcal{S})$ denotes the set of probability distributions over $\mathcal{S}$. The mappings $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ and $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ represent the transition kernel and the reward function, respectively. We sometimes also abuse the notation and view r as a vector of dimension $|\mathcal{S}||\mathcal{A}|$. The constant $\gamma \in (0, 1)$ is the discount factor. The policy of the follower is denoted by $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$, where $\pi(s, \cdot) \in \Delta(\mathcal{A})$ for all $s \in \mathcal{S}$. The set of all policies is denoted by Π.

The leader may modify the reward function of the follower by allocating rewards in the environment. Let $\mathcal{D} = \{(s^{(1)}, a^{(1)}), \ldots, (s^{(K)}, a^{(K)})\} \subseteq \mathcal{S} \times \mathcal{A}$ be the nonempty set of candidate state-action pairs at which reward can be allocated. The reward allocation is described by a function $x : \mathcal{D} \to \mathbb{R}$, where $x(s^{(i)}, a^{(i)})$ is the amount of reward allocated at $(s^{(i)}, a^{(i)})$. For convenience, we sometimes abuse the notation and treat x as a vector of

dimension $|\mathcal{D}|$ such that $x_i = x(s^{(i)}, a^{(i)})$ is the ith component of x. A reward allocation x is said to be *admissible* if x satisfies the following conditions:

- The total amount of allocated resource cannot exceed a given budget $C > 0$: $\sum_{i=1}^{|\mathcal{D}|} x_i \le C$.
- The amount of resource allocated to each state is nonnegative: $x_i \ge 0$, $i = 1, \ldots, |\mathcal{D}|$.

The set of admissible reward allocations is denoted by $\mathcal{X} = \left\{x \mid \sum_{i=1}^{|\mathcal{D}|} x_i \le C,\ x \succeq 0\right\}$, where $\succeq$ denotes the entrywise inequality between two vectors. An admissible reward allocation x induces a new reward function r_2^x of the follower, where $r_2^x(s,a) = r(s,a) + x(s,a)$ for any $(s,a) \in \mathcal{D}$, and $r_2^x(s,a) = r_2(s,a)$ when $(s,a) \in (\mathcal{S} \times \mathcal{A}) \setminus \mathcal{D}$. The reward function of the leader is denoted by $r_1 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which is assumed to be nonnegative and bounded. Without loss of generality, we assume that $0 \le r_1(s,a) \le 1$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Under a given policy π of the follower, the leader's value function is defined by $V_1^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t r_1(s_t, a_t) \mid s_0 = s\right]$, and the follower's value function is defined by $V_2^\pi(s;x) = \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t r_2^x(s_t, a_t) \mid s_0 = s\right]$. The expected payoffs of the leader and the follower are given by $\mathbb{E}_{s\sim\rho}\left[V_1^\pi(s)\right]$ and $\mathbb{E}_{s\sim\rho}\left[V_2^\pi(s;x)\right]$, respectively. The goal of the leader is to maximize her payoff by influencing the policy of the follower through the reward allocation x. Mathematically, this can be cast as the following optimization problem:

$$\begin{aligned} \underset{x,\pi}{\text{maximize}} \quad & \mathbb{E}_{s\sim\rho}\left[V_1^\pi(s)\right] \\ \text{subject to} \quad & \pi \in \arg\max_\pi \mathbb{E}_{s\sim\rho}\left[V_2^\pi(s;x)\right], \\ & x \in \mathcal{X}. \end{aligned} \tag{1}$$

We do not assume that $\arg\max_\pi \mathbb{E}_{s\sim\rho}\left[V_2^\pi(s;x)\right]$ is a singleton set. When the optimal policy of the follower is not unique under a reward allocation x, the problem formulation in (1) assumes that the follower will choose from the set of optimal policies in favor of the leader. This assumption will be revisited in Section 4 and remains a central focus throughout the paper.

3.2 Reformulation via the Occupancy Measure

While the expected payoff in an MDP is a nonlinear function of the policy π, it can be rewritten as a linear function by a change of variables. As shown in Section 5.1, writing the expected payoff as a linear function is useful in revealing important geometric structures of (1). The change of variables involves replacing the policy π with $m\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defined by

$$m(s,a) \triangleq \mathbb{E}_{\pi, s_0\sim\rho}\left[\sum_{t=0}^\infty \gamma^t \mathbb{P}(s_t = s, a_t = a)\right], \tag{2}$$

known as the *occupancy measure* induced by π. It can be shown that m is a valid occupancy measure if and only if $m(s,a) \ge 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and

$$\sum_{a\in\mathcal{A}} m(s,a) = \rho(s) + \gamma \sum_{s',a'} \mathcal{T}(s',a',s) m(s',a') \quad \forall s \in \mathcal{S}. \tag{3}$$

For convenience, we abuse the notation and treat m as a vector of dimension $|\mathcal{S}||\mathcal{A}|$, ρ as a vector of dimension $|\mathcal{S}|$, and succinctly write (3) as $Am = \rho$ for some matrix A. Denote the set of valid occupancy measures by

$$\mathcal{M} = \{m \mid m \succeq 0,\ Am = \rho\}.$$

For the direction from policies to occupancy measures, however, there could be many policies that induce the same occupancy measure.

Proposition 1. *Given any $m \in \mathcal{M}$, a policy π induces m if and only if*

$$\pi(s,a) = \begin{cases} m(s,a)/\sum_{a' \in \mathcal{A}} m(s,a') & \text{if } \sum_{a' \in \mathcal{A}} m(s,a') \neq 0, \\ \pi_0(s,a) & \text{otherwise} \end{cases} \tag{4}$$

for some policy π_0.

Proof. See Appendix A.1.1. □

Define $\langle r_2^x, m \rangle \triangleq \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} r_2^x(s,a) m(s,a)$. It can be shown (see Appendix A.2) that the expected payoff of the follower becomes a linear function of m:

$$\mathbb{E}_{s \sim \rho}\left[V_2^\pi(s;x)\right] = \langle r_2^x, m \rangle.$$

Similarly, the expected payoff of the leader satisfies $\mathbb{E}_{s \sim \rho}\left[V_1^\pi(s)\right] = \langle r_1, m \rangle$. For any reward allocation x, define the set of best responses of the follower to x by

$$\mathsf{BR}(x) \triangleq \arg\max_{m \in \mathcal{M}} \langle r_2^x, m \rangle.$$

Because $\mathcal{M}$ is a polyhedron, computing a best response of the follower is a linear program. Reformulating an MDP as a linear program is a well-established result. Readers may refer to Puterman (1994, Section 6.9) for details. Using the occupancy measure, the reward design problem in (1) can be equivalently reformulated as

$$\begin{array}{ll} \underset{x \in \mathcal{X}, m \in \mathcal{M}}{\text{maximize}} & \langle r_1, m \rangle \\ \text{subject to} & m \in \mathsf{BR}(x). \end{array} \tag{5}$$

We will hereafter denote the optimal value of (5) by $v_1^\star$ and an optimal solution of (5) by $(x^\star, m^\star)$, where $x^\star$ is called an *optimal allocation* and $m^\star$ an *optimal occupancy measure*.

3.3 Computing an Optimal Reward Allocation

The reward design problem in (5) can be solved by an MILP (Ma et al., 2023). In the following, we will briefly review the procedure for completeness. For a given reward allocation x, the condition $m \in \mathsf{BR}(x)$ is equivalent to that m is an optimal solution of the following problem:

$$\begin{array}{ll} \underset{m}{\text{maximize}} & \langle r_2^x, m \rangle \\ \text{subject to} & Am = \rho, \qquad m \succeq 0, \end{array} \tag{6}$$

where the constraints come from the condition $m \in \mathcal{M}$. Because problem (6) is a linear program and is always feasible, it is known that m is optimal if and only if there exists (a dual variable) v such that (m, v) satisfies the Karush–Kuhn–Tucker (KKT) conditions:

$$Am = \rho, \quad m \succeq 0, \quad A^T v - r_2^x \succeq 0, \quad m \perp A^T v - r_2^x. \tag{7}$$

The reward design problem in (5) can then be rewritten as

$$\underset{x,m,v}{\text{maximize}} \quad \langle r_1, m \rangle \tag{8a}$$

$$\text{subject to} \quad Am = \rho, \qquad m \succeq 0, \qquad A^T v - r_2^x \succeq 0, \tag{8b}$$

$$m \perp A^T v - r_2^x, \tag{8c}$$

$$x \in \mathcal{X}. \tag{8d}$$

The complementary slackness constraint in (8c) can be reformulated as affine constraints with integer variables (Vielma, 2015). Because other constraints are affine, and the objective is linear, problem (8) can be reformulated as an MILP.

3.4 Notation

A policy π is called *deterministic* if for any $s \in \mathcal{S}$, there exists $a \in \mathcal{A}$ such that $\pi(s,a) = 1$. A policy is called *randomized* if it is not deterministic. The set of all deterministic policies is denoted by Π_{det}. For any $\pi \in \Pi$, we sometimes denote by m^π the occupancy measure *induced* by π according to (2). An occupancy measure is called *deterministic* (resp. *randomized*) if it is induced by a deterministic (resp. randomized) policy. The set of deterministic occupancy measures is denoted by $\mathcal{M}_{\text{det}}$. For any $m \in \mathcal{M}$, denote the set of policies of the form (4) by $\Pi(m)$, which is the set of policies that induce m according to Proposition 1.

For a given (possibly randomized) policy π, we abuse the notation of Π_{det} and define

$$\Pi_{\text{det}}(\pi) \triangleq \{\pi' \in \Pi_{\text{det}} \mid \pi'(s,a) = 0 \text{ when } \pi(s,a) = 0\}.$$

In plain words, at any $s \in \mathcal{S}$, a policy $\pi' \in \Pi_{\text{det}}(\pi)$ is only allowed take an action that π takes with nonzero probability. The set of occupancy measures induced by policies in $\Pi_{\text{det}}(\pi)$ is denoted by $\mathcal{M}_{\text{det}}(\pi) \triangleq \{m^{\pi'} \mid \pi' \in \Pi_{\text{det}}(\pi)\}$ with an abuse of notation. Such a definition has a nice property that for any $\pi_1, \pi_2 \in \Pi(m)$, it holds that $\mathcal{M}_{\text{det}}(\pi_1) = \mathcal{M}_{\text{det}}(\pi_2)$. Interested reader may refer to Appendix A.1.2.

For a given reward function $r\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, let h be the mapping such that $h(r) : \mathcal{D} \to \mathbb{R}$ is the restriction of r to $\mathcal{D}$, i.e., $h(r)$ satisfies $h(r)(s,a) = r(s,a)$ for all $(s,a) \in \mathcal{D}$.

4 Sensitivity Issues of Optimal Allocations

In problem (5), the maximization over m implicitly assumes that the follower chooses a best response that maximizes the expected payoff of the leader. Under this assumption, the expected payoff received by the leader under an allocation x is given by the optimal value of the problem

$$\begin{array}{ll} \underset{m \in \mathcal{M}}{\text{maximize}} & \langle r_1, m \rangle \\ \text{subject to} & m \in \mathsf{BR}(x). \end{array} \tag{9}$$

In practice, however, the follower may choose arbitrarily from the set of best responses. In the worst case for the leader, the follower may choose a best response that is most unfavorable to the leader. In this case, the leader's expected payoff under the allocation x is given by the optimal value of the problem

$$\begin{array}{ll} \underset{m \in \mathcal{M}}{\text{minimize}} & \langle r_1, m \rangle \\ \text{subject to} & m \in \mathsf{BR}(x). \end{array} \tag{10}$$

We define the optimal value of (9) as the *optimistic value* of x, denoted by $\mathsf{OptiVal}(x)$, and the optimal value of (10) as the *pessimistic value* of x, denoted by $\mathsf{PessVal}(x)$. We also refer to $\sup_{x \in X} \mathsf{OptiVal}(x)$ as the *optimal optimistic value* and $\sup_{x \in X} \mathsf{PessVal}(x)$ as the *optimal pessimistic value*. In general, when the best response of the follower under an allocation x is not unique, the optimistic value $\mathsf{OptiVal}(x)$ and the pessimistic value $\mathsf{PessVal}(x)$ may differ. In addition, the optimal optimistic value and the optimal pessimistic value may also differ, as illustrated in Example 2. In all the examples, for simplicity, the rewards are assigned based on reaching certain states.

Example 2 (Optimistic and pessimistic values)**.** Consider a 4×4 grid world in Figure 1. The follower starts from the position $(1,2)$ and can move in any direction or stay in the same place by taking an action in $\mathcal{A} = \{\text{left}, \text{right}, \text{up}, \text{down}, \text{stay}\}$. The true goals are at $(4,1)$ and $(4,3)$. The leader is allowed to allocate the reward at $(3,3)$. To simplify the presentation, we choose a discount factor of 1 and require that the follower take a shortest path when multiple paths lead to the same cumulative reward. (The shortest-path requirement can be removed if

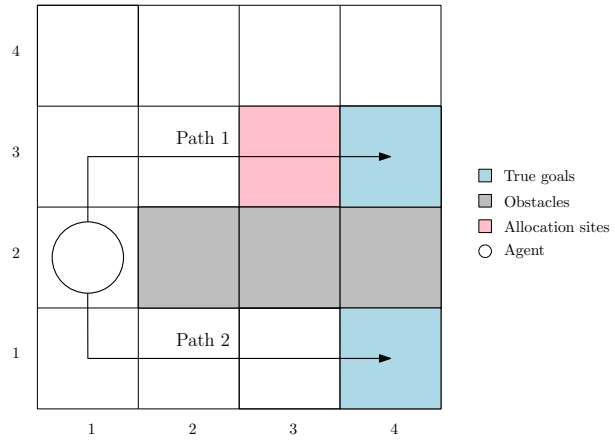

Fig. 1. Reward design problem with one state admits allocated reward. The reward for arriving at $(4, 1)$ is 3 and that for arriving at $(4, 3)$ is 2. At most 1 unit of resource can be allocated toward the reward at $(3, 3)$. The follower starts at $(1, 2)$, and all the transitions are deterministic.

.

the discount factor is chosen to be less than 1 and the allocation budget is modified accordingly.) The reward for arriving at $(4, 1)$ is 3, and that for arriving at $(4, 3)$ is 2. The allocation budget is 1, and $(4, 1)$ and $(4, 3)$ are absorbing states, meaning that the follower cannot leave the state once entering them. The leader will get a reward of 1 once the follower arrives at $(3, 3)$, the state with allocated reward.

First, consider the optimistic case where the follower breaks any tie in favor of the leader. Denote by x_1 the allocation strategy that spends the entire budget of 1 at $(3, 3)$. Under x_1, choosing Path 1 to arrive at $(3, 3)$ and $(4, 3)$ in 4 steps gives the follower a total payoff of 3. Path 2 has the same payoff for the follower because the true goal at $(4, 1)$ is in the path. Since the game ends in 4 steps, 3 is the maximum payoff that the follower can achieve, implying that two paths are the only best responses. The leader, however, will obtain different payoffs when the follower chooses different paths. If the follower reaches the allocated reward at $(3, 3)$ with probability 1 by following Path 1, the leader will receive a payoff of 1. In comparison, if the follower reaches the true goal at $(4, 1)$ by following Path 2, the leader will receive no payoff. Under the optimistic assumption, the follower will choose Path 1. Therefore, the leader's payoff for allocating all the budget at $(3, 3)$ is 1. For any other allocation strategy, the allocation at $(3, 3)$ is less than 1, and the follower's payoff for following Path 1 is less than the payoff for following Path 2. The follower will then follow Path 2, in which case the payoff for the leader is 0. Therefore, x_1 is an optimal allocation strategy, and the optimal optimistic value of the game is 1.

Next, consider the pessimistic case. It can be seen that the follower will always follow Path 2 under any admissible allocation. When the allocation at $(3, 3)$ is less than 1, the payoff for the leader remains the same as in the optimistic case because the best response of the follower is unique and is following Path 2. When the allocation at $(3, 3)$ is 1, the follower will still take Path 2 under the pessimistic assumption. This is because both paths are best responses of the follower, but Path 2 leads to a lower payoff for the leader. Therefore, the leader will always receive a payoff of 0 in the pessimistic case regardless of the allocation strategy, which implies that the optimal pessimistic value of the game is 0.

In summary, this example shows that the optimistic value and pessimistic value under x_1 are different: $\mathrm{OptiVal}(x_1) = 1$ and $\mathrm{PessVal}(x_1) = 0$. In addition, the optimal optimistic value of the game is 1, whereas the optimal pessimistic value of the game is 0.

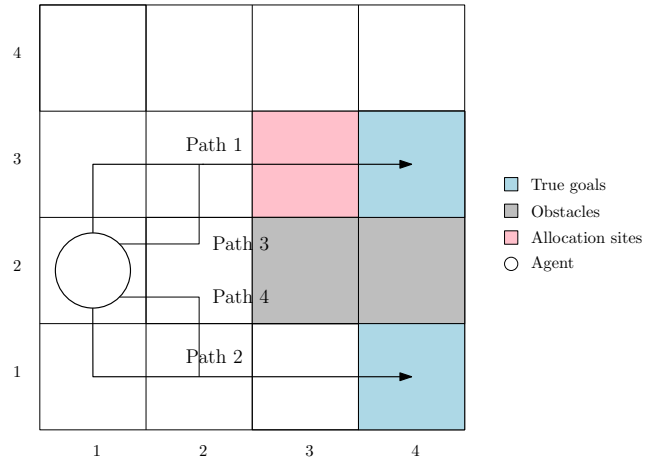

Fig. 2. Reward design problem with two obstacles. Other settings are similar to Example 2.

Besides the handcrafted setup in Example 2, our numerical experiments in Section 9.2 suggest that the optimal allocation obtained using the method in Section 3.3 also exhibits sensitivity to nonunique best responses. This motivates us to search for an optimal allocation $x^\star$ that is robust to nonunique best responses of the follower. Mathematically, this requires finding $x^\star$ that satisfies

$$\mathsf{OptiVal}(x^\star) = \mathsf{PessVal}(x^\star) = v_1^\star.$$

The issue of sensitivity to tie-breaking in Stackelberg games is a known issue in the literature, where a folklore solution (see page 12 in von Stengel and Zamir (2004)) is to perturb the optimal allocation slightly such that the best response becomes unique. Nevertheless, two issues prevent the immediate application of this solution. First, it remains unclear how such a perturbation should be obtained algorithmically. Indeed, when the perturbation is not chosen appropriately, the induced best response may be to the disadvantage of the leader and yield the pessimistic value. Second, the best response may remain nonunique regardless of how the optimal allocation is perturbed, as illustrated in Example 3.

Example 3 (Best response is never unique)**.** Consider the environment in Figure 2, which has a similar setting as Example 2 with the exception that one obstacle has been removed. Both Path 1 and Path 3 go through the allocated reward at $(3, 3)$ and the true goal at $(4, 3)$ but never reach the true goal at $(4, 1)$. Therefore, for the follower, the payoff for choosing Path 1 and Path 3 are always the same under any reward allocation. This implies that Path 1 can never be a unique best response. For a similar reason, for the follower, the payoff for choosing Path 2 and Path 4 are always the same under any reward allocation. This implies that Path 2 can never be a unique best response.

In the meantime, it is not difficult to see that either Path 1 or Path 2 must be a best response under any allocation. Therefore, the existence of Paths 3 and 4 shows that it is impossible to induce a unique best response regardless of the allocation strategy.

In general, a robust solution $x^\star$ satisfying $\mathsf{OptiVal}(x^\star) = \mathsf{PessVal}(x^\star) = v_1^\star$ may not always exist. One key factor influencing its existence is the set $\mathcal{D}$ of state-action pairs that allow reward allocation. If allocating reward at state-action pairs in $\mathcal{D}$ does not alter how the follower visits the state-action pairs of interest to the leader, then finding a robust solution may be impossible. The following example illustrates this scenario.

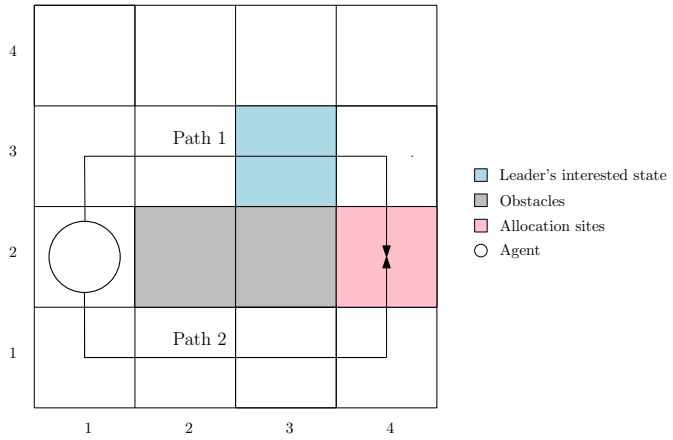

Fig. 3. The reward design problem that has no robust solution. The leader is interested in the state $(3, 3)$, and the leader's reward is 1 in this state and 0 in all other states. The reward allocation is only admissible at $(4, 2)$. There is no true goal for the follower.

Example 4 (Robust solution does not exist)**.** Consider a 4×4 grid world in Figure 3. The follower starts from $(1, 2)$. The action space of the follower and the transition kernel are the same as Example 2, where $\mathcal{A} = \{\text{left}, \text{right}, \text{up}, \text{down}, \text{stay}\}$, and the transitions are deterministic and follow the direction of the action. There is no true goal for the follower. The leader is interested in the state $(3, 3)$: When the follower enters $(3, 3)$, the leader receives a payoff of 1. The leader's reward is 0 in all other states. The reward allocation is only allowed at $(4, 2)$, and the allocation budget is 1. Once the follower enters $(4, 2)$, the follower will receive the reward allocated by the leader and cannot leave the state.

Suppose a positive reward is allocated at $(4, 2)$. In this case, both Path 1 and Path 2 lead the follower to $(4, 2)$, the only state that the follower can get rewarded. Consequently, regardless of the reward allocation, the follower remains indifferent between taking Path 1 or Path 2, and no solution is robust to the nonunique best responses of the follower.

To avoid the case that robustness is inherently impossible, an important assumption is imposed on the leader's reward function:

Assumption 5. *The support of r_1 is a subset of the state-action pairs that allow reward allocation, i.e.,* $\{(s, a) \in \mathcal{S} \times \mathcal{A} \mid r_1(s, a) \neq 0\} \subseteq \mathcal{D}$.

Intuitively, this assumption on the leader's reward function can be understood as a *controllability* condition. It requires that any state-action pair of interest to the leader must permit reward allocation, which can be used to steer the follower's behavior toward the leader's desired outcomes. It can be seen that Assumption 5 fails to hold in Example 4: For the state $(3, 3)$, the leader's reward is nonzero, but reward allocation is not permitted. Assumption 5 is required in almost all the subsequent results, except for those in Section 6.

5 Interior-Point Allocations Are Robust

We introduce the concept of an *optimal interior-point allocation*, which is a special type of optimal allocations, and show that any optimal interior-point allocation is robust to nonunique best responses of the follower. Namely, by using an optimal interior-point allocation, the leader is always guaranteed to receive the optimal expected payoff of the game regardless of how the follower breaks ties. Moreover, an optimal interior-point allocation is

able to provide other forms of robustness guarantees. One is robustness to uncertainty in the follower's reward function: It ensures that the leader still receives the optimal payoff without exactly knowing how the follower perceives the modified reward function. Another form is robustness to bounded rationality in the follower: It ensures that the leader does not lose much payoff when the follower starts to behave irrationally by responding with a slightly suboptimal policy.

5.1 Optimal Interior-Point Allocations

Since any allocation $x \in \mathbb{R}^{|\mathcal{D}|}$ corresponds to one or more best responses described by $\mathsf{BR}(x)$, it is possible to divide $\mathbb{R}^{|\mathcal{D}|}$ into (possibly overlapping) regions based on the corresponding best response. Each region is called an *allocation region* and is identified by a unique occupancy measure.

Definition 6 (Allocation region)**.** Let $m \in \mathcal{M}$ be an occupancy measure. The *allocation region* of m is defined by

$$\mathcal{P}_m = \{x \in \mathbb{R}^{|\mathcal{D}|} \mid m \in \mathsf{BR}(x)\} = \{x \in \mathbb{R}^{|\mathcal{D}|} \mid \langle r_2^x, m\rangle \geq \langle r_2^x, m'\rangle \text{ for all } m' \in \mathcal{M}\}.$$

Furthermore, the allocation region of an optimal occupancy measure is called an *optimal allocation region*.

The term allocation region highlights the fact that $\mathcal{P}_m$ is a subset of $\mathbb{R}^{|\mathcal{D}|}$, the space where allocation vectors live. For $x^\star$ to be an optimal allocation, it is necessary from Definition 6 that $x^\star$ belongs to some optimal allocation region. However, when $x^\star$ is on the boundary of an optimal allocation region, the best responses under $x^\star$ are not unique and may lead to undesirable sensitivity to tie-breaking. This motivates us to examine the interior points of an allocation region, which we refer to as *interior-point allocations*.

Definition 7 (Interior-point allocation)**.** Let $m \in \mathcal{M}$ be an occupancy measure. An allocation vector $x \in \mathcal{X}$ is called an *interior-point allocation of* $\mathcal{P}_m$ if there exists $c > 0$ such that

$$x + cv \in \mathcal{P}_m \quad \text{for all } \|v\|_1 \leq 1. \tag{11}$$

The largest constant c for (11) to hold is called the *margin* of x. An interior-point allocation of an optimal allocation region is called an *optimal interior-point allocation*.

Because all norms are equivalent in a finite-dimensional space, the choice of the norm in (11) does not affect the definition of interior-point allocation: An interior-point allocation in one norm must also be an interior-point allocation in another norm. The choice of $\|\cdot\|_1$ is for the convenience of computation. For computing an optimal interior-point allocation, the condition in (11) can be rewritten equivalently as finitely many constraints given by (17) in Section 8.

5.2 Robustness to Nonunique Best Responses

We will show that an optimal interior-point allocation $x^\star$ always yields the optimal value $v_1^\star$ of the game regardless of how the follower breaks ties. In other words, the pessimistic value of $x^\star$ is equal to the optimistic value of $x^\star$.

Proposition 8 (Robustness to nonunique best responses)**.** *Suppose that Assumption 5 holds. If $x^\star$ is an optimal interior-point allocation, then* $\mathsf{OptiVal}(x^\star) = \mathsf{PessVal}(x^\star) = v_1^\star$.

Proof. See Appendix B.1. □

The importance of an interior-point allocation can be understood as follows. Suppose that $x^\star$ is an optimal interior-point allocation of $\mathcal{P}_{m_1}$ for some optimal occupancy measure m_1. This implies $m_1 \in \mathsf{BR}(x^\star)$. When $\mathsf{OptiVal}(x^\star) \neq \mathsf{PessVal}(x^\star)$, there must exist $m_2 \in \mathsf{BR}(x^\star)$ with $\langle r_1, m_1\rangle > \langle r_1, m_2\rangle$. Let $x(\epsilon) = x^\star - \epsilon h(r_1)$, where $h(r_1)$ is the restriction of r_1 to $\mathcal{D}$ as defined in Section 3.4. It follows that $r_2^{x(\epsilon)} = r_2^{x^\star} - \epsilon r_1$. Because $m_1, m_2 \in \mathsf{BR}(x^\star)$, it holds that $\langle r_2^{x^\star}, m_1\rangle = \langle r_2^{x^\star}, m_2\rangle$. Thus, for any $\epsilon > 0$, it holds that $\langle r_2^{x(\epsilon)}, m_1\rangle < \langle r_2^{x(\epsilon)}, m_2\rangle$

or, equivalently, $x(\epsilon) \notin \mathcal{P}_{m_1}$. This implies that no neighborhood of $x^\star$ is contained in $\mathcal{P}_{m_1}$, which contradicts with the fact that $x^\star$ is an interior-point allocation of $\mathcal{P}_{m_1}$.

While Proposition 8 shows that finding an optimal interior-point allocation is sufficient for achieving robustness to nonunique best responses of the follower, the existence of an optimal interior-point allocation is, in fact, also *necessary* for robustness to nonunique best responses. We shall postpone the discussion on necessity until Theorem 17.

5.3 Robustness to Uncertain Reward Perception of the Follower

Besides robustness to nonunique best responses, one may be interested in other notions of robustness motivated by practical considerations. One such notion is robustness to uncertainty in how the follower perceives the modified reward function. The original formulation in (5) assumes that the modified reward function of the follower is exactly r_2^x when the allocation is x. However, the follower may perceive the modified reward differently as $r_2^{x+\delta}$, where $\delta \in \mathbb{R}^{|\mathcal{D}|}$ represents uncertainty in how the follower perceives reward modifications. In light of such uncertainty, one reasonable goal is to seek an allocation $x^\star$ that is robust to any uncertainty δ up to a certain magnitude.

The following proposition shows that an optimal interior-point allocation offers robustness to uncertainty in reward perception of the follower.

Proposition 9 (Robustness to reward perception of the follower)**.** *Suppose that Assumption 5 holds. If $x^\star$ is an optimal interior-point allocation with margin $c > 0$, then* $\mathsf{OptiVal}(x^\star + \delta) = \mathsf{PessVal}(x^\star + \delta) = v_1^\star$ *for all* $\|\delta\|_1 < c$.

Proof. Since $x^\star$ is an optimal interior-point allocation, there exists m such that $x^\star + cv \in \mathcal{P}_m$ for all $\|v\|_1 \leq 1$ and $\langle r_1, m\rangle = v_1^\star$. Then $x^\star + \delta$ is in the interior of $\mathcal{P}_m$ when $\|\delta\|_1 < c$. According to Proposition 8, $\mathsf{OptiVal}(x^\star + \delta) = \mathsf{PessVal}(x^\star + \delta) = v_1^\star$. □

It can be seen that Proposition 9 implies Proposition 8: By setting $\delta = 0$, the result in Proposition 9 recovers the one in Proposition 8. In other words, robustness to uncertainty in the follower's perceived reward is a stronger notion than robustness to nonunique best responses.

5.4 Robustness to a Boundedly Rational Follower

In practice, the follower may not be able to solve the MDP to optimality and may instead produce a response that is only near-optimal, a phenomenon known as *bounded rationality* (Simon, 1955). One common model of bounded rationality is *quantal response* (Luce, 1959), which assumes that the decision-maker takes suboptimal actions with probabilities that diminish exponentially as the corresponding payoffs decrease (McFadden, 1976). More concretely, under a reward allocation x, a boundedly rational follower attempts to choose a policy that maximizes the payoff

$$\langle r_2^x, m\rangle - \tau \cdot \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} m(s,a) \log \frac{m(s,a)}{\sum_{a'\in\mathcal{A}} m(s,a')}, \tag{12}$$

where m is the occupancy measure induced by the policy of the follower, $\tau > 0$ is a constant. The form of the payoff in (12) is inspired by the objectives used in entropy-regularized MDPs (Neu et al., 2017). The level of irrationality is modeled through the constant τ. When $\tau = 0$, the model recovers the rational case, where the follower will choose a best response. At the other extreme, as $\tau \to \infty$, the role of r_2^x diminishes, and the follower is inclined to make the occupancy measure spread uniformly among all state-action pairs regardless of r_2^x. Since $\tau \cdot \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} m(s,a) \log \frac{m(s,a)}{\sum_{a'\in\mathcal{A}} m(s,a')}$ is strictly convex (Neu et al., 2017, Proposition 1) in m, the function in (12) admits a unique maximizer.

It can be shown that an optimal interior-point allocation is robust to *unmodeled* bounded rationality of the follower.

Proposition 10 (Robustness to bounded rationality)**.** *Suppose that Assumption 5 holds, and that* $(x^\star, m^\star)$ *is an optimal solution to the reward design problem in (5), with* $x^\star$ *being an interior point of* $\mathcal{P}_{m^\star}$*. Let*

$$m_\tau^\star = \arg\max_{m \in \mathcal{M}} \left\{ \langle r_2^{x^\star}, m \rangle - \tau \cdot \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} m(s,a) \log \frac{m(s,a)}{\sum_{a' \in \mathcal{A}} m(s,a')} \right\} \tag{13}$$

and $\mathcal{M}_{\text{det}}{}^\star = \mathcal{M}_{\text{det}} \cap \text{BR}(x^\star)$*. Then for any* $\tau > 0$*, it holds that*

$$\langle r_1, m_\tau^\star \rangle \geq \left(1 - \frac{2\tau}{b(1-\gamma)} \log |\mathcal{A}|\right) \langle r_1, m^\star \rangle, \tag{14}$$

where γ *is the discount factor, and* $b = \langle r_2^{x^\star}, m^\star \rangle - \max_{m \in \mathcal{M}_{\text{det}} \setminus \mathcal{M}_{\text{det}}^\star} \langle r_2^{x^\star}, m \rangle$*.*

Proof. See Appendix C.1. □

Since $\mathcal{P}_{m^\star}$ is an optimal allocation region, the allocation $x^\star$ is an optimal interior-point allocation. The left side of (14) is the expected payoff of the leader when the follower is boundedly rational at level τ. Proposition 10 provides a quantitative performance characterization on any optimal interior-point allocation in the presence of a near-rational follower, i.e., when τ is small. (The bound in (14) becomes vacuous when τ is large enough for the right side to become negative.) Specifically, the leader is guaranteed to receive an expected payoff not much worse than the optimal payoff $v_1^\star$ against a completely rational follower. In other words, even when incorrectly treating a near-rational follower as completely rational, the leader will still receive a payoff not far from her prediction.

The discrepancy between the predicted payoff and the actual payoff depends on b, which measures the gap between the optimal payoff and the second-best payoff of a follower who only plays deterministic policies. The constant b depends not only on the configuration of the MDP but also on $x^\star$. As b decreases, the follower becomes more indifferent between a best response and a suboptimal response. This implies that a boundedly rational follower is less likely to play an optimal policy as desired by the leader, leading to a potential decrease in the payoff of the leader.

A similar analysis can be carried out for other models of bounded rationality. For instance, another model of the payoff function of the follower, which is also used in entropy-regularized MDPs (Neu et al., 2017), is given by

$$\langle r_2^x, m \rangle + \tau \cdot \text{Ent}(m), \tag{15}$$

where $\text{Ent}(m) \triangleq -\sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} m(s,a) \log m(s,a)$ is the entropy of m. For the payoff function in (15), an optimal interior-point allocation also provably provides a robustness guarantee similar to the one in Proposition 10; see Proposition 24 in Appendix C.2 for details. The third model of bounded rationality assumes that the follower can take any δ-optimal response, which is a response that yields a payoff within δ from the optimal payoff (Gan et al., 2023). Appendix C.3 discusses how an optimal interior-point allocation is robust to δ-optimal responses.

6 Choosing an Optimal Allocation Region

When an optimal interior-point allocation exists, some optimal allocation region $\mathcal{P}^\star$ must have a nonempty interior. If $\mathcal{P}^\star$ can be identified, then an optimal interior-point allocation can be found by picking any interior point of $\mathcal{P}^\star$. However, is it possible to choose an arbitrary optimal allocation region? If not, is it necessary to examine all optimal allocation regions or only a subset? These questions will be answered in this section.

6.1 Optimal Allocation Regions May Have an Empty Interior

Since an optimal occupancy measure $m^\star$ can be computed by solving the optimization problem in (8), it may be tempting to use the corresponding region $\mathcal{P}_{m^\star}$ as the candidate optimal allocation region $\mathcal{P}^\star$. However, this procedure may fail because not all optimal allocation regions have a nonempty interior.

Example 11 (Empty interior)**.** Consider a similar setting as Example 2. As shown in Figure 4, aside from $(3,3)$, we also allow reward allocated at $(3,1)$. The transition kernel at $(1,2)$ is modified as follows: If the follower chooses right, he will arrive at $(1,3)$ or $(1,1)$, each with probability 0.5. If the follower chooses up (resp. down), he will arrive at $(1,3)$ (resp. $(1,1)$). If the follower chooses left or stay, he will remain at $(1,2)$. For any other state, the transition kernel remains deterministic. Denote by x_1 and x_2 the amount of resource allocated to the reward at $(3,3)$ and $(3,1)$, respectively. The total allocation budget C satisfies $C > 1$. Here, we also assume that the transition kernel at the states with the allocated reward only allows the follower to keep going right. Therefore, the follower cannot go back after arriving at the state with allocated reward.

Define a policy π_r of the follower such that π_r always chooses right in any state. As a result, because the follower starts at $(1,2)$, by following π_r and taking right, the follower may end up in two possible paths: 1) The follower arrives at $(1,3)$ with probability 0.5 and subsequently follows Path 1; 2) the follower arrives at $(1,1)$ with probability 0.5 and subsequently follows Path 2.

Let $\mathcal{X}_1 = \{x \mid x \succeq 0,\ x_1 = x_2 + 1\}$ and m_r be the occupancy measure induced by π_r. We will show that $\mathcal{P}_{m_\mathrm{r}} = \mathcal{X}_1$. For any $x \in \mathcal{X}_1$, the payoff for following Path 1 and that for following Path 2 are identical: In either case, the follower will receive a payoff of $2 + x_1$. Thus, although the follower does not follow either Path 1 or 2 deterministically under π_r, the follower is still always guaranteed to receive a payoff of $2 + x_1$. This is also the largest payoff possible because the follower get into absorbing states after 4 steps. Therefore, π_r is an optimal policy under x, or equivalently $x \in \mathcal{P}_{m_\mathrm{r}}$. On the other hand, for any $x \notin \mathcal{X}_1$, the corresponding payoff of Path 1 will differ from that of Path 2. Let π_u (resp. π_d) be a policy that chooses up (resp. down) at $(2,1)$ and follows Path 1 (resp. Path 2) onward, and m_u (resp. m_d) be the induced occupancy measure. The payoff under π_r will be strictly lower than the payoff of the better policy between π_u and π_d. The suboptimality of π_r implies $x \notin \mathcal{P}_{m_\mathrm{r}}$.

Consider instead $\mathcal{X}_2 = \{x \mid x \succeq 0,\ x_1 \geq x_2 + 1\}$. It is not difficult to verify that $\mathcal{P}_{m_\mathrm{u}} = \mathcal{X}_2$.

Both m_r and m_u are optimal occupancy measures because the follower is guaranteed to reach a state with allocated reward with probability 1, giving the leader the maximum payoff of 1. However, while $\mathcal{P}_{m_\mathrm{u}}$ has a nonempty interior, the interior of $\mathcal{P}_{m_\mathrm{r}}$ is empty because its dimension is 1.

6.2 Allocation Regions of Deterministic Occupancy Measures

Since not all optimal allocation regions have a nonempty interior, which region should be chosen to produce an optimal interior-point allocation? At first glance, it appears that one needs to examine every optimal allocation region or, equivalently, every optimal occupancy measure until a region with a nonempty interior is identified. Nevertheless, we will show that it suffices to focus on *deterministic* optimal occupancy measures.

Let $(x^\star, m^\star)$ be an optimal solution of the reward design problem in (5). Let $\pi^\star$ be a policy that induces $m^\star$. This immediately implies that $\pi^\star$ is an optimal policy of the follower. The following proposition ensures that any $\pi \in \Pi_{\mathrm{det}}(\pi^\star)$ is also an optimal policy of the follower.

Proposition 12. *Let $(x^\star, m^\star)$ be an optimal solution of the reward design problem in (5) and $\pi^\star$ be a policy that induces $m^\star$, i.e., $\pi^\star \in \Pi(m^\star)$. For any $\pi \in \Pi_{\mathrm{det}}(\pi^\star)$, the induced occupancy measure m^π is a best response of $x^\star$, i.e. $m^\pi \in \mathrm{BR}(x^\star)$ or, equivalently, $x^\star \in \mathcal{P}_{m^\pi}$.*

Proof. See Appendix D.1. □

Not only is any policy $\pi \in \Pi_{\mathrm{det}}(\pi^\star)$ optimal for the follower, but the policy is also optimal for the leader even though the reward function r_1 of the leader generally differs from $r_2^{x^\star}$.

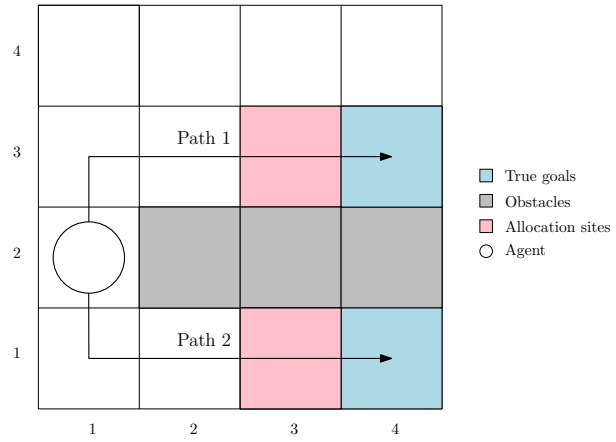

Fig. 4. A reward design problem with two states that allow reward allocation, and the budget is $C > 1$. All the transitions are deterministic except at $(1, 2)$. If the follower chooses right at $(2, 2)$, he will end up in $(1, 3)$ or $(1, 1)$ with equal probability; otherwise, the follower will move deterministically according to his action.

Proposition 13. *Under the same conditions in Proposition 12, the induced occupancy measure m^π also satisfies $\langle r_1, m^\pi \rangle = v_1^\star$.*

Proof. See Appendix D.2. □

Proposition 13 not only shows the existence of deterministic optimal occupancy measures but also suggests an algorithm to find them. If the given optimal occupancy measure $m^\star$ is randomized, one can first recover a randomized policy $\pi^\star$ from $m^\star$. Then, a deterministic policy $\pi \in \Pi_{\text{det}}(\pi^\star)$ can be constructed by examining the actions taken by $\pi^\star$ at each $s \in \mathcal{S}$: Whenever more than one action is taken by $\pi^\star$ with nonzero probability, an arbitrary one of these actions will be assigned to π to ensure $\pi \in \Pi_{\text{det}}(\pi^\star)$. The induced deterministic occupancy measure m^π from π is guaranteed to be optimal according to Proposition 13.

Theorem 14. *Let $m^\star$ be a randomized optimal occupancy measure and $\pi^\star \in \Pi(m^\star)$. For any $m \in \mathcal{M}_{\text{det}}(\pi^\star)$, it holds that $\mathcal{P}_{m^\star} \subseteq \mathcal{P}_m$, and $\mathcal{P}_m$ is an optimal allocation region.*

Proof. Consider any $x \in \mathcal{P}_{m^\star}$. Because $m \in \mathcal{M}_{\text{det}}(\pi^\star)$, m is induced by a policy in $\Pi_{\text{det}}(\pi^\star)$ by definition. From Proposition 12, we know that $x \in \mathcal{P}_m$, implying $\mathcal{P}_{m^\star} \subseteq \mathcal{P}_m$. In addition, Proposition 13 ensures that m is an optimal occupancy measure, implying that $\mathcal{P}_m$ is an optimal allocation region. □

As implied by Theorem 14, if $\mathcal{P}_{m^\star}$ has a nonempty interior, so does $\mathcal{P}_m$ for any (deterministic) occupancy measure $m \in \mathcal{M}_{\text{det}}(\pi^\star)$, where $\pi^\star$ is an arbitrary policy that induces $m^\star$. Thus, only allocation regions of deterministic optimal occupancy measures need to be examined to find an optimal interior-point allocation. Recall that the set $\mathcal{M}$ of occupancy measures is generally uncountably infinite. Theorem 14 reduces the relevant occupancy measures to a finite set $\mathcal{M}_{\text{det}}$. In theory, one can find an optimal interior-point allocation by examining every allocation region of a deterministic occupancy measure in $\mathcal{M}_{\text{det}}$ to see whether the occupancy measure is optimal and whether the region has a nonempty interior. However, such an exhaustive search can be computationally prohibitive in practice, as the size of $\mathcal{M}_{\text{det}}$ is $|\mathcal{A}|^{|\mathcal{S}|}$. This may be unavoidable in the worst case because the allocation region of a deterministic optimal occupancy measure may still has an empty interior. For instance, although π_{r} in Example 11 is a deterministic policy, its corresponding allocation region $\mathcal{P}_{m_{\text{r}}}$ has an empty interior. We postpone the computational issue until Section 8, in which we show that an optimal interior-point allocation

can be obtained by solving an MILP. Nevertheless, the reduction of the relevant occupancy measures to $\mathcal{M}_{\text{det}}$ by Theorem 14 has important theoretical implications and will be used in Section 7 to establish conditions for the existence of optimal interior-point allocations.

7 Existence of Optimal Interior-Point Allocations

While Section 5 shows that optimal interior-point allocations offer robustness, it remains a question whether an optimal interior-point allocation is guaranteed to exist. Indeed, as shown in Section 6, although it suffices to examine deterministic occupancy measures, the allocation region of a deterministic occupancy measure may still have an empty interior. As will be shown in this section, the allocation budget plays an important role in the existence of an optimal interior-point allocation. Moreover, the existence of an optimal interior-point allocation is not only sufficient (Proposition 8) but also necessary for the existence of a robust allocation against nonunique best responses of the follower.

7.1 Influence of Allocation Budget

Before presenting the general result, we would like to use an example to illustrate how the existence of an optimal interior-point allocation can be affected by the allocation budget.

Example 15. Consider the environment in Example 2. Let π_{u} be the policy that follows Path 1 and m_{u} the occupancy measure induced by π_{u}. Without considering the allocation budget, we will show that an allocation x is optimal for the leader if and only if $x \geq 1$. When $x \geq 1$, from the perspective of the follower, the total payoff of Path 1 is not less than that of Path 2. Thus, π_{u} is optimal for the follower and will give the leader the maximum payoff of 1. On the other hand, when $x < 1$, the follower will choose Path 2 over Path 1 and give the leader a suboptimal payoff of 0. This also shows that m_{u} is the only optimal occupancy measure and that the corresponding optimal allocation region is given by $\mathcal{P}_{m_{\text{u}}} = \{x \mid x \geq 1\}$.

Next, we will show how the allocation budget C affects the existence of an optimal interior-point allocation. When $C > 1$, the leader can choose any $x \in (1, C]$, which lies within the interior of $\mathcal{P}_{m_{\text{u}}}$. Because $\mathcal{P}_{m_{\text{u}}}$ is an optimal allocation region, x must be an optimal interior-point allocation. However, an optimal interior-point allocation does not exist when $C = 1$: The only optimal allocation the leader can choose is $x = 1$, which does not belong to the interior of $\mathcal{P}_{m_{\text{u}}}$.

Example 15 shows that the allocation budget is an important factor in the existence of an optimal interior-point allocation. It suggests that an optimal interior-point allocation may fail to exist when any optimal allocation must exhaust the allocation budget. To examine whether the budget needs to be exhausted, consider the following problem:

$$\begin{aligned} \underset{x \in \mathcal{X},\ m \in \mathcal{M}}{\text{maximize}} \quad & C - \sum_{i=1}^{|\mathcal{D}|} x_i \\ \text{subject to} \quad & \langle r_1, m \rangle = v_1^\star, \\ & m \in \text{BR}(x). \end{aligned} \tag{16}$$

The constraints in (16) ensure that (x, m) is an optimal solution to the reward design problem in (5). Thus, the optimal value of (16) is 0 if and only if any optimal allocation must exhaust the budget. The following theorem shows that the optimal value of problem (16) determines the existence of an optimal interior-point allocation.

Theorem 16. *Suppose that Assumption 5 holds. The reward design problem in (5) admits at least one optimal interior-point allocation if and only if the optimal value of problem (16) is strictly positive.*

Proof. See Appendix E.1. □

In the absence of Assumption 5, an interior-point allocation may still exist even if the conditions of Theorem 16 do not hold. However, in such cases, an interior-point allocation may lack the robustness discussed in Section 5.1. For example, consider allocating a reward of 0.5 at $(4, 2)$ in Example 4 when the budget is 1. While this is an optimal interior-point allocation as mentioned, no allocation in this example is robust to nonunique best responses.

7.2 Necessity for Robustness to Nonunique Best Responses

Example 15 also hints at a relationship between the existence of an optimal interior-point allocation and robustness to nonunique best responses (introduced in Section 5.2). When $C = 1$, under the only optimal allocation $x = 1$, the best responses of the follower is not unique: It is optimal for the follower to choose either Path 1 or Path 2. However, the former will give the leader a payoff of 1, whereas the latter will give a payoff of 0. In other words, an optimal allocation that is robust to nonunique best responses does not exist. Such nonexistence happens to coincide with the nonexistence of optimal interior-point allocations.

In fact, this is more than a coincidence. The existence of an optimal interior-point allocation is a sufficient and necessary condition for the existence of an allocation robust to nonunique best responses. Notice that the sufficiency has already been established in Proposition 8. The following theorem focuses only on the necessity.

Theorem 17 (Necessity of optimal interior-point allocations)**.** *Suppose that Assumption 5 holds. If there exists an optimal allocation $x^\star$ satisfying* $\mathsf{OptiVal}(x^\star) = \mathsf{PessVal}(x^\star) = v_1^\star$*, then an optimal interior-point allocation must exist.*

Proof. See Appendix E.2. □

Theorem 17 does not, however, guarantee that $x^\star$ is an optimal interior-point allocation when $x^\star$ satisfies $\mathsf{OptiVal}(x^\star) = \mathsf{PessVal}(x^\star) = v_1^\star$. For instance, consider any allocation $x^\star \in \mathcal{X}_1$ in Example 11. Due to Theorem 14, only optimal allocation regions of deterministic occupancy measures need to be examined. The allocation $x^\star$ belongs to three optimal allocation regions of deterministic occupancy measures: $\mathcal{P}_{m_\mathrm{r}}$, $\mathcal{P}_{m_\mathrm{u}}$, and $\mathcal{P}_{m_\mathrm{d}}$. It is not difficult to see that $x^\star$ is robust to nonunique best responses. However, $x^\star$ is not an interior point of any optimal allocation region. First, $x^\star$ cannot be an interior point of $\mathcal{P}_{m_\mathrm{r}}$ because Example 11 has already shown that $\mathcal{P}_{m_\mathrm{r}}$ has an empty interior. Second, $x^\star$ is not in the interior of $\mathcal{P}_{m_\mathrm{u}}$ or $\mathcal{P}_{m_\mathrm{d}}$ since it is possible to perturb $x^\star$ arbitrarily small to make m_u or m_d no longer a best response of the follower.

8 Computing an Optimal Interior-Point Allocation

For finding an optimal interior-point allocation, Theorem 14 shows that only a finite number of optimal allocation regions (i.e., the ones of deterministic occupancy measures) need to be examined. Nevertheless, as discussed at the end of Section 6.2, it may be impractical to enumerate all the allocation regions of deterministic occupancy measures. This section presents a practical method for computing an optimal interior-point allocation via MILP.

Suppose that the optimal value $v_1^\star$ of problem (5) has been obtained using the procedure described in Section 3.3. Our goal is to find (x, m) that satisfies the following conditions:

(1) x is an admissible allocation: $x \in \mathcal{X}$.
(2) m is an optimal occupancy measure: $m \in \mathcal{M}$, and $\langle r_1, m\rangle = v_1^\star$.
(3) x is an interior point of $\mathcal{P}_m$: There exists $c > 0$ such that (11) holds.

An issue with condition (11) is that the condition would lead to infinitely many constraints because it requires checking every v satisfying $\|v\|_1 \le 1$. Denote by $\{e_i\}_{i=1}^{|\mathcal{D}|}$ the standard basis of $\mathbb{R}^{|\mathcal{D}|}$. Recall that the unit ℓ_1-norm ball in $\mathbb{R}^{|\mathcal{D}|}$ is the convex hull of $\{\pm e_i\}_{i=1}^{|\mathcal{D}|}$. Since $\mathcal{P}_m$ is a convex set, condition (11) is equivalent to

$$x + ce_i \in \mathcal{P}_m, \quad x - ce_i \in \mathcal{P}_m, \quad i = 1, \ldots, |\mathcal{D}|, \tag{17}$$

or equivalently,

$$m \in \mathrm{BR}(x + ce_i), \quad m \in \mathrm{BR}(x - ce_i), \quad i = 1, \ldots, |\mathcal{D}|. \tag{18}$$

According to the KKT conditions in (7), condition (18) holds if and only if there exist v_i^+ and v_i^- for $i = 1, \ldots, |\mathcal{D}|$ such that

$$Am = \rho, \qquad m \succeq 0, \tag{19}$$

$$A^T v_i^+ - r_2^{x+ce_i} \succeq 0, \quad A^T v_i^- - r_2^{x-ce_i} \succeq 0, \quad i = 1, \ldots, |\mathcal{D}|, \tag{20}$$

$$m \perp A^T v_i^+ - r_2^{x+ce_i}, \quad m \perp A^T v_i^- - r_2^{x-ce_i}, \quad i = 1, \ldots, |\mathcal{D}|. \tag{21}$$

Rather than an arbitrary optimal interior-point allocation, it is actually possible to find an optimal interior-point allocation with the *maximum* margin by solving the following optimization problem:

$$\begin{aligned} &\underset{x,m,c,v_i^+,v_i^-}{\text{maximize}} \quad c \\ &\text{subject to} \quad x \in \mathcal{X}, \quad \langle r_1, m \rangle = v_1^\star, \quad (19)\text{–}(21). \end{aligned} \tag{22}$$

Let $(x^\star, m^\star, c^\star)$ be an optimal solution of problem (22). If $c^\star > 0$, then $x^\star$ is an interior point of the optimal allocation region $\mathcal{P}_{m^\star}$ and hence is an optimal interior-point allocation with margin $c^\star$.

Similar to (8c), the complementarity constraints in (21) can be reformulated as affine constraints with integer variables. Since all the remaining constraints in (22) are affine, problem (22) can be reformulated as an MILP and solved by off-the-shelf solvers including Gurobi (Gurobi Optimization, LLC, 2024) and CPLEX (IBM, 2022).

9 Numerical Experiments

In this section, we numerically validate the robustness of the maximum-margin optimal interior-point allocation given by (22) and compare it with an arbitrary optimal allocation given by (8). Denote by x_{IP} the optimal interior-point allocation given by (22), and denote by x_{MILP} an arbitrary optimal allocation given by (8). We sometimes refer to x_{IP} as the *interior-point solution* and x_{MILP} as the *MILP solution*. The robustness of allocation is evaluated in three different environments: a 6×6 grid world, a 10×10 grid world, and a probabilistic attack graph. We will check two notions of robustness: robustness to nonunique best responses and robustness to a boundedly rational attacker. The remaining notion of robustness, i.e., robustness to uncertain reward perception of the attacker, will be demonstrated by the computed margin of allocation. All numerical experiments were performed on a Macbook Air laptop computer with an Apple M2 processor and 8 GB RAM running macOS Sonoma 14.3.1. The interior-point solutions and MILP solutions in different environments are computed using the Python MIP package with Gurobi 11.0.0. Code for the numerical experiments is provided at the URL in Footnote[1].

9.1 Environments

We evaluate the robustness of reward allocations in three distinct environments: two grid worlds of different size and a probabilistic attack graph. Each environment is modeled as an MDP in which a defender (leader) allocates rewards and an attacker (follower) navigates toward a goal. The environments are designed to test the defender's performance under different spatial and structural settings. We describe each environment in detail below.

9.1.1 6×6 Grid World. The 6×6 grid world illustrated in Figure 5 is modeled as an MDP with finite state and action spaces. Each cell is associated with a 2-tuple that represents the horizontal and vertical coordinates of the cell: The coordinate of the lower-left cell is $(1, 1)$, and that of the upper-right cell is $(6, 6)$. The grid world consists of a defender and an attacker, who play the roles of the leader and the follower, respectively. The attacker always starts from the state $(1, 3)$ and can choose one of the four directions, left, right, up, and down, as the

[1] https://github.com/fribuilder/robust-reward-design

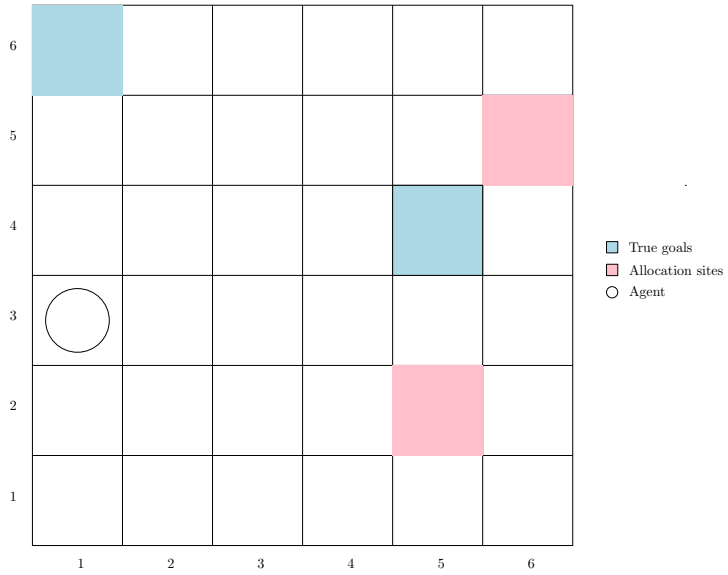

Fig. 5. The 6×6 grid world. The true goals are at $(1, 6)$ and $(5, 4)$, and the defender can allocate rewards at $(5, 2)$ and $(6, 5)$. The attacker always starts from $(1, 3)$.

desired direction to move. The attacker will transit to the adjacent cell in the desired direction with probability 0.8 if the new cell remains within the boundary. The attacker may also move laterally to a new cell in one of the two directions, each with probability 0.1. The attacker will never move opposite to the desired direction. In any case, if the new cell is out of bounds, then the attacker will stay at the current cell. The grid world has two true goals at $(1, 6)$ and $(5, 4)$ and two allocated rewards set up by the defender at $(5, 2)$ and $(6, 5)$. From the perspective of the attacker, the allocated rewards are indistinguishable from the true targets. The perceived reward of the attacker for reaching any of the true goals is 1. The perceived reward for reaching an allocated reward is equal to the allocated resource therein by the defender. The allocated resource at each allocated reward must be nonnegative, and the total budget of allocation is 4. The MDP ends when the attacker arrives at a true goal or an allocated reward. The goal of the defender is to find an allocation strategy that maximizes the probability for the attacker to receive an allocated reward. This goal is captured by the reward function of the defender: $r_1(s, a) = 1\ \forall (s, a) : s \in \mathcal{S}_d$, and $r_1(s, a) = 0\ \forall (s, a) : s \notin \mathcal{S}_d$, which is consistent with the form of r_1 introduced in Section 3.1. Using the notation from Section 3, the problem setup is summarized as follows:

- State space $\mathcal{S} = \{1, 2, \cdots, 6\}^2$ represents the cells in the 6×6 grid world.
- Action space $\mathcal{A} = \{\text{left}, \text{right}, \text{up}, \text{down}\}$ represents the four desired directions of movement.
- The set $\mathcal{S}_d = \{s_1, s_2\}$, where $s_1 = (1, 4)$ and $s_2 = (4, 5)$, represents the cells at which the allocated rewards are located.
- The reward function of the attacker is given by r, and that of the defender is given by r_1. Both functions are defined in Section 3.1.

9.1.2 10×10 *Grid World.* The transition kernel, the total budget of allocation, and actions of the attacker in the 10×10 grid world (Figure 6) are the same as those in the 6×6 grid world. The true goals are at $(6, 9)$, $(8, 5)$, and $(10, 3)$, and the allocated rewards are at $s_1 = (8, 1)$, $s_2 = (7, 4)$, and $s_3 = (6, 9)$. The attacker always starts from the state $(2, 6)$.

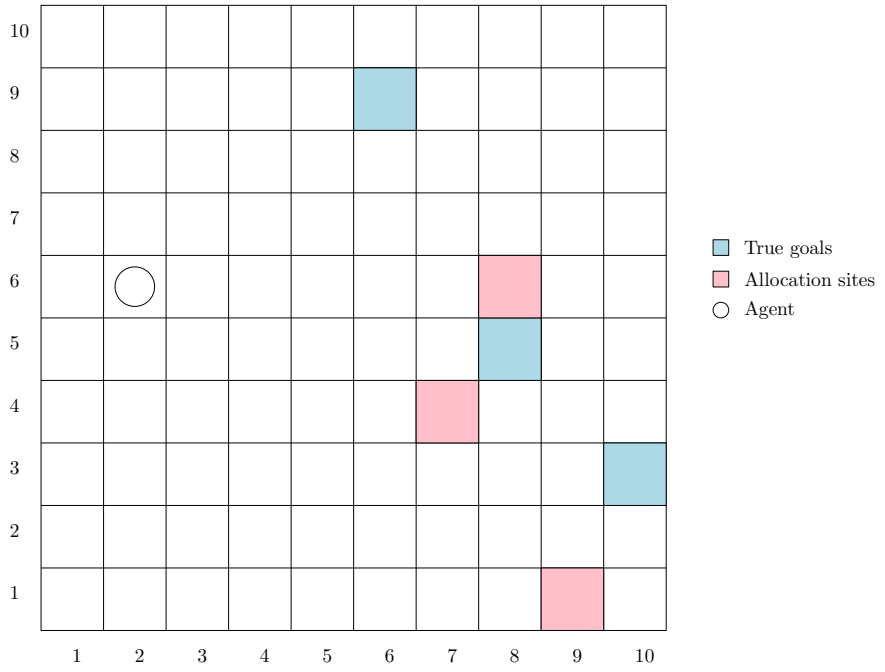

Fig. 6. The 10×10 grid world. The true goals are at $(6, 9)$, $(8, 5)$ and $(10, 3)$, and the defender can allocate rewards at $(8, 1)$, $(7, 4)$ and $(6, 9)$. The attacker always starts from $(2, 6)$.

9.1.3 Probabilistic Attack Graph. Figure 7 shows the probabilistic attack graph used in our numerical experiments. The probabilistic attack graph is an MDP whose state space consists of all the nodes in the graph. The attacker has four actions $\{a, b, c, d\}$. The transition probabilities of the MDP are defined by the edges of the graph. For clarity, the graph only shows the possible transitions given action a, where a thick (resp. thin) arrow represents a high (resp. low) transition probability. For example, when the attacker is at node 0 and takes action a, the transition probabilities are given by $\mathcal{T}(0, a, 1) = 0.7$ and $\mathcal{T}(0, a, i) = 0.1$ for $i = 2, 3, 4$. The attacker always starts from state 0. The only true goal is node 10. The reward allocations are set at $\{11, 12\}$. Similar to the grid world environments, the perceived reward of the attacker for reaching the true goal is 1. The perceived reward for reaching an allocated reward is equal to the allocated resource therein by the defender. The total budget of allocation is 4. The reward function r_1 of the defender is the same as the one used in the grid world environments.

9.2 Robustness to Nonunique Best Responses

We first validate the robustness of the interior-point allocation to nonunique best responses. According to Proposition 8, the interior-point solution guarantees that any best response of the follower yields the optimal payoff for the defender. To numerically validate this property, we develop a relaxation-based verification procedure and apply it to each of the three environments introduced in Section 9.1.

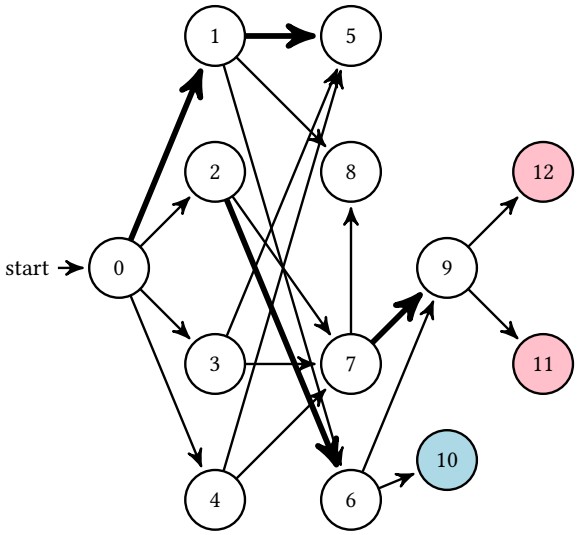

Fig. 7. A probabilistic attack graph.

9.2.1 Method of Validation. Our goal is to verify Proposition 8 that the interior-point solution x_{IP} has the property that all its best responses are optimal, i.e., $\text{OptiVal}(x_{\text{IP}}) = \text{PessVal}(x_{\text{IP}}) = v_1^\star$. Since $\text{OptiVal}(x_{\text{IP}}) = v_1^\star$ by the definition of x_{IP}, it remains to show that $\text{PessVal}(x_{\text{IP}}) = v_1^\star$. Although $\text{PessVal}(x_{\text{IP}})$ can be obtained by solving the problem in (10) in theory, the constraint may become infeasible due to finite numerical precision. Instead, given an optimal allocation $x^\star$ of the reward design problem in (5), we solve the relaxed problem

$$\begin{aligned} \underset{m \in \mathcal{M}}{\text{minimize}} \quad & \langle r_1, m \rangle \\ \text{subject to} \quad & \max_{m' \in \mathcal{M}} \langle r_2^{x^\star}, m' \rangle - \langle r_2^{x^\star}, m \rangle \le \epsilon \end{aligned} \tag{23}$$

for different values of ϵ. Denote the optimal value of (23) by $v_\epsilon(x^\star)$. If $\text{PessVal}(x^\star) = v_1^\star$, then $v_\epsilon(x^\star)$ can be made arbitrarily close to $v_1^\star$ by choosing ϵ sufficiently small.

Proposition 18. *Let $x^\star$ be an optimal allocation. Then $\lim_{\epsilon \to 0^+} v_\epsilon(x^\star) = v_1^\star$ if and only if $x^\star$ is robust to nonunique best responses, i.e.,* $\text{OptiVal}(x^\star) = \text{PessVal}(x^\star) = v_1^\star$.

Proof. See Appendix B.2. □

9.2.2 6×6 Grid World. For the 6×6 grid world, the MILP defined in (8) gives an optimal reward allocation $x_{\text{MILP}} = (1.946,\ 1.774)$. The defender's expected payoff is 0.433. In comparison, the maximum-margin optimal interior-point allocation given by (22) is $x_{\text{IP}} = (2.122,\ 1.869)$. The defender's expected payoff under x_{IP} is the same as the MILP solution. The margin of the interior-point solution is 0.087.

We then computed $v_\epsilon(x_{\text{MILP}})$ and $v_\epsilon(x_{\text{IP}})$ under different values of ϵ by solving problem (23). The results are shown in Table 2 and Figure 8. As ϵ approaches 0, the value of $v_\epsilon(x_{\text{IP}})$ also approaches $v_1^\star = 0.433$. By Proposition 18, the results support Proposition 8 that the interior-point solution is robust to nonunique best responses. In comparison, x_{MILP} is not an interior-point allocation. The value of $v_\epsilon(x_{\text{MILP}})$ appears to converge as ϵ approaches 0, with the gap $v_1^\star - v_\epsilon(x_{\text{MILP}}) > 0.03$ even when $\epsilon = 10^{-10}$, showing that x_{MILP} is not robust to nonunique best responses.

Table 2. 6×6 grid world case: The optimal values of (23) for the MILP solution and the interior-point solution under different values of ϵ.

ϵ	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
$v_\epsilon(x_{\text{MILP}})$	0.107	0.195	0.378	0.396	0.398
$v_\epsilon(x_{\text{IP}})$	0.379	0.425	0.432	0.432	0.433

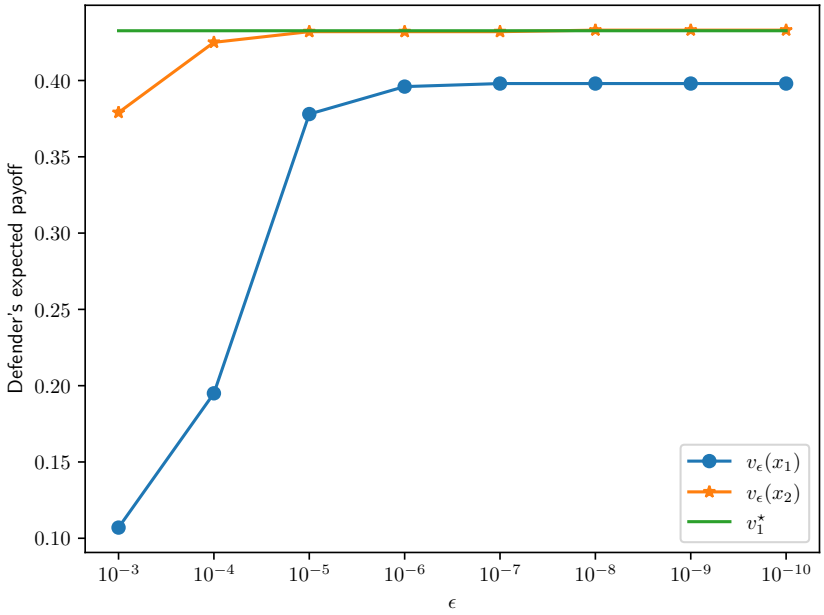

Fig. 8. 6×6 grid world case: The optimal values of (23) for the MILP solution and the interior-point solution under different values of ϵ.

9.2.3 10×10 Grid World. For the 10×10 grid world, the optimal solution given by the MILP is $x_{\text{MILP}} = (1.429,\ 0.000,\ 1.466)$, and the interior-point solution is $x_{\text{IP}} = (1.836,\ 0.000,\ 2.164)$. The optimal expected payoffs of the defender are both 0.495 given two reward allocations x_{MILP} and x_{IP}. The margin of the interior-point solution is 0.061.

The optimal values of (23) for x_{MILP} and x_{IP} under different ϵ are given in (3). In this case, the interior-point solution remains robust to nonunique best responses according to the numerical results. The MILP solution may also be an optimal interior-point allocation, but it is difficult to conclude definitively due to finite numerical precision. If the defender's payoff finally approaches 0.495 under x_{MILP}, then x_{MILP} must be robust to nonunique best responses due to Proposition 18.

9.2.4 Attack Graph. For the probabilistic attack graph, the MILP solution is $x_{\text{MILP}} = (1.218,\ 0)$, and the interior-point solution is $x_{\text{IP}} = (2.667, 1.333)$ with a margin of 1.333. Both solutions lead to a payoff of 0.655 for the defender. Table 4 shows the optimal values of (23) for x_{MILP} and x_{IP} under different ϵ. The numerical results imply that the interior-point solution is robust to nonunique best responses, whereas the MILP solution is not.

Table 3. 10×10 grid world case: The optimal values of (23) for the MILP solution and the interior-point solution under different values of ϵ.

ϵ	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-6}
$v_\epsilon(x_{\text{MILP}})$	0.491	0.494	0.494	0.494	0.494
$v_\epsilon(x_{\text{IP}})$	0.494	0.495	0.495	0.495	0.495

Table 4. Attack graph case: The optimal values of (23) for the MILP solution and the interior-point solution under different values of ϵ.

ϵ	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-6}
$v_\epsilon(x_{\text{MILP}})$	0.247	0.248	0.248	0.248	0.248
$v_\epsilon(x_{\text{IP}})$	0.654	0.654	0.655	0.655	0.655

Table 5. 6×6 grid world case: Expected payoffs of the defender for the interior-point solution and the MILP solution when facing attackers with different levels of rationality. The optimal payoff against a rational attacker ($\tau = 0$) is given by $v_1^\star = 0.433$.

τ	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Optimal solution	1.31×10^{-3}	0.429	0.433	0.433	0.433
Interior-point solution	1.29×10^{-3}	0.429	0.433	0.433	0.433
MILP solution	8.63×10^{-4}	0.335	0.324	0.385	0.414

9.3 Robustness to a Boundedly Rational Attacker

For each environment introduced in Section 9.1, we computed the defender's payoff against a boundedly rational attacker. Our model of bounded rationality assumes that the attacker solves an entropy-regularized MDP in (12). The parameter τ in (12) reflects the level of rationality of the attacker: The smaller the value of τ, the more rational the attacker.

For the 6×6 grid world, Table 5 gives the expected payoffs of the defender when facing attackers with different levels of rationality for the interior-point solution x_{IP} (Row 3) and the MILP solution x_{MILP} (Row 4), respectively. From Proposition 10, when the interior-point solution is used for allocation, a lower bound for the defender's payoff should approach $v_1^\star = \langle r_1, m^\star \rangle$ as $\tau \to 0$. Recall from Section 9.2.2 that the optimal expected payoff against a rational attacker is given by $v_1^\star = 0.433$. It can be seen that the theoretical result given by Proposition 10 is consistent with the numerical results in Row 3 of Table 5. In comparison, the MILP solution does not enjoy the same payoff guarantee. The payoff for the MILP solution remains noticeably below $v_1^\star$ even when τ is as small as 10^{-5}. In addition, the MILP solution performs consistently worse than the interior-point solution for all values of τ.

For comparison, Row 2 of Table 5 shows the optimal expected payoffs of the defender against a boundedly rational attacker, where the defender is assumed to know the actual value of τ and choose the reward allocation accordingly. The expected payoff of the defender for the interior-point solution is found to be close to the optimal payoff. In other words, the interior-point solution leads to a near-optimal payoff even when the defender incorrectly assumes a rational attacker.

Table 6. 10×10 grid world case: Expected payoffs of the defender for the interior-point solution and the MILP solution when facing attackers with different levels of rationality. The optimal payoff against a rational attacker ($\tau = 0$) is given by $v_1^\star = 0.495$.

τ	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Optimal solution	0.194	0.489	0.495	0.495	0.495
Interior point solution	1.21×10^{-3}	0.489	0.495	0.495	0.495
MILP solution	1.78×10^{-4}	0.484	0.495	0.495	0.495

Table 7. Attack graph case: Expected payoffs of the defender for the interior-point solution and the MILP solution when facing attackers with different levels of rationality. The optimal payoff against a rational attacker ($\tau = 0$) is given by $v_1^\star = 0.655$.

τ	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
Optimal solution	0.620	0.655	0.655	0.655	0.655
Interior-point solution	0.583	0.655	0.655	0.655	0.655
MILP solution	0.407	0.446	0.446	0.446	0.500

Similar relationships between the payoffs were observed for the 10×10 grid world and the probabilistic attack graph. The numerical results are shown in Table 6 and Table 7.

9.4 Scalability

This subsection presents the scalability analysis of solving both the robust optimization problem (22) and its non-robust counterpart (8). Instead of solving (22) directly, we applied a bisection method to identify the largest feasible margin c, where each step involved solving an MILP to verify feasibility. The bisection was initialized with lower and upper bounds of 0 and C, respectively, where C is the total reward budget. Typically, the bisection terminated within 10 to 20 iterations.

The numerical experiments were conducted in a grid world environment for 5 different sizes. For each grid size, we recorded the running time per feasibility check from 15 instances, each with a different initial state distribution ρ. We observed that feasibility checks tended to incur longer and more variable running times when the proposed margin is near-feasible, i.e., close to the optimal margin. To accurately characterize scalability, the proposed margin was kept away from the optimal value in all the experiments. For comparison, we also recorded the time required to solve the corresponding non-robust problem (8). The results across different grid sizes are summarized in Table 8, where instances that exceed 20 minutes are excluded from the calculation of the average and median running times.

Overall, for all the grid sizes tested, the running time for checking the feasibility of the robust allocation problem is comparable to that of solving the non-robust allocation problem. For small grid sizes, the feasibility check rarely exceeded 20 minutes. Since a bisection search for the optimal margin typically terminated in 10 to 20 feasibility checks, the robust allocation problem could be solved within several minutes. For larger grid sizes, although the running time for some feasibility checks remained moderate, an increasing portion of the instances began to exceed 20 minutes, thus adversely affecting the total running time for solving the robust allocation problem. The issue of longer running times is further exacerbated by the fact that feasibility checks take longer to complete when the proposed margin is near-feasible, which is inevitable in the last few steps of the bisection search.

Table 8. Average and median running times (in seconds) and percentage of instances taking longer than 20 minutes in a grid world environment. Both the times for the robust allocation problem (22) and the non-robust allocation problem (8) are shown. The running time for the robust allocation problem is only for checking the feasibility of a proposed margin c. The robust allocation problem can typically be solved with 10 to 20 feasibility checks in a bisection search.

Grid Size	Robust Allocation (Feasibility)			Non-Robust Allocation		
	Avg (s)	Median (s)	% > 20 min	Avg (s)	Median (s)	% > 20 min
6×6	0.2	0.2	0%	0.1	0.1	0%
10×10	2.0	1.8	0%	0.9	0.9	0%
12×12	6.6	5.4	20%	6.7	2.05	0%
15×15	7.2	5.6	40%	6.0	3.0	7%
20×20	18.6	16.0	60%	30.5	10.6	20%

10 Future Directions

Scalability. The numerical experiments from Section 9 show that our approach is currently limited to relatively small reward design problems. Because the standard reward design problem without robustness considerations is known to be NP-hard (Zhang and Parkes, 2008), we do not expect the robust reward design problem to admit an efficient algorithm in general. An interesting future research direction is to explore efficient algorithms for computing an approximate solution to the robust reward design problem. Recent work by Ben-Porat et al. (2024) gave a polynomial-time approximation algorithm for solving the standard reward design problem. Combining ideas from their approach with our formulation may serve as a promising direction for developing a polynomial-time approximation algorithm for reward design with robustness guarantees. Another interesting direction is to study the use of function approximation or a parametric policy class, both of which have been recognized as common techniques for handling large, even infinite, state and action spaces. Our main results (e.g., Theorem 17) cannot be immediately extended because they rely on the fact that the set of deterministic policies is finite, which no longer holds in general beyond tabular MDPs.

Learning-based approaches. Our setting assumes perfect knowledge of the transition dynamics faced by the follower. A more general setting is when the transition dynamics are fixed but not known a priori. Without knowing the transition dynamics, one cannot compute the set of best responses for a given reward allocation and therefore cannot determine the allocation regions. However, if repeated interaction with the follower is allowed, it may be possible to gain information about the allocation regions from the trajectory of states and actions of the follower. Naively, when the leader fixes the reward allocation x, a corresponding best response $m \in \mathrm{BR}(x)$ can be estimated accurately from the empirical distribution computed over sufficiently many rounds of interaction. This would imply $x \in \mathcal{P}_m$ with high probability and thus reveal information about $\mathcal{P}_m$. As more information about different allocation regions is gathered, the leader can expect to eventually learn an optimal reward allocation with high probability. In fact, the viewpoint based on best-response regions has been used in previous work (Blum et al., 2014; Letchford et al., 2009) on learning in normal-form Stackelberg games and has led to efficient algorithms for computing an optimal strategy for the leader. In theory, these algorithms can be applied to reward design for MDPs because the problem of reward design can be equivalently viewed as a normal-form Stackelberg game. However, the number of corresponding pure strategies played by the follower would equal to the number of deterministic policies, which grows exponentially with the problem size, making existing learning algorithms impractical to apply. It remains an open question how our MILP reformulation can be combined with existing learning algorithms for normal-form Stackelberg games to yield a practical algorithm that learns a robust reward allocation for MDPs.

11 Conclusions

We study the problem of reward design for Markov decision processes in a leader-follower setup, where the leader is tasked with modifying the follower's reward function to induce a policy favorable to the leader. Existing methods for reward design typically rely on exact knowledge of the follower's reward function and can be sensitive to modeling errors. Motivated by the issue of sensitivity, we present a new method of reward design with robustness guarantees when the follower's reward function is unknown. Our method can be viewed as a formal algorithmic counterpart of the folklore solution for achieving robustness based on perturbing the leader's strategy. The reward modification in our method is based on the concept of *optimal interior-point allocation*. We show that an optimal interior-point allocation provably offers robustness to three types of uncertainties, including 1) how the follower breaks ties in the presence of nonunique best responses, 2) inexact knowledge of how the follower perceives reward modifications, and 3) bounded rationality of the follower.

We also study the existence of an optimal interior-point allocation when the leader's reward function adopts a special form. One complication in finding an optimal interior-point allocation is that the corresponding optimal allocation region cannot be predetermined because some may have an empty interior. Nevertheless, our result shows that it suffices to focus on optimal allocation regions of deterministic occupancy measures. This characterization of the optimal allocation regions leads to two sufficient and necessary conditions for the existence of an optimal interior-point allocation. One examines whether the allocation budget needs to be exhausted to achieve the optimal leader's payoff, which can be verified numerically. The other requires the existence of an allocation robust to nonunique best responses of the follower, which establishes the central role that optimal interior-point allocations play in achieving robustness. When the leader uses a general reward function, most results can be generalized by replacing the optimal interior-point allocation with the *optimal interior-point reward function*. However, a sufficient and necessary condition for the existence of an optimal interior-point reward function remains an open question.

We further show that an optimal interior-point allocation can be found by mixed-integer linear programming if such an allocation exists. In fact, the mixed-integer linear program can be used to find an optimal interior-point allocation with the largest margin. Numerical experiments have been conducted in several simulation environments inspired by problems in cybersecurity. The experiments not only validate the theoretical guarantees on robustness but also show that our method scales to problems of practical size.

Acknowledgments

Research was sponsored by the Army Research Office and was accomplished under Grant Numbers W911NF-22-1-0034 and W911NF-22-1-0166. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. We would like to thank Tamer Başar and Haifeng Xu for helpful discussions.

References

Tamer Başar. 2024. Inducement of Desired Behavior via Soft Policies. *International Game Theory Review* 26, 02 (2024), 2440002. https://doi.org/10.1142/S0219198924400024

Kiarash Banihashem, Adish Singla, Jiarui Gan, and Goran Radanovic. 2022. Admissible Policy Teaching Through Reward Design. In *AAAI Conference on Artificial Intelligence*, Vol. 36. 6037–6045. https://doi.org/10.1609/aaai.v36i6.20550

Tamer Başar. 1982. General Theory for Stackelberg Games with Partial State Information. *Large Scale Systems* 3, 1 (1982), 47–56.

Tamer Başar. 1984. Affine Incentive Schemes for Stochastic Systems with Dynamic Information. *SIAM Journal on Control and Optimization* 22, 2 (March 1984), 199–210. https://doi.org/10.1137/0322015

Omer Ben-Porat, Yishay Mansour, Michal Moshkovitz, and Boaz Taitler. 2024. Principal-Agent Reward Shaping in MDPs. In *AAAI Conference on Artificial Intelligence*, Vol. 38. 9502–9510. https://doi.org/10.1609/aaai.v38i9.28805

Martino Bernasconi, Matteo Castiglioni, Alberto Marchesi, and Mirco Mutti. 2023. Persuading Farsighted Receivers in MDPs: The Power of Honesty. In *Advances in Neural Information Processing Systems*, Vol. 36. 14987–15014. https://proceedings.neurips.cc/paper_files/paper/2023/file/30b28eb87fe7a6c4af8520293317d4c6-Paper-Conference.pdf

Avrim Blum, Nika Haghtalab, and Ariel D. Procaccia. 2014. Learning Optimal Commitment to Overcome Insecurity. In *Advances in Neural Information Processing Systems*, Vol. 27. 1826–1834. https://proceedings.neurips.cc/paper_files/paper/2014/file/d3eea90e8e9cff58fc84a4c40d22d95a-Paper.pdf

Patrick Bolton and Mathias Dewatripont. 2005. *Contract Theory*. MIT Press, Cambridge, MA and London, England.

Michele Breton, Abderrahmane Alj, and Alain Haurie. 1988. Sequential Stackelberg Equilibria in Two-Person Games. *Journal of Optimization Theory and Applications* 59, 1 (Oct. 1988), 71–97. https://doi.org/10.1007/BF00939867

Colin F. Camerer, Teck-Hua Ho, and Juin-Kuan Chong. 2004. A Cognitive Hierarchy Model of Games. *The Quarterly Journal of Economics* 119, 3 (Aug. 2004), 861–898. https://doi.org/10.1162/0033553041502225

Derya H. Cansever and Tamer Başar. 1982. A Minimum Sensitivity Approach to Incentive Design Problems. In *IEEE Conference on Decision and Control*. 158–163. https://doi.org/10.1109/CDC.1982.268419

Derya H. Cansever and Tamer Başar. 1985. Optimum/Near-Optimum Incentive Policies for Stochastic Decision Problems Involving Parametric Uncertainty. *Automatica* 21, 5 (Sept. 1985), 575–584. https://doi.org/10.1016/0005-1098(85)90006-8

Thomas E. Carroll and Daniel Grosu. 2009. A Game Theoretic Investigation of Deception in Network Security. In *International Conference on Computer Communications and Networks*. 1–6.

Souradip Chakraborty, Amrit Bedi, Alec Koppel, Huazheng Wang, Dinesh Manocha, Mengdi Wang, and Furong Huang. 2024. PARL: A Unified Framework for Policy Alignment in Reinforcement Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=ByR3NdDSZB

Siyu Chen, Donglin Yang, Jiayang Li, Senmiao Wang, Zhuoran Yang, and Zhaoran Wang. 2022. Adaptive Model Design for Markov Decision Process. In *International Conference on Machine Learning*, Vol. 162. 3679–3700. https://proceedings.mlr.press/v162/chen22ab.html

Vincent Conitzer and Tuomas Sandholm. 2006. Computing the Optimal Strategy to Commit To. In *ACM Conference on Electronic Commerce*. 82–90. https://doi.org/10.1145/1134707.1134717

Stephan Dempe and Alain B. Zemkoho. 2020. *Bilevel Optimization: Advances and next Challenges*. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-030-52119-6

Sam Devlin and Daniel Kudenko. 2011. Theoretical Considerations of Potential-Based Reward Shaping for Multi-agent Systems. In *International Conference on Autonomous Agents and Multiagent Systems*. 225–232. https://dl.acm.org/doi/10.5555/2030470.2030503

Marcel Frigault, Lingyu Wang, Anoop Singhal, and Sushil Jajodia. 2008. Measuring Network Security Using Dynamic Bayesian Network. In *ACM Workshop on Quality of Protection*. 23–30. https://doi.org/10.1145/1456362.1456368

Jiarui Gan, Minbiao Han, Jibang Wu, and Haifeng Xu. 2023. Robust Stackelberg Equilibria. In *ACM Conference on Economics and Computation*. 735. https://doi.org/10.1145/3580507.3597680

Jiarui Gan, Minbiao Han, Jibang Wu, and Haifeng Xu. 2024. Generalized Principal-Agency: Contracts, Information, Games and Beyond. arXiv:2209.01146

Jiarui Gan, Rupak Majumdar, Goran Radanovic, and Adish Singla. 2022. Bayesian Persuasion in Sequential Decision-Making. In *AAAI Conference on Artificial Intelligence*, Vol. 36. 5025–5033.

Sanford J. Grossman and Oliver D. Hart. 1983. An Analysis of the Principal-Agent Problem. *Econometrica* 51, 1 (1983), 7–45. http://www.jstor.org/stable/1912246

Gurobi Optimization, LLC. 2024. Gurobi Optimizer Reference Manual. https://www.gurobi.com

Oliver Hart and Bengt Holmström. 1987. The Theory of Contracts. In *Advances in Economic Theory* (1 ed.). Cambridge University Press, 71–156. https://doi.org/10.1017/CCOL0521340446.003

Rattikorn Hewett and Phongphun Kijsanayothin. 2008. Host-Centric Model Checking for Network Vulnerability Analysis. In *Annual Computer Security Applications Conference*. 225–234. https://doi.org/10.1109/ACSAC.2008.15

John R. Hicks. 1935. Marktform und Gleichgewicht. *The Economic Journal* 45, 178 (June 1935), 334–336. https://doi.org/10.2307/2224643

Yu-Chi Ho, Peter B. Luh, and Geert Jan Olsder. 1982. A Control-Theoretic View on Incentives. *Automatica* 18, 2 (March 1982), 167–179. https://doi.org/10.1016/0005-1098(82)90106-6

IBM. 2022. IBM ILOG CPLEX 22.1.1 User's Manual. https://www.ibm.com/docs/en/icos/22.1.1?topic=optimizers-users-manual-cplex

Somesh Jha, Oleg Sheyner, and Jeannette Wing. 2002. Two Formal Analyses of Attack Graphs. In *IEEE Computer Security Foundations Workshop*. 49–63. https://doi.org/10.1109/CSFW.2002.1021806

Sham Kakade and John Langford. 2002. Approximately Optimal Approximate Reinforcement Learning. In *International Conference on Machine Learning*. 267–274. https://dl.acm.org/doi/10.5555/645531.656005

Emir Kamenica and Matthew Gentzkow. 2011. Bayesian Persuasion. *American Economic Review* 101, 6 (2011), 2590–2615. https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590

Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. 2009. Learning and Approximating the Optimal Strategy to Commit To. In *International Symposium on Algorithmic Game Theory*. 250–262. https://doi.org/10.1007/978-3-642-04645-2_23

Shiau Hong Lim, Huan Xu, and Shie Mannor. 2013. Reinforcement Learning in Robust Markov Decision Processes. In *Advances in Neural Information Processing Systems*, Vol. 26. https://proceedings.neurips.cc/paper_files/paper/2013/file/0deb1c54814305ca9ad266f53bc82511-Paper.pdf

R Duncan Luce. 1959. *Individual Choice Behavior*. John Wiley.

Haoxiang Ma, Shuo Han, Charles Kamhoua, and Jie Fu. 2023. Optimal Resource Allocation for Proactive Defense with Deception in Probabilistic Attack Graphs. In *Decision and Game Theory for Security*, Vol. 14167. Springer Nature Switzerland, 215–233. https://doi.org/10.1007/978-3-031-50670-3_11

Daniel McFadden. 1976. *Quantal Choice Analysis: A Survey*. NBER Chapters. National Bureau of Economic Research, Inc. 363–390 pages.

Richard D. McKelvey and Thomas R. Palfrey. 1995. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior* 10, 1 (July 1995), 6–38. https://doi.org/10.1006/game.1995.1023

Roger B Myerson. 1982. Optimal Coordination Mechanisms in Generalized Principal–Agent Problems. *Journal of Mathematical Economics* 10, 1 (June 1982), 67–81. https://doi.org/10.1016/0304-4068(82)90006-4

Gergely Neu, Anders Jonsson, and Vicenç Gómez. 2017. A Unified View of Entropy-Regularized Markov Decision Processes. arXiv:1705.07798

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. 1999. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In *International Conference on Machine Learning*. 278–287. http://dl.acm.org/doi/10.5555/645528.657613

Thanh H. Nguyen, Mason Wright, Michael P. Wellman, and Satinder Singh. 2018. Multistage Attack Graph Security Games: Heuristic Strategies, with Empirical Game-Theoretic Analysis. *Security and Communication Networks* 2018, 1 (2018), 2864873. https://doi.org/10.1155/2018/2864873

Arnab Nilim and Laurent El Ghaoui. 2005. Robust Control of Markov Decision Processes with Uncertain Transition Matrices. *Operations Research* 53, 5 (Sept. 2005), 780–798. https://doi.org/10.1287/opre.1050.0216

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2022. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=JYtwGwIL7ye

Praveen Paruchuri, Jonathan P. Pearce, Janusz Marecki, Milind Tambe, Fernando Ordonez, and Sarit Kraus. 2008. Efficient Algorithms to Solve Bayesian Stackelberg Games for Security Applications. In *National Conference on Artificial Intelligence*, Vol. 3. 1559–1562. https://dl.acm.org/doi/10.5555/1620270.1620332

Martin L. Puterman. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (1st ed.). John Wiley & Sons, Inc.

Xingsheng Qin, Frank Jiang, Mingcan Cen, and Robin Doss. 2023. Hybrid Cyber Defense Strategies Using Honey-X: A Survey. *Computer Networks* 230 (July 2023), 109776. https://doi.org/10.1016/j.comnet.2023.109776

Lillian J. Ratliff, Roy Dong, Shreyas Sekar, and Tanner Fiez. 2019. A Perspective on Incentive Design: Challenges and Opportunities. *Annual Review of Control, Robotics, and Autonomous Systems* 2 (2019), 305–338. http://dx.doi.org/10.1146/annurev-control-053018-023634

Stephen A. Ross. 1973. The Economic Theory of Agency: The Principal's Problem. *The American Economic Review* 63, 2 (1973), 134–139. http://www.jstor.org/stable/1817064

Bernard Salanié. 2005. *The Economics of Contracts: A Primer* (2nd ed.). MIT Press, Cambridge, MA.

Herbert A. Simon. 1955. A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics* 69, 1 (Feb. 1955), 99–118. https://doi.org/10.2307/1884852

Vinzenz Thoma, Barna Pasztor, Andreas Krause, Giorgia Ramponi, and Yifan Hu. 2024. Contextual Bilevel Reinforcement Learning for Incentive Alignment. In *Advances in Neural Information Processing Systems*, Vol. 37. 127369–127435. https://proceedings.neurips.cc/paper_files/paper/2024/file/e66309ead63bc1410d2df261a28f602d-Paper-Conference.pdf

Juan Pablo Vielma. 2015. Mixed Integer Linear Programming Formulation Techniques. *SIAM Rev.* 57, 1 (Jan. 2015), 3–57. https://doi.org/10.1137/130915303

Bernhard von Stengel and Shmuel Zamir. 2004. *Leadership with Commitment to Mixed Strategies*. Technical Report LSE-CDAM-2004-01. London School of Economics, London, UK.

Jibang Wu, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Zhuoran Yang, Michael I. Jordan, and Haifeng Xu. 2022. Sequential Information Design: Markov Persuasion Process and Its Efficient Reinforcement Learning. In *ACM Conference on Economics and Computation*. 471–472. https://doi.org/10.1145/3490486.3538313

Rong Yang, Fernando Ordonez, and Milind Tambe. 2012. Computing Optimal Strategy Against Quantal Response in Security Games. In *International Conference on Autonomous Agents and Multiagent Systems*, Vol. 2. 847–854.

Guanghui Yu and Chien Ju Ho. 2022. Environment Design for Biased Decision Makers. In *International Joint Conference on Artificial Intelligence*. 592–598. https://doi.org/10.24963/ijcai.2022/84

Haoqi Zhang, Yiling Chen, and David Parkes. 2009a. A General Approach to Environment Design With One Agent. In *International Joint Conference on Artificial Intelligence*. 2002–2008. https://dl.acm.org/doi/10.5555/1661445.1661765

Haoqi Zhang and David Parkes. 2008. Value-Based Policy Teaching With Active Indirect Elicitation. In *National Conference on Artificial Intelligence*, Vol. 1. 208–214. https://dl.acm.org/doi/abs/10.5555/1619995.1620030

Haoqi Zhang, David C. Parkes, and Yiling Chen. 2009b. Policy Teaching Through Reward Function Learning. In *ACM Conference on Electronic Commerce*. 295–304. https://doi.org/10.1145/1566374.1566417

Ying-Ping Zheng and Tamer Başar. 1982. Existence and Derivation of Optimal Affine Incentive Schemes for Stackelberg Games with Partial Information: A Geometric Approach. *Internat. J. Control* 35, 6 (June 1982), 997–1011. https://doi.org/10.1080/00207178208922667

A Basic Results About MDPs

Appendix A provides basic theoretical results on MDPs that support the main analysis in the paper. We begin by formalizing the relationship between policies and their induced occupancy measures, which underlies our reformulations of the reward design problem from an optimization problem over policies to one over occupancy measures. We the show that the value function can be expressed as a linear function of the occupancy measure and present a useful property of the occupancy measure.

A.1 Policies and Occupancy Measures

We present technical results that clarify how a policy induces an occupancy measure and vice versa, and highlight the special structure of deterministic occupancy measures. These results are critical for validating the linear reformulation of the follower's optimization problem and for characterizing solution sets.

A.1.1 Proof of Proposition 1.

Proof. By definition

$$\begin{aligned}
m(s,a) &= \mathbb{E}_{\pi,s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s,a_t=a)\right]\\
&= \mathbb{E}_{\pi,s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s)\cdot\pi(s,a)\right]\\
&= \pi(s,a)\cdot\mathbb{E}_{\pi,s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s)\right]\\
&= \pi(s,a)\cdot\mathbb{E}_{\pi,s_0\sim\rho}\left[\sum_{t=0}^{\infty}\sum_{a\in\mathcal{A}}\gamma^t\mathbb{P}(s_t=s,a_t=a)\right]\\
&= \pi(s,a)\cdot\sum_{a\in\mathcal{A}}\mathbb{E}_{\pi,s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s,a_t=a)\right]\\
&= \pi(s,a)\cdot\sum_{a\in\mathcal{A}}m(s,a).
\end{aligned}$$

The exchangeability of the summation is due to the absolute convergence of $\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s,a_t=a)$. If $\sum_a m(s,a)\neq 0$, the definition requires that $\pi(s,a)=\frac{m(s,a)}{\sum_a m(s,a)}$. Otherwise, $\pi(s,\cdot)$ can be an arbitrary probability distribution over $\mathcal{A}$. □

A.1.2 Occupancy Measures Induced by Deterministic Policies.

Proposition 19. *For any $\pi_1,\pi_2\in\Pi(m)$, it holds that $\mathcal{M}_{\text{det}}(\pi_1)=\mathcal{M}_{\text{det}}(\pi_2)$.*

Proof. Let $\mathcal{S}_0=\{s\in\mathcal{S}\mid\sum_{a\in\mathcal{A}}m(s,a)=0\}$. Observe that for any $\pi_1,\pi_2\in\Pi(m)$,

$$\pi_1(s,a)=\pi_2(s,a)=\frac{m(s,a)}{\sum_a m(s,a)}\quad\forall s\in\mathcal{S}\setminus\mathcal{S}_0. \tag{24}$$

Suppose a policy $\pi^1_{\text{det}}\in\Pi_{\text{det}}(\pi_1)$ induces the occupancy measure $m^1_{\text{det}}\in\mathcal{M}_{\text{det}}(\pi_1)$. Consider another policy π^2_{det} defined by

$$\pi^2_{\text{det}}(s,\cdot)\triangleq\begin{cases}\pi^1_{\text{det}}(s,\cdot) & \text{if } s\in\mathcal{S}\setminus\mathcal{S}_0,\\ \pi_{\text{det}}(s,\cdot) & \text{if } s\in\mathcal{S}_0,\end{cases} \tag{25}$$

where π_{det} is an arbitrary deterministic policy in $\Pi_{\text{det}}(\pi_2)$. Since π^1_{det} and π_{det} are both deterministic policies, it is not hard to see that π^2_{det} is also a deterministic policy.

We shall first show that $\pi^2_{\text{det}} \in \Pi_{\text{det}}(\pi_2)$. By definition, it suffices to show that $\pi^2_{\text{det}}(s,a) = 0$ when $\pi_2(s,a) = 0$. When $\pi_2(s,a) = 0$ and $s \in \mathcal{S}_0$, it holds that $\pi_{\text{det}}(s,a) = 0$ by the definition of $\Pi_{\text{det}}(\pi_2)$, which implies that $\pi^2_{\text{det}}(s,a) = 0$. On the other hand, when $\pi_2(s,a) = 0$ and $s \in \mathcal{S} \setminus \mathcal{S}_0$, it follows from (24) that $\pi_1(s,a) = 0$. From the definition of $\Pi_{\text{det}}(\pi_1)$, we know $\pi^1_{\text{det}}(s,a) = 0$, which implies that $\pi^2_{\text{det}}(s,a) = \pi^1_{\text{det}}(s,a) = 0$ according to (25).

Meanwhile, since $\pi^2_{\text{det}}(s,\cdot) = \pi^1_{\text{det}}(s,\cdot)$ when $s \in \mathcal{S} \setminus \mathcal{S}_0$, it follows from Proposition 1 that the occupancy measures induced by π^2_{det} and π^1_{det} are equal. Because π^1_{det} is selected arbitrarily from $\Pi_{\text{det}}(\pi_1)$, it follows that $\mathcal{M}_{\text{det}}(\pi_1) \subseteq \mathcal{M}_{\text{det}}(\pi_2)$. The same procedure can be used to show $\mathcal{M}_{\text{det}}(\pi_2) \subseteq \mathcal{M}_{\text{det}}(\pi_1)$, which completes the proof. □

Lemma 20. *$\mathcal{M}_{\text{det}}$ consists of all the vertices of $\mathcal{M}$.*

Before the proof, recall that for a polyhedron P, a point $x \in P$ is an extreme point if one cannot find two points $y, z \in P$ that are different from x, and a constant $\alpha \in (0,1)$ such that $x = \alpha y + (1-\alpha)z$. A point $x \in P$ is a vertex if there exists a vector c such that $\langle c, x\rangle > \langle c, y\rangle$ for all $y \in P \setminus \{x\}$. A polytope is a bounded polyhedron. It is well known that for a polytope, a vertex is equivalent to an extreme point.

Proof. Consider any $m_{\text{det}} \in \mathcal{M}_{\text{det}}$. We shall first show that m_{det} cannot be expressed as any convex combination of two other occupancy measures and hence is an extreme point of $\mathcal{M}$. Assume, for the sake of contradiction, that

$$m_{\text{det}} = \alpha m_1 + (1-\alpha) m_2$$

for some $m_1, m_2 \in \mathcal{M} \setminus \{m_{\text{det}}\}$, where $m_1 \neq m_2$, and $\alpha \in (0,1)$. Denote by π_{det}, π_1, and π_2 any policies that induce m_{det}, m_1, and m_2, respectively. Consider any $s \in \mathcal{S}$ such that $\sum_{a\in\mathcal{A}} m_{\text{det}}(s,a) \neq 0$. Because π_{det} is a deterministic policy, there exists a unique action $a^\star \in \mathcal{A}$ such that $\pi_{\text{det}}(s,a^\star) = 1$ and $\pi_{\text{det}}(s,a) = 0$ for all $a \in \mathcal{A}\setminus\{a^\star\}$. Consider any $a \in \mathcal{A}\setminus\{a^\star\}$. Since $\pi_{\text{det}}(s,a) = m_{\text{det}}(s,a)/\sum_a m_{\text{det}}(s,a)$, it follows that $m_{\text{det}}(s,a) = 0$. Because $m_1, m_2 \succeq 0$, this implies $m_1(s,a) = m_2(s,a) = 0$. It follows that π_1 and π_2 satisfy $\pi_1(s,a) = \pi_2(s,a) = 0$. Since $\sum_{a'\in\mathcal{A}} \pi_i(s,a') = 1$ for $i \in \{1,2\}$, it must hold that $\pi_1(s,a^\star) = \pi_2(s,a^\star) = 1$, which implies that $\pi_1(s,a) = \pi_2(s,a) = \pi_{\text{det}}(s,a)$ for all $a \in \mathcal{A}$. Since $s \in \mathcal{S}$ is arbitrarily selected, we know that $\pi_1(s,a) = \pi_2(s,a) = \pi_{\text{det}}(s,a)$ holds for all (s,a) such that $\sum_{a\in\mathcal{A}} m_{\text{det}}(s,a) \neq 0$. From Proposition 1, the two policies π_1, π_2 induce the same occupancy measures as π_{det}, i.e., $m_1 = m_2 = m_{\text{det}}$, which leads to a contradiction.

Suppose some $m \notin \mathcal{M}_{\text{det}}$ is a vertex of $\mathcal{M}$. From the definition of vertices, there is a vector y, viewed as a reward function, such that m is the unique occupancy measure that achieves the maximum expected reward. However, this is impossible since for any reward function in an MDP, there must exist a deterministic optimal policy. □

A.2 Expected Payoff and Occupancy Measure

Lemma 21. *Let x be a reward allocation. Suppose the follower uses a policy π, which induces an occupancy measure m. The expected payoff of the follower satisfies $\mathbb{E}_{s\sim\rho}\left[V_2^\pi(s;x)\right] = \langle r_2^x, m\rangle$.*

Proof. The expected cumulative reward of the follower satisfies

$$\begin{aligned}
\mathbb{E}_{s_0\sim\rho}\left[V_2^\pi(s)\right] &= \mathbb{E}_{\pi,s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t r_2^x(s_t,a_t)\right] \\
&= \mathbb{E}_{\pi,s_0\sim\rho}\left[\sum_{t=0}^{\infty}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\gamma^t\mathbb{P}(s_t=s,a_t=a)r_2^x(s,a)\right] \\
&= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathbb{E}_{\pi,s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s,a_t=a)\right]r_2^x(s,a) \\
&= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} m(s,a)r_2^x(s,a).
\end{aligned}$$

The exchangeability of the summation is from the absolute convergence of $\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s,a_t=a)$. □

A.3 A Useful Property of Occupancy Measures

Lemma 22. *Let m be an occupancy measure. Then $\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} m(s,a) = 1/(1-\gamma)$.*

Proof. By the definition of occupancy measures,

$$m(s,a) = \mathbb{E}_{s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s,a_t=a\mid s_0)\right].$$

Sum over the states and actions to obtain

$$\begin{aligned}
\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} m(s,a) &= \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathbb{E}_{s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s,a_t=a\mid s_0)\right] \\
&= \mathbb{E}_{s_0\sim\rho}\left[\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s,a_t=a\mid s_0)\right] \\
&= \mathbb{E}_{s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathbb{P}(s_t=s,a_t=a\mid s_0)\right] \\
&= \mathbb{E}_{s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t\right] \\
&= \frac{1}{1-\gamma}.
\end{aligned}$$

It was possible to change the order of summation because $\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=s,a_t=a\mid s_0)\le 1/(1-\gamma)$, which is an absolutely convergent series. □

B Proofs on the Robustness to Nonunique Best Response

Appendix B provides the proof and technical details for the robustness analysis of interior-point allocations under nonunique best responses and justifies the correctness of using (23) to verify robustness in numerical experiments.

B.1 Proof of Proposition 8

Lemma 23. *Let v be a vector in $\mathbb{R}^{|\mathcal{D}|}$. Suppose $\bar{x}$ is an interior point of an allocation region $\mathcal{P}_m$, i.e., $\bar{x} + cv \in \mathcal{P}_m$ for some constant $c > 0$ and for all $\|v\|_1 \leq 1$. For any $m' \in \mathcal{M}$, it holds that*

$$\langle r_2^{\bar{x}}, m - m' \rangle \geq c \cdot \max_{(s,a)\in\mathcal{D}} |m(s,a) - m'(s,a)| .$$

Proof. Consider any $m' \in \mathcal{M}$. It follows from the definition of $\mathcal{P}_m$ that

$$\begin{aligned} 0 &\leq \min_{\|v\|_1=1} \left\{ \langle r_2^{\bar{x}+cv}, m \rangle - \langle r_2^{\bar{x}+cv}, m' \rangle \right\} \\ &= \min_{\|v\|_1=1} \langle r_2^{\bar{x}+cv}, m - m' \rangle \\ &= \min_{\|v\|_1=1} \left\{ \langle r_2^{\bar{x}}, m - m' \rangle + c\langle \delta^v, m - m' \rangle \right\} \\ &= \langle r_2^{\bar{x}}, m - m' \rangle + c \min_{\|v\|_1=1} \langle \delta^v, m - m' \rangle, \end{aligned}$$

where δ^v is defined by

$$\delta^v(s,a) = \begin{cases} v(s,a) & \text{if } (s,a) \in \mathcal{D}, \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

Therefore, we have

$$\begin{aligned} \langle r_2^{\bar{x}}, m - m' \rangle &\geq -c \min_{\|v\|_1=1} \langle \delta^v, m - m' \rangle \\ &= c \max_{\|v\|_1=1} \langle \delta^v, m' - m \rangle \\ &\geq c \cdot \max_{(s,a)\in\mathcal{D}} |m(s,a) - m'(s,a)| , \end{aligned}$$

which completes the proof. □

Proof of Proposition 8. Since $x^\star$ is an optimal interior-point allocation, there exists an optimal occupancy measure $m^\star$ such that $x^\star$ is an interior point of $\mathcal{P}_{m^\star}$. By Lemma 23, there exists $c > 0$ such that

$$\langle r_2^{x^\star}, m^\star - m \rangle \geq c \cdot \max_{(s,a)\in\mathcal{D}} \left| m(s,a) - m^\star(s,a) \right| \quad \forall m \in \mathcal{M}.$$

When $m \in \mathrm{BR}(x^\star)$, it holds that $\langle r_2^{x^\star}, m \rangle = \langle r_2^{x^\star}, m^\star \rangle$, i.e., $\langle r_2^{x^\star}, m^\star - m \rangle = 0$. The left-hand side of the inequality evaluates to zero. This implies

$$m(s,a) - m^\star(s,a) = 0 \quad \forall (s,a) \in \mathcal{D}.$$

By Assumption 5, it holds that $\langle r_1, m - m^\star \rangle = \sum_{(s,a)\in\mathcal{D}} r_1(s,a)(m(s,a) - m^\star(s,a)) = 0$, implying $\langle r_1, m \rangle = \langle r_1, m^\star \rangle$. □

B.2 Proof of Proposition 18

Proof. We first show the *if* direction. Suppose that $x^\star$ satisfies $\langle r_1, m \rangle = v_1^\star$ for all $m \in \mathrm{BR}(x^\star)$. Consider any $\epsilon > 0$ and m satisfying $\langle r_2^{x^\star}, m^\star \rangle - \langle r_2^{x^\star}, m \rangle \leq \epsilon$. Let $b = \langle r_2^{x^\star}, m^\star \rangle - \max_{m\in\mathcal{M}_{\mathrm{det}}\setminus\mathrm{BR}(x^\star)} \langle r_2^{x^\star}, m \rangle$ and $\mathcal{M}_{\mathrm{det}}^\star = \mathcal{M}_{\mathrm{det}} \cap \mathrm{BR}(x^\star)$. By Lemma 20, we can express m as a convex combination of deterministic occupancy measures: $m = \sum_{i=1}^h \lambda_i m_i + \sum_{j=1}^l \lambda_j m'_j$, where $m_i \in \mathcal{M}_{\mathrm{det}}^\star$ for $i = 1, 2, \ldots, h$, $m'_j \in \mathcal{M}_{\mathrm{det}} \setminus \mathcal{M}_{\mathrm{det}}^\star$ for $j = 1, 2, \ldots, l$, $\lambda_i, \lambda_j \geq 0$

and $\sum_{i=1}^{h}\lambda_i+\sum_{j=1}^{l}\lambda_j=1$. For any $i\in\{1,2,\ldots,h\}$, because $m_i\in\mathcal{M}_{\text{det}}^{\star}$, it holds that $\langle r_2^{x^\star},m_i\rangle=\langle r_2^{x^\star},m^\star\rangle$. For any $j=\{1,2,\ldots,l\}$, because $m'_j\in\mathcal{M}_{\text{det}}\setminus\mathcal{M}_{\text{det}}^{\star}=\mathcal{M}_{\text{det}}\setminus\text{BR}(x^\star)$, it holds that $\langle r_2^{x^\star},m^\star\rangle-\langle r_2^{x^\star},m'_j\rangle\geq b$. Therefore,

$$\begin{aligned}\langle r_2^{x^\star},m^\star\rangle-\langle r_2^{x^\star},m\rangle&=\langle r_2^{x^\star},m^\star\rangle-\langle r_2^{x^\star},\sum_{i=1}^{h}\lambda_i m_i\rangle-\langle r_2^{x^\star},\sum_{j=1}^{l}\lambda_j m'_j\rangle\\&=\langle r_2^{x^\star},m^\star\rangle-\sum_{i=1}^{h}\lambda_i\langle r_2^{x^\star},m_i\rangle-\sum_{j=1}^{l}\lambda_j\langle r_2^{x^\star},m'_j\rangle\\&=\sum_{i=1}^{h}\lambda_i\langle r_2^{x^\star},m^\star-m_i\rangle+\sum_{j=1}^{l}\lambda_j\langle r_2^{x^\star},m^\star-m'_j\rangle\\&=\sum_{j=1}^{l}\lambda_j\langle r_2^{x^\star},m^\star-m'_j\rangle\\&\geq b\cdot\sum_{j=1}^{l}\lambda_j.\end{aligned}$$

This implies that $\epsilon\geq b\cdot\sum_{j=1}^{l}\lambda_j$, i.e., $\sum_{j=1}^{l}\lambda_j\leq\epsilon/b$. In the meantime, because $m_i\in\mathcal{M}_{\text{det}}^{\star}$, it follows from the given assumption that $\langle r_1,m_i\rangle=v_1^\star=\langle r_1,m^\star\rangle$. Thus,

$$\begin{aligned}\langle r_1,m^\star\rangle-\langle r_1,m\rangle&=\langle r_1,m^\star\rangle-\langle r_1,\sum_{i=1}^{h}\lambda_i m_i\rangle-\langle r_1,\sum_{j=1}^{l}\lambda_j m'_j\rangle\\&=\sum_{i=1}^{h}\lambda_i\langle r_1,m^\star-m_i\rangle+\sum_{j=1}^{l}\lambda_j\langle r_1,m^\star-m'_j\rangle\\&=\sum_{j=1}^{l}\lambda_j\langle r_1,m^\star-m'_j\rangle.\end{aligned}$$

When $\epsilon\to 0^+$, it follows from $\sum_{j=1}^{l}\lambda_j\leq\epsilon/b$ that $\sum_{j=1}^{l}\lambda_j\to 0^+$, which further implies that $\langle r_1,m^\star\rangle-\langle r_1,m\rangle\to 0^+$. Since m is selected arbitrarily such that $\langle r_2^{x^\star},m^\star\rangle-\langle r_2^{x^\star},m\rangle\leq\epsilon$, this implies that $v_1^\star-v_\epsilon(x^\star)\to 0^+$ when $\epsilon\to 0^+$.

For the *only if* direction, we assume for the sake of contradiction that there exists $m\in\text{BR}(x^\star)$ such that $\langle r_1,m^\star\rangle-\langle r_1,m\rangle=c>0$. Then m is a feasible solution to (23), which implies $v_1^\star-v_\epsilon(x^\star)\geq c$ for any $\epsilon>0$. This contradicts the given assumption that $v_1^\star-v_\epsilon(x^\star)\to 0^+$ when $\epsilon\to 0^+$. □

C Proofs on the Robustness to Bounded Rationality

Appendix C contains proofs supporting the robustness of interior-point allocations when the follower exhibits bounded rationality. For each of the three models of irrational behavior, we derive a lower bound on the leader's cumulative reward relative to the fully rational case.

C.1 Model of Bounded Rationality

In the original entropy-regularized MDP, the follower chooses a policy π to maximize

$$V_\tau^\pi(\rho)\triangleq V^\pi(\rho)+\tau\cdot\mathcal{H}(\rho,\pi),$$

where $V^\pi(\rho) = \mathbb{E}_{s\sim\rho}\left[V_2^\pi(s)\right]$ and

$$\mathcal{H}(\rho,\pi) \triangleq \mathbb{E}_{\pi,s_0\sim\rho}\left[\sum_{t=0}^{\infty} -\gamma^t \log\pi(s_t,a_t)\right].$$

Let m be the occupancy measure induced by π. It then holds that

$$\mathcal{H}(\rho,\pi) = -\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} m(s,a)\log\pi(s,a) = -\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} m(s,a)\log\frac{m(s,a)}{\sum_{a'\in\mathcal{A}} m(s,a')},$$

where the last equality follows from Proposition 1. Combine Lemma 21 to obtain the payoff function in (12).

Proof of Proposition 10. Notice that

$$\begin{aligned}
&-\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left(m^\star(s,a)\log\frac{m^\star(s,a)}{\sum_{a'\in\mathcal{A}} m^\star(s,a')}\right)\\
&\quad= -\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left(\left(\sum_{a\in\mathcal{A}} m^\star(s,a)\right)\left(\frac{m^\star(s,a)}{\sum_{a'\in\mathcal{A}} m^\star(s,a')}\right)\log\frac{m^\star(s,a)}{\sum_{a'\in\mathcal{A}} m^\star(s,a')}\right)\\
&\quad= -\sum_{s\in\mathcal{S}}\left(|\mathcal{A}|\sum_{a\in\mathcal{A}} m^\star(s,a)\right)\sum_{a\in\mathcal{A}}\left(\left(\frac{m^\star(s,a)}{\sum_{a'\in\mathcal{A}} m^\star(s,a')}\right)\log\frac{m^\star(s,a)}{\sum_{a'\in\mathcal{A}} m^\star(s,a')}\right)\\
&\quad\le \sum_{s\in\mathcal{S}}\left(|\mathcal{A}|\sum_{a\in\mathcal{A}} m^\star(s,a)\cdot\frac{1}{|\mathcal{A}|}\log|\mathcal{A}|\right)\\
&\quad= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} m^\star(s,a)\log|\mathcal{A}|\\
&\quad= \frac{1}{1-\gamma}\log|\mathcal{A}|.
\end{aligned}$$

This implies

$$\tau\cdot\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} m^\star(s,a)\log\frac{m^\star(s,a)}{\sum_{a'\in\mathcal{A}} m^\star(s,a')} - \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} m_\tau^\star(s,a)\log\frac{m_\tau^\star(s,a)}{\sum_{a'\in\mathcal{A}} m_\tau^\star(s,a')}\right) \le 2\tau\cdot\frac{1}{1-\gamma}\cdot\log|\mathcal{A}|. \tag{27}$$

It follows from the optimality of $m_\tau^\star$ that

$$\tau\cdot\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} m^\star(s,a)\log\frac{m^\star(s,a)}{\sum_{a'\in\mathcal{A}} m^\star(s,a')} - \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} m_\tau^\star(s,a)\log\frac{m_\tau^\star(s,a)}{\sum_{a'\in\mathcal{A}} m_\tau^\star(s,a')}\right) \ge \langle r_2^{x^\star}, m^\star\rangle - \langle r_2^{x^\star}, m_\tau^\star\rangle. \tag{28}$$

Combine (27) and (28) to obtain

$$\langle r_2^{x^\star}, m^\star\rangle - \langle r_2^{x^\star}, m_\tau^\star\rangle \le 2\tau\cdot\frac{1}{1-\gamma}\cdot\log|\mathcal{A}|. \tag{29}$$

Let $h = |\mathcal{M}^\star_{\text{det}}|$ and $l = |\mathcal{M}_{\text{det}} \setminus \mathcal{M}^\star_{\text{det}}|$. Since $m^\star_\tau \in \mathcal{M}$, according to Lemma 20, we can write $m^\star_\tau = \sum_{i=1}^h \lambda_i m_i + \sum_{j=1}^l \lambda_j m'_j$, where $m_i \in \mathcal{M}^\star_{\text{det}}$ for $i = 1, 2, \ldots, h$, $m'_j \in \mathcal{M}_{\text{det}} \setminus \mathcal{M}^\star_{\text{det}}$ for $j = 1, 2, \ldots, l$, $\lambda_i, \lambda_j \geq 0$ and $\sum_{i=1}^h \lambda_i + \sum_{j=1}^l \lambda_j = 1$. We can rewrite the left side of (29) as

$$\begin{aligned}
\langle r_2^{x^\star}, m^\star \rangle - \langle r_2^{x^\star}, m^\star_\tau \rangle &= \langle r_2^{x^\star}, m^\star \rangle - \langle r_2^{x^\star}, \sum_{i=1}^h \lambda_i m_i + \sum_{j=1}^l \lambda_j m'_j \rangle \\
&= (1 - \sum_{i=1}^h \lambda_i) \langle r_2^{x^\star}, m^\star \rangle - \sum_{j=1}^l \lambda_j \langle r_2^{x^\star}, m'_j \rangle \\
&= \sum_{j=1}^l \lambda_j \langle r_2^{x^\star}, m^\star \rangle - \sum_{j=1}^l \lambda_j \langle r_2^{x^\star}, m'_j \rangle \\
&= \sum_{j=1}^l \lambda_j \langle r_2^{x^\star}, m^\star - m'_j \rangle.
\end{aligned}$$

The second equality is because $\langle r_2^{x^\star}, m^\star \rangle = \langle r_2^{x^\star}, m_i \rangle$ for $i = 1, 2, \ldots, h$, which is obtained by the assumption that $m_i \in \mathcal{M}^\star_{\text{det}} = \mathcal{M}_{\text{det}} \cap \text{BR}(x^\star) \subseteq \text{BR}(x^\star)$. It then follows that

$$\sum_{j=1}^l \lambda_j \langle r_2^{x^\star}, m^\star - m'_j \rangle \leq 2\tau \cdot \frac{1}{1-\gamma} \cdot \log |\mathcal{A}|. \tag{30}$$

Since $\mathcal{M}_{\text{det}} \setminus \mathcal{M}^\star_{\text{det}}$ is a finite set and hence $\max_{m \in \mathcal{M}_{\text{det}} \setminus \mathcal{M}^\star_{\text{det}}} \langle r_2^{x^\star}, m \rangle = b$ exists, one can obtain

$$b \sum_{j=1}^l \lambda_j \leq \sum_{j=1}^l \lambda_j \langle r_2^{x^\star}, m^\star - m'_j \rangle.$$

Combine with (30) to obtain $b \sum_{j=1}^l \lambda_j \leq 2\tau \cdot \frac{1}{1-\gamma} \cdot \log |\mathcal{A}|$, i.e., $\sum_{j=1}^l \lambda_j \leq 2\tau \cdot \frac{1}{b(1-\gamma)} \cdot \log |\mathcal{A}|$. Therefore,

$$\begin{aligned}
\langle r_1, m^\star_\tau \rangle &= \langle r_1, \sum_{i=1}^h \lambda_i m_i + \sum_{j=1}^l \lambda_j m'_j \rangle \\
&= \sum_{i=1}^h \lambda_i \langle r_1, m^\star \rangle + \sum_{j=1}^l \lambda_j \langle r_1, m'_j \rangle \\
&\geq \sum_{i=1}^h \lambda_i \langle r_1, m^\star \rangle \\
&= (1 - \sum_{j=1}^l \lambda_j) \langle r_1, m^\star \rangle \\
&\geq \left(1 - 2\tau \cdot \frac{1}{b(1-\gamma)} \cdot \log |\mathcal{A}|\right) \langle r_1, m^\star \rangle.
\end{aligned}$$

The second equality follows from the assumption that $x^\star$ is an optimal interior-point allocation. From Proposition 8, any best response of $x^\star$ is an optimal occupancy measure. Thus, $\langle r_1, m_i \rangle = \langle r_1, m^\star \rangle$ since $m_i \in \text{BR}(x^\star)$.

□

C.2 Alternative Model of Bounded Rationality

Proposition 24. *Suppose that Assumption 5 holds and that $(x^\star, m^\star)$ is an optimal solution to the reward design problem in (5) with $x^\star$ being an interior point of $\mathcal{P}_{m^\star}$. Define $\mathcal{M}^\star_{\text{det}} \triangleq \mathcal{M}_{\text{det}} \cap \text{BR}(x^\star)$, the set of optimal deterministic occupancy measures under $x^\star$. Let*

$$m^\star_\tau = \arg\max_{m \in \mathcal{M}} \left\{ \langle r_2^{x^\star}, m \rangle - \tau \cdot \text{Ent}(m) \right\}.$$

Then for any $\tau > 0$, it holds that

$$\langle r_1, m^\star_\tau \rangle \geq \left(1 - \frac{2\tau}{b(1-\gamma)} |\log\left(|\mathcal{S}||\mathcal{A}| \cdot (1-\gamma)\right)|\right) \langle r_1, m^\star \rangle,$$

where γ is the discount factor, and $b = \langle r_2^{x^\star}, m^\star \rangle - \max_{m \in \mathcal{M}_{\text{det}} \setminus \mathcal{M}^\star_{\text{det}}} \langle r_2^{x^\star}, m \rangle$.

Proof. Notice that

$$\begin{aligned} \tau \cdot \left(\text{Ent}(m^\star) - \text{Ent}(m^\star_\tau)\right) &\leq \tau |\text{Ent}(m^\star) + \text{Ent}(m^\star_\tau)| \\ &= 2\tau \left(\frac{1}{1-\gamma} |\log\left(|\mathcal{S}||\mathcal{A}| \cdot (1-\gamma)\right)|\right). \end{aligned}$$

The upper bound of $\text{Ent}(m^\star) + \text{Ent}(m^\star_\tau)$ is derived from the facts that $m \in \mathcal{M}$ is nonnegative and the sum over states and actions is bounded by $\frac{1}{1-\gamma}$ by Lemma 22. The rest of the proof is similar to the proof of Proposition 10. □

C.3 δ-optimal Response as a Model for Bounded Rationality

Consider a follower who chooses a worst response for the leader among all the δ-optimal responses. Denote by $\text{BR}_\delta(x) \triangleq \left\{ m \mid \max_{m' \in \mathcal{M}} \langle r_2^x, m' \rangle - \langle r_2^x, m \rangle \leq \delta \right\}$ the set of δ-optimal responses under a reward allocation x. The payoff of the leader under x is given by

$$\min_{m \in \text{BR}_\delta(x)} \langle r_1, m \rangle.$$

The following proposition shows that an optimal interior-point allocation is robust against δ-optimal responses.

Proposition 25. *Suppose that Assumption 5 holds and that $(x^\star, m^\star)$ is an optimal solution to the reward design problem in (5), with $x^\star$ being an interior point of $\mathcal{P}_{m^\star}$. Let $m^\star_\delta \in \arg\min_{m \in \text{BR}_\delta(x^\star)} \langle r_1, m \rangle$ and $\mathcal{M}^\star_{\text{det}} = \mathcal{M}_{\text{det}} \cap \text{BR}(x^\star)$. Then for any $\delta > 0$, the value $\langle r_1, m^\star_\delta \rangle$ satisfies*

$$\langle r_1, m^\star_\delta \rangle \geq \left(1 - \frac{\delta}{b}\right) \langle r_1, m^\star \rangle,$$

where $b = \langle r_2^{x^\star}, m^\star \rangle - \max_{m \in \mathcal{M}_{\text{det}} \setminus \mathcal{M}^\star_{\text{det}}} \langle r_2^{x^\star}, m \rangle$.

Proof. Notice that

$$\langle r_2^{x^\star}, m^\star \rangle - \langle r_2^{x^\star}, m^\star_\delta \rangle \leq \delta$$

according to the definition of $\text{BR}_\delta(x^\star)$. The rest of the proof is similar to the proof of Proposition 10. □

D Proofs on the Optimality of Deterministic Occupancy Measures

Appendix D gives proofs of Propositions 12 and 13, which show that one can restrict attention to deterministic occupancy measures without loss of optimality.

D.1 Proof of Proposition 12

Lemma 26. *Let Q^{opt} be the optimal Q-function of an MDP, and π^{opt} be any deterministic greedy policy with respect to Q^{opt}, i.e., for all $s \in \mathcal{S}$, it holds that $\pi^{\mathrm{opt}}(s,a) = 1$ for some $a \in \arg\max_{a' \in \mathcal{A}} Q^{\mathrm{opt}}(s,a')$. For any optimal policy π (not necessarily deterministic) of the MDP, it holds that*

$$\mathbb{E}_{a \sim \pi(s,\cdot)} \left[Q^{\mathrm{opt}}(s,a) \right] = \max_{a' \in \mathcal{A}} Q^{\mathrm{opt}}(s,a') \quad \forall s \in \mathcal{S}.$$

Proof. Let $d_s^\pi(s') = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s' \mid s_0 = s)$. By the performance difference lemma (Kakade and Langford, 2002),

$$\begin{aligned} V^\pi(s) - V^{\mathrm{opt}}(s) &= \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_s^\pi} \left[\mathbb{E}_{a \sim \pi(s',\cdot)} Q^{\mathrm{opt}}(s',a) - V^{\mathrm{opt}}(s') \right] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_s^\pi} \left[\mathbb{E}_{a \sim \pi(s',\cdot)} Q^{\mathrm{opt}}(s',a) - \max_{a' \in \mathcal{A}} Q^{\mathrm{opt}}(s',a') \right]. \end{aligned}$$

Meanwhile, it follows from the optimality of π that $V^\pi(s) - V^{\mathrm{opt}}(s) = 0$, which implies that

$$\mathbb{E}_{s' \sim d_s^\pi} \left[\mathbb{E}_{a \sim \pi(s',\cdot)} Q^{\mathrm{opt}}(s',a) - \max_{a' \in \mathcal{A}} Q^{\mathrm{opt}}(s',a') \right] = 0.$$

Because $\mathbb{E}_{a \sim \pi(s',\cdot)} Q^{\mathrm{opt}}(s',a) - \max_{a' \in \mathcal{A}} Q^{\mathrm{opt}}(s',a') \le 0$ for all $s' \in \mathcal{S}$, it must hold that

$$\mathbb{E}_{a \sim \pi(s',\cdot)} Q^{\mathrm{opt}}(s',a) - \max_{a' \in \mathcal{A}} Q^{\mathrm{opt}}(s',a') = 0$$

for all $s' \in \mathcal{S}$. □

Proof of Proposition 12. If $\pi^\star \in \Pi_{\mathrm{det}}$, then $\Pi_{\mathrm{det}}(\pi^\star) = \{\pi^\star\}$ by definition. Therefore, any $\pi \in \Pi_{\mathrm{det}}(\pi^\star)$ trivially satisfies $\langle r_2^{x^\star}, m^\pi \rangle = \langle r_2^{x^\star}, m^\star \rangle = \max_{m' \in \mathcal{M}} \langle r_2^{x^\star}, m' \rangle$.

Consider the case when $\pi^\star \notin \Pi_{\mathrm{det}}$. Let V_2^{opt} and Q_2^{opt} be the optimal value function and the optimal Q-function of the follower's MDP, respectively. Because $\pi^\star$ is an optimal policy, it follows from Lemma 26 that

$$\mathbb{E}_{a \sim \pi^\star(s,\cdot)} Q_2^{\mathrm{opt}}(s,a) = \max_{a' \in \mathcal{A}} Q_2^{\mathrm{opt}}(s,a') \quad \forall s \in \mathcal{S}. \tag{31}$$

Hence, if $\pi^\star(s,a) \neq 0$ for some $(s,a) \in \mathcal{S} \times \mathcal{A}$, then

$$Q_2^{\mathrm{opt}}(s,a) = \max_{a' \in \mathcal{A}} Q_2^{\mathrm{opt}}(s,a') = V_2^{\mathrm{opt}}(s). \tag{32}$$

For any $s \in \mathcal{S}$, let

$$\mathcal{A}_s^\star = \{a \mid \pi^\star(s,a) \in (0,1)\},$$

and consider the following two cases.

Case 1: Consider $s \in \mathcal{S}$ such that $\mathcal{A}_s^\star$ is nonempty. Since $\pi \in \Pi_{\mathrm{det}}(\pi^\star)$ is deterministic, there exists some $a^\star \in \mathcal{A}$ such that $\pi(s,a^\star) = 1$. From the definition of $\Pi_{\mathrm{det}}(\pi^\star)$, it follows that $\pi^\star(s,a^\star) \neq 0$. Use (32) to obtain $Q_2^{\mathrm{opt}}(s,a^\star) = V_2^{\mathrm{opt}}(s)$. This implies

$$\mathbb{E}_{a \sim \pi(s,\cdot)} Q_2^{\mathrm{opt}}(s,a) = Q_2^{\mathrm{opt}}(s,a^\star) = V_2^{\mathrm{opt}}(s).$$

Case 2: Consider $s \in \mathcal{S}$ such that $\mathcal{A}_s^\star$ is empty, i.e., $\pi^\star(s,a) \in \{0,1\}$ for all $a \in \mathcal{A}$. Since $\pi \in \Pi_{\mathrm{det}}(\pi^\star)$, one must have $\pi(s,a) = \pi^\star(s,a)$ for all $a \in \mathcal{A}$. Use (31) to obtain

$$\mathbb{E}_{a \sim \pi(s,\cdot)} Q_2^{\mathrm{opt}}(s,a) = \mathbb{E}_{a \sim \pi^\star(s,\cdot)} Q_2^{\mathrm{opt}}(s,a) = \max_{a' \in \mathcal{A}} Q_2^{\mathrm{opt}}(s,a') = V_2^{\mathrm{opt}}(s).$$

Therefore, for all $s \in \mathcal{S}$, it holds that $\mathbb{E}_{a\sim\pi(s,\cdot)} Q_2^{\text{opt}}(s,a) = V_2^{\text{opt}}(s)$. Applying the performance difference lemma again, one can obtain that for all $s_0 \in \mathcal{S}$,

$$V_2^{\pi}(s_0) - V_2^{\text{opt}}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi}}\left[\mathbb{E}_{a\sim\pi(s,\cdot)} Q_2^{\text{opt}}(s,a) - V_2^{\text{opt}}(s)\right] = 0.$$

This implies $\mathbb{E}_{s_0\sim\rho}\left[V_2^{\pi}(s_0; x^\star)\right] = \mathbb{E}_{s_0\sim\rho}\left[V_2^{\text{opt}}(s_0; x^\star)\right]$, equivalently, $\langle r_2^{x^\star}, m^\pi\rangle = \langle r_2^{x^\star}, m^\star\rangle = \max_{m'\in\mathcal{M}}\langle r_2^{x^\star}, m'\rangle$, from which one obtains $m^\pi \in \text{BR}(x^\star)$. □

D.2 Proof of Proposition 13

Proof. If $\pi^\star \in \Pi_{\text{det}}$, then $\Pi_{\text{det}}(\pi^\star) = \{\pi^\star\}$, and the proposition holds trivially. Consider the case when $\pi^\star \notin \Pi_{\text{det}}$. For any $s \in \mathcal{S}$, let

$$\mathcal{A}_s = \{a \in \mathcal{A} \mid \pi^\star(s,a) \in (0,1)\}.$$

Because $\pi^\star \notin \Pi_{\text{det}}$, there exists $s' \in \mathcal{S}$ such that $\mathcal{A}_{s'}$ is nonempty. Choose $a' \in \mathcal{A}_{s'}$ arbitrarily. Define a policy π' as follows: $\pi'(s,\cdot) = \pi^\star(s,\cdot)$ when $s \neq s'$, and $\pi'(s,a') = 1$ when $s = s'$. We shall first show that $\pi' \in \text{BR}(x^\star)$. By Lemma 26, since $\pi^\star$ is optimal,

$$\mathbb{E}_{a\sim\pi^\star(s',\cdot)}\left[Q_2^{\text{opt}}(s',a)\right] = \max_{a\in\mathcal{A}} Q_2^{\text{opt}}(s',a).$$

This implies that $Q_2^{\text{opt}}(s',a) = \max_{a\in\mathcal{A}} Q_2^{\text{opt}}(s',a)$ for all $a \in \mathcal{A}_{s'}$. Because $a' \in \mathcal{A}_{s'}$, it holds that

$$\mathbb{E}_{a\sim\pi'(s',\cdot)}\left[Q_2^{\text{opt}}(s',a)\right] = Q_2^{\text{opt}}(s',a') = \max_{a\in\mathcal{A}} Q_2^{\text{opt}}(s',a).$$

Meanwhile, for (s,a) such that $s \neq s'$, since $\pi'(s,a) = \pi^\star(s,a)$,

$$\mathbb{E}_{a\sim\pi'(s,\cdot)}\left[Q_2^{\text{opt}}(s,a)\right] = \mathbb{E}_{a\sim\pi^\star(s,\cdot)}\left[Q_2^{\text{opt}}(s,a)\right] = \max_{a\in\mathcal{A}} Q_2^{\text{opt}}(s,a).$$

By the performance difference lemma, π' is an optimal policy of the follower.

Define $V_1^\star \triangleq V_1^{\pi^\star}$ and $Q_1^\star \triangleq Q_1^{\pi^\star}$. Because π' is an optimal policy of the follower and hence a feasible solution to (1), it follows from the optimality of $\pi^\star$ that

$$V_1^{\pi'}(s_0) - V_1^\star(s_0) \leq 0. \tag{33}$$

By the performance difference lemma,

$$V_1^{\pi'}(s_0) - V_1^\star(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^{\pi'}}\left[\mathbb{E}_{a\sim\pi'(s,\cdot)} Q_1^\star(s,a) - V_1^\star(s)\right]. \tag{34}$$

Since $\pi'(s,\cdot) = \pi^\star(s,\cdot)$ for any $s \neq s'$, we know that $\mathbb{E}_{a\sim\pi'(s,\cdot)} Q_1^\star(s,a) - V_1^\star(s) = 0$ for $s \neq s'$. It then follows from (33) and (34) that $\mathbb{E}_{a\sim\pi'(s,\cdot)} Q_1^\star(s,a) - V_1^\star(s) \leq 0$ when $s = s'$, i.e.,

$$Q_1^\star(s',a') \leq V_1^\star(s').$$

Since $s' \in \mathcal{S}$ and $a' \in \mathcal{A}_{s'}$ were chosen arbitrarily, it holds that $Q_1^\star(s,a) \leq V_1^\star(s)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ such that $\pi^\star(s,a) \in (0,1)$. However, the inequality can never be achieved, else

$$\mathbb{E}_{a\sim\pi^\star(s,\cdot)} Q_1^\star(s,a) - V_1^\star(s) < 0,$$

which violates the Bellman consistency equation. In other words, $Q_1^\star(s,a) = V_1^\star(s)$ when $\pi^\star(s,a) \in (0,1)$. In addition, $Q_1^\star(s,a) = V_1^\star(s)$ when $\pi^\star(s,a) = 1$. Combine the two conditions to obtain $Q_1^\star(s,a) = V_1^\star(s)$ when

$\pi^\star(s,a) \neq 0$. Consider any $\pi \in \Pi_{\text{det}}(\pi^\star)$. When $\pi(s,a) = 1$, it must hold that $\pi^\star(s,a) \neq 0$, and consequently $Q_1^\star(s,a) = V_1^\star(s)$. For any $s_0 \in \mathcal{S}$, by the performance difference lemma,

$$V_1^\pi(s_0) - V_1^\star(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^\pi}\left[\mathbb{E}_{a\sim\pi(s,\cdot)}Q_1^\star(s,a) - V_1^\star(s)\right]$$
$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{s_0}^\pi}\left[V_1^\star(s) - V_1^\star(s)\right] = 0.$$

This implies $\mathbb{E}_{s_0\sim\rho}\left[V_1^\pi(s_0)\right] = \mathbb{E}_{s_0\sim\rho}\left[V_1^\star(s_0)\right]$ or, equivalently, $\langle r_1, m^\pi\rangle = v_1^\star$. □

E Proofs of the Existence of Optimal Interior-point Allocation

In Appendix E, we first give a proof of Theorem 16, which establishes a sufficient condition for the existence of an optimal interior-point allocation via the solution of the margin-maximization problem (16). Then, we give a proof of Theorem 17, which further shows that an interior-point solution must exist if there exists an optimal allocation that is robust against the follower's nonunique best responses.

E.1 Proof of Theorem 16

Lemma 27. *Let X and Y be two sets. Suppose that X is convex with a nonempty interior and that Y is open. Then either $X \cap Y = \emptyset$, or $X \cap Y$ has a nonempty interior.*

Proof. The proof is reproduced from a post on StackExchange (see Footnote[2]). Within this proof, for a given set X, we shall denote its interior by X°, its boundary by ∂X, its complement by X^C, and its closure by $\overline{X}$.

We first show that $\partial X = \partial(X^\circ)$. Let $c \in X^\circ$ and $x \in \partial X$. Without loss of generality, we assume that $x = 0$ by a change of coordinates. Then $\lambda c \in X$ for $0 \leq \lambda \leq 1$ by convexity. Because $c \in X^\circ$, there is an open set $U \subseteq X^\circ$ that contains c. Define $U_\lambda \triangleq \{y \mid y = \lambda u,\ u \in U\}$. It is not hard to verify that when $0 < \lambda \leq 1$, the set U_λ is open, contains λc, and $U_\lambda \subseteq X$ by the convexity of X, which implies $\lambda c \in X^\circ$ when $0 < \lambda \leq 1$. This shows $x \in \overline{X^\circ}$. Meanwhile, since $x \in \partial X$, it holds that $x \in \overline{X^C} \subseteq \overline{X^{\circ C}}$. We obtain that $x \in \overline{X^\circ} \cap \overline{X^{\circ C}} = \partial(X^\circ)$, which implies $\partial X \subseteq \partial(X^\circ)$. For the other direction, recall that $\overline{X^C} = X^{\circ C}$ for any X. Use this result to obtain $\overline{X} = \overline{(X^C)^C} = X^{C\circ C}$, $\overline{X^\circ} = \overline{(X^{\circ C})^C} = X^{\circ C\circ C}$, and $\overline{X^{\circ C}} = X^{\circ\circ C} = X^{\circ C}$. Thus,

$$\partial X = \overline{X} \cap \overline{X^C} = X^{\circ C} \cap X^{C\circ C},$$

and

$$\partial(X^\circ) = \overline{X^\circ} \cap \overline{X^{\circ C}} = X^{\circ C} \cap X^{\circ C\circ C}.$$

From the fact $X^{C\circ C} \supseteq X^{\circ C\circ C}$, we obtain $\partial X \supseteq \partial(X^\circ)$.

Suppose that $X \cap Y$ is nonempty, and let $x \in X \cap Y$. If $x \in X^\circ$, then $X^\circ \cap Y \neq \emptyset$ is a nonempty open set in $X \cap Y$. If $x \in \partial X$, then it follows from the previous discussion that $x \in \partial(X^\circ)$. Because Y is open, there exists an open ball $B(x) \subseteq Y$ centered at x. Moreover, $B(x) \cap X^\circ \neq \emptyset$ since $x \in \partial(X^\circ)$. Therefore, $B(x) \cap X^\circ$ is a nonempty open set in $X \cap Y$. □

Lemma 28. *Given a function $v\colon \mathcal{D} \to \mathbb{R}$ and an occupancy measure m, let δ^v be defined by (26). Then $|\langle \delta^v, m\rangle| \leq \|v\|_1 \max_{(s,a)\in\mathcal{D}} m(s,a)$.*

Proof. According to the definition,

$$\langle \delta^v, m\rangle = \sum_{(s,a)\in\mathcal{D}} \delta^v(s,a)m(s,a) = \sum_{(s,a)\in\mathcal{D}} v(s,a)m(s,a). \tag{35}$$

Apply Hölder's inequality to complete the proof. □

[2] https://math.stackexchange.com/q/2701244

Proof of Theorem 16. (Sufficiency) Let (x^0, m^0) be an optimal solution of (16) such that $C - \sum_{(s,a)\in\mathcal{D}} x^0(s,a) > 0$. Define the set

$$\mathcal{M}^\star(x^0) \triangleq \{m \mid \langle r_1, m\rangle = \langle r_1, m^0\rangle\} \cap (\mathcal{M}_{\text{det}} \cap \text{BR}(x^0)) \tag{36}$$

and the constant

$$b \triangleq \langle r_2^{x^0}, m^0\rangle - \max_{m'\in\mathcal{M}_{\text{det}}\setminus\text{BR}(x^0)} \langle r_2^{x^0}, m'\rangle > 0.$$

Consider $x' = x^0 + \epsilon\cdot\frac{(1-\gamma)}{2|\mathcal{D}|}\cdot h(r_1)$, where $\epsilon = \min\left\{b, \frac{1}{(1-\gamma)}\left(C - \sum_{(s,a)\in\mathcal{D}} x^0(s,a)\right)\right\} > 0$. Notice that $0 \le r_1(s,a) \le 1$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, implying $\sum_{(s,a)\in\mathcal{D}} h(r_1)(s,a) \le |\mathcal{D}|$. It follows that $x' \in \mathcal{X}$ because

$$\begin{aligned}
C - \sum_{(s,a)\in\mathcal{D}} x'(s,a) &\ge C - \sum_{(s,a)\in\mathcal{D}} x^0(s,a) - |\mathcal{D}|\cdot\epsilon\cdot\frac{(1-\gamma)}{2|\mathcal{D}|} \\
&\ge C - \sum_{(s,a)\in\mathcal{D}} x^0(s,a) - \frac{1}{(1-\gamma)}\left(C - \sum_{(s,a)\in\mathcal{D}} x^0(s,a)\right)\cdot\frac{(1-\gamma)}{2} \\
&= \frac{1}{2}\left(C - \sum_{(s,a)\in\mathcal{D}} x^0(s,a)\right) \\
&> 0.
\end{aligned}$$

For any $m \in \mathcal{M}$, it holds that

$$\begin{aligned}
\langle r_2^{x'}, m\rangle &= \langle r_2^{x^0+\epsilon\cdot\frac{(1-\gamma)}{2|\mathcal{D}|}\cdot h(r_1)}, m\rangle \\
&= \langle r_2^{x^0}, m\rangle + \epsilon\cdot\frac{(1-\gamma)}{2|\mathcal{D}|}\cdot\langle r_1, m\rangle.
\end{aligned} \tag{37}$$

Furthermore, when $m \in \mathcal{M}_{\text{det}} \setminus \text{BR}(x^0)$, it follows from (37) that

$$\begin{aligned}
\langle r_2^{x'}, m\rangle &= \langle r_2^{x^0}, m\rangle + \epsilon\cdot\frac{(1-\gamma)}{2|\mathcal{D}|}\cdot\langle r_1, m\rangle \\
&\le \langle r_2^{x^0}, m\rangle + \epsilon\cdot\frac{(1-\gamma)}{2|\mathcal{D}|}\cdot\frac{1}{1-\gamma} \\
&\le \langle r_2^{x^0}, m\rangle + \epsilon\cdot\frac{1}{2|\mathcal{D}|} \\
&\le \langle r_2^{x^0}, m\rangle + \frac{\epsilon}{2} \\
&\le \langle r_2^{x^0}, m\rangle + \frac{b}{2} \\
&\le \langle r_2^{x^0}, m^0\rangle - \frac{b}{2} \\
&\le \langle r_2^{x'}, m^0\rangle - \frac{b}{2},
\end{aligned} \tag{38}$$

where the first inequality is from Lemma 22. In the meantime, consider any $\tilde{m} \in \left(\mathcal{M}_{\text{det}} \cap \text{BR}(x^0)\right) \setminus \mathcal{M}^\star(x^0)$. It follows that $\langle r_1, \tilde{m}\rangle \ne \langle r_1, m^0\rangle$ and $\langle r_2^{x^0}, \tilde{m}\rangle = \langle r_2^{x^0}, m^0\rangle$ from the definition of $\mathcal{M}^\star(x^0)$. Since (x^0, m^0) is an

optimal solution of the reward design problem in (5), it holds that $\langle r_1, m^0\rangle > \langle r_1, \tilde{m}\rangle$. Hence, for any $\tilde{m} \in \big(\mathcal{M}_{\text{det}} \cap \text{BR}(x^0)\big) \setminus \mathcal{M}^\star(x^0)$, it follows from (37) that

$$\begin{aligned}\langle r_2^{x'}, \tilde{m}\rangle &= \langle r_2^{x^0}, \tilde{m}\rangle + \epsilon \cdot \frac{(1-\gamma)}{2|\mathcal{D}|} \cdot \langle r_1, \tilde{m}\rangle \\ &< \langle r_2^{x^0}, \tilde{m}\rangle + \epsilon \cdot \frac{(1-\gamma)}{2|\mathcal{D}|} \cdot \langle r_1, m^0\rangle \\ &= \langle r_2^{x^0}, m^0\rangle + \epsilon \cdot \frac{(1-\gamma)}{2|\mathcal{D}|} \cdot \langle r_1, m^0\rangle \\ &= \langle r_2^{x'}, m^0\rangle.\end{aligned} \tag{39}$$

Let $c = \langle r^{x'}, m^0\rangle - \max_{m' \in (\mathcal{M}_{\text{det}} \cap \text{BR}(x^0)) \setminus \mathcal{M}^\star(x^0)} \langle r^{x'}, m'\rangle$. It follows from (39) that $c > 0$. Let $d = \min\left\{\frac{1}{2}b, c\right\} > 0$. Recall that $\mathcal{M}_{\text{det}} = \big(\mathcal{M}_{\text{det}} \setminus \text{BR}(x^0)\big) \cup \big(\big(\mathcal{M}_{\text{det}} \cap \text{BR}(x^0)\big) \setminus \mathcal{M}^\star(x^0)\big) \cup \mathcal{M}^\star(x^0)$. For any $m' \in \mathcal{M}_{\text{det}} \setminus \mathcal{M}^\star(x^0) = \big(\mathcal{M}_{\text{det}} \setminus \text{BR}(x^0)\big) \cup \big(\big(\mathcal{M}_{\text{det}} \cap \text{BR}(x^0)\big) \setminus \mathcal{M}^\star(x^0)\big)$, from (38) and (39), we obtain that

$$\begin{aligned}\langle r_2^{x'}, m'\rangle &\le \langle r_2^{x'}, m^0\rangle - \min\left\{\frac{1}{2}b, c\right\} \\ &= \langle r_2^{x'}, m^0\rangle - d.\end{aligned} \tag{40}$$

We will continue the proof by considering two mutually exclusive cases.
Case 1: Suppose for all $m \in \mathcal{M}^\star(x^0)$, it holds that $h(m) = h(m^0)$ or equivalently

$$m(s,a) = m^0(s,a). \tag{41}$$

for all $(s,a) \in \mathcal{D}$. For any $m \in \mathcal{M}^\star(x^0)$, it follows from (36) that $\langle r_2^{x^0}, m^0\rangle = \langle r_2^{x^0}, m\rangle$ and $\langle r_1, m^0\rangle = \langle r_1, m\rangle$. Use (37) to obtain

$$\begin{aligned}\langle r_2^{x'}, m^0\rangle &= \langle r_2^{x^0}, m^0\rangle + \epsilon \cdot \frac{(1-\gamma)}{2|\mathcal{D}|} \cdot \langle r_1, m^0\rangle \\ &= \langle r_2^{x^0}, m\rangle + \epsilon \cdot \frac{(1-\gamma)}{2|\mathcal{D}|} \cdot \langle r_1, m\rangle \\ &= \langle r_2^{x'}, m\rangle.\end{aligned} \tag{42}$$

Given a function $v\colon \mathcal{D} \to \mathbb{R}$, define δ^v as in (26). In the following, we sometimes abuse the notation and treat v as a vector in $\mathbb{R}^{|\mathcal{D}|}$. Define $x'(v) \triangleq x' + dv$. For any $m \in \mathcal{M}^\star(x^0)$ and $v \in \mathbb{R}^{|\mathcal{D}|}$,

$$\langle r_2^{x'(v)}, m^0\rangle = \langle r_2^{x'+dv}, m^0\rangle = \langle r_2^{x'}, m^0\rangle + d \cdot \langle \delta^v, m^0\rangle = \langle r_2^{x'}, m\rangle + d \cdot \langle \delta^v, m^0\rangle. \tag{43}$$

From (35) and (41),

$$\langle \delta^v, m^0\rangle = \sum_{(s,a)\in\mathcal{D}} v(s,a) m^0(s,a) = \sum_{(s,a)\in\mathcal{D}} v(s,a) m(s,a) = \langle \delta^v, m\rangle. \tag{44}$$

Substitute (44) into (43) to obtain

$$\langle r_2^{x'(v)}, m^0\rangle = \langle r_2^{x'}, m\rangle + d \cdot \langle \delta^v, m\rangle = \langle r_2^{x'(v)}, m\rangle. \tag{45}$$

Furthermore, when $\|v\|_1 \leq \frac{1}{2}(1-\gamma)$, for any $m' \in \mathcal{M}_{\text{det}} \setminus \mathcal{M}^\star(x^0)$, it holds that

$$\begin{aligned}
\langle r_2^{x'(v)}, m' \rangle &= \langle r_2^{x'+dv}, m' \rangle \\
&= \langle r_2^{x'}, m' \rangle + d \cdot \langle \delta^v, m' \rangle \\
&\leq \langle r_2^{x'}, m' \rangle + d \cdot \frac{1}{2}(1-\gamma) \cdot \max_{(s,a)\in\mathcal{D}} m'(s,a) \\
&\leq \langle r_2^{x'}, m' \rangle + \frac{1}{2} d \\
&\leq \langle r_2^{x'}, m^0 \rangle - \frac{1}{2} d \\
&\leq \langle r_2^{x'}, m^0 \rangle - d \cdot \frac{1}{2}(1-\gamma) \cdot \max_{(s,a)\in\mathcal{D}} m^0(s,a) \\
&\leq \langle r_2^{x'}, m^0 \rangle + d \cdot \langle \delta^v, m \rangle \\
&= \langle r_2^{x'+dv}, m^0 \rangle \\
&= \langle r_2^{x'(v)}, m^0 \rangle.
\end{aligned} \tag{46}$$

The first inequality is from Lemma 28. From (45) and (46), it follows that $\langle r_2^{x'(v)}, m^0 \rangle \geq \langle r_2^{x'(v)}, m \rangle$ for any $m \in \mathcal{M}_{\text{det}}$ and $\|v\|_1 < (1-\gamma)/2$. This implies that the open ball

$$\{x' + dv \mid \|v\|_1 < \frac{1}{2}(1-\gamma)\}$$

is contained in $\mathcal{P}_{m^0}$. Since $x' \in \mathcal{X}$, and m^0 is an optimal occupancy measure, x' must be an optimal interior-point allocation.

Case 2: Consider the case when condition (41) in Case 1 does not hold. Recall the lexicographic order on real-valued vectors: Let $x, y \in \mathbb{R}^{|\mathcal{D}|}$. The vector x is said to be lexicographically greater than y, denoted by $x \succ_{\text{lex}} y$, if there exists $i \in \{1, \ldots, |\mathcal{D}|\}$ such that

$$x_j = y_j \quad \text{for all } j < i \quad \text{and} \quad x_i > y_i.$$

Arrange the elements in $\mathcal{M}^\star(x^0)$ in descending lexicographic order according to their restrictions on $\mathcal{D}$: $h(m^1) = \cdots = h(m^K) \succ_{\text{lex}} h(m^{K+1}) \ldots$, where $m^1, \cdots \in \mathcal{M}^\star(x^0)$. Then there must exist $w \in \mathbb{R}^{|\mathcal{D}|}$ such that $\langle w, h(m^1) \rangle = \cdots = \langle w, h(m^K) \rangle > \langle w, h(m^{K+1}) \rangle \ldots$ or equivalently $\langle \delta^w, m^1 \rangle = \cdots = \langle \delta^w, m^K \rangle > \langle \delta^w, m^{K+1} \rangle \ldots$. For sufficiently small $\epsilon' > 0$, it can be shown that $x'' = x' + \epsilon' \delta^w$ is an optimal interior-point allocation as follows. Let $k \in \{1, \ldots, K\}$. For any $k' > K$, it follows from the fact $m^k, m^{k'} \in \mathcal{M}^\star(x^0)$ and (42) that

$$\begin{aligned}
\langle r_2^{x''}, m^k \rangle &= \langle r_2^{x'}, m^k \rangle + \epsilon' \cdot \langle \delta^w, m^k \rangle \\
&= \langle r_2^{x'}, m^{k'} \rangle + \epsilon' \cdot \langle \delta^w, m^k \rangle \\
&> \langle r_2^{x'}, m^{k'} \rangle + \epsilon' \cdot \langle \delta^w, m^{k'} \rangle \\
&= \langle r_2^{x''}, m^{k'} \rangle.
\end{aligned} \tag{47}$$

Also, for any $m' \in \mathcal{M}_{\text{det}} \setminus \mathcal{M}^\star(x^0)$, it follows from (40) that

$$\begin{aligned}\langle r_2^{x''}, m^k\rangle &= \langle r_2^{x'}, m^k\rangle + \epsilon' \cdot \langle \delta^w, m^k\rangle \\ &= \langle r_2^{x'}, m^0\rangle + \epsilon' \cdot \langle \delta^w, m^k\rangle \\ &\geq \langle r_2^{x'}, m'\rangle + d + \epsilon' \cdot \langle \delta^w, m^k\rangle \\ &> \langle r_2^{x'}, m'\rangle\end{aligned} \tag{48}$$

when ϵ' is sufficiently small. Conditions (47) and (48) imply $\mathcal{M}^\star(x'') = \{m^1, \ldots, m^K\}$ when ϵ' is sufficiently small. Then, using the proof idea from Case 1, it can be shown that x'' is an optimal interior-point allocation.

(Necessity) Assume toward a contradiction that $x^\star$ is an optimal interior-point allocation in $\mathcal{P}_m$ for some optimal occupancy measure m. Then there must exist an open set $B(x^\star)$ containing $x^\star$ with $B(x^\star) \subseteq \mathcal{P}_m$. Since $\mathcal{X}$ is convex, the set $B(x^\star) \cap \mathcal{X}$ has a nonempty interior by Lemma 27. On the other hand, from the given condition, any $x \in \mathcal{P}_m$ must satisfy $C - \sum_{i=1}^{|\mathcal{D}|} x_i = 0$ and is hence on the boundary of $\mathcal{X}$. This implies that the interior of $\mathcal{P}_m \cap \mathcal{X}$ is empty. Since $B(x^\star) \subseteq \mathcal{P}_m$, it follows that the interior of $B(x^\star) \cap \mathcal{X}$ is also empty, leading to a contradiction. □

E.2 Proof of Theorem 17

Proof. If $\sum_{i=1}^{|\mathcal{D}|} x_i^\star < C$, then there exists an optimal interior-point allocation according to Theorem 16.

If $\sum_{i=1}^{|\mathcal{D}|} x_i^\star = C$, let $m^\star \in \text{BR}(x^\star)$, and define

$$\text{BR}_{\text{det}}^\star \triangleq \mathcal{M}_{\text{det}} \cap \text{BR}(x^\star) = \left\{m \in \mathcal{M}_{\text{det}} \mid \langle r_2^{x^\star}, m\rangle = \langle r_2^{x^\star}, m^\star\rangle\right\}$$

and $b \triangleq \langle r_2^{x^\star}, m^\star\rangle - \max_{m \in \mathcal{M}_{\text{det}} \setminus \text{BR}_{\text{det}}^\star} \langle r_2^{x^\star}, m\rangle > 0$. Because any best response $m \in \text{BR}(x^\star)$ satisfies $\langle r_1, m\rangle = v_1^\star$, it holds that $\langle r_1, m\rangle = \langle r_1, m^\star\rangle$ for all $m \in \text{BR}_{\text{det}}^\star$. Since $\sum_{i=1}^{|\mathcal{D}|} x_i^\star = C > 0$, there exists $j \in \{1, \ldots, |\mathcal{D}|\}$ such that $x_j^\star > 0$. Choose c such that $0 < c < \min\{x_j^\star, b(1-\gamma)\}$, and define $x' \in \mathbb{R}^{|\mathcal{D}|}$ such that $x'_j = x_j^\star - c$ and $x'_i = x_i^\star$ for $i \neq j$. Choose any $m^0 \in \arg\max_{m \in \text{BR}_{\text{det}}^\star} \langle r_2^{x'}, m\rangle$. It can be shown that (x', m^0) is a feasible solution of (16) as follows. For any $m \in \mathcal{M}_{\text{det}} \setminus \text{BR}_{\text{det}}^\star$,

$$\begin{aligned}\langle r_2^{x'}, m^0\rangle - \langle r_2^{x'}, m\rangle &= \langle r_2^{x^\star}, m^0\rangle - cm^0(s^{(j)}, a^{(j)}) - \langle r_2^{x^\star}, m\rangle + cm(s^{(j)}, a^{(j)}) \\ &\geq \langle r_2^{x^\star}, m^0\rangle - cm^0(s^{(j)}, a^{(j)}) - \langle r_2^{x^\star}, m\rangle \\ &\geq b - cm^0(s^{(j)}, a^{(j)}) \\ &\geq b - c \cdot \frac{1}{1-\gamma} \\ &> 0,\end{aligned}$$

where the third inequality follows from Lemma 22. From the choice of m^0, it follows that $\langle r_2^{x'}, m^0\rangle \geq \langle r_2^{x'}, m\rangle$ for all $m \in \mathcal{M}_{\text{det}}$ or equivalently $m^0 \in \text{BR}(x')$ since at least one deterministic occupancy measure must be optimal under $r_2^{x'}$. Moreover, $m^0 \in \text{BR}_{\text{det}}^\star$ implies that $\langle r_1, m^0\rangle = v_1^\star$, and $c < x_j^\star$ implies that $x' \in \mathcal{X}$. Since $C - \sum_{i=1}^{|\mathcal{D}|} x'_i = C - \sum_{i=1}^{|\mathcal{D}|} x_i^\star + c = c > 0$, the optimal value of (16) must be strictly positive. It then follows from Theorem 16 that there exists an optimal interior-point allocation. □

Received 15 May 2025; accepted 23 August 2025