Journal of Artificial Intelligence Research 72 (2021) 507-531 Submitted 11/2020; published 10/2021

Optimally Deceiving a Learning Leader in Stackelberg Games

Georgios Birmpas GEBIRBAS@GMAIL.COM
Sapienza University of Rome
Rome, Italy

Jiarui Gan JRGAN@MPI-SWS.ORG
Max Planck Institute for Software Systems
Kaiserslautern, Germany

Alexandros Hollender ALEXANDROS.HOLLENDER@CS.OX.AC.UK
University of Oxford
Oxford, United Kingdom

Francisco J. Marmolejo-Cossío FJMARMOL@SEAS.HARVARD.EDU
Harvard University
Cambridge, MA, USA

Ninad Rajgopal NINAD.RAJGOPAL@WARWICK.AC.UK
University of Warwick
Coventry, United Kingdom

Alexandros A. Voudouris ALEXANDROS.VOUDOURIS@ESSEX.AC.UK
University of Essex
Colchester, United Kingdom

Abstract

Recent results have shown that algorithms for learning the optimal commitment in a Stackelberg game are susceptible to manipulation by the follower. These learning algorithms operate by querying the best responses of the follower, who consequently can deceive the algorithm by using fake best responses, typically by responding according to fake payoffs that are different from the actual ones. For this strategic behavior to be successful, the main challenge faced by the follower is to pinpoint the fake payoffs that would make the learning algorithm output a commitment that benefits them the most. While this problem has been considered before, the related literature has only focused on a simple setting where the follower can only choose from a finite set of payoff matrices, thus leaving the general version of the problem unanswered. In this paper, we fill this gap by showing that it is always possible for the follower to *efficiently* compute (near-)optimal fake payoffs, for various scenarios of learning interaction between the leader and the follower. Our results also establish an interesting connection between the follower's deception and the leader's maximin utility: through deception, the follower can induce almost any (fake) Stackelberg equilibrium if and only if the leader obtains at least their maximin utility in this equilibrium.

1. Introduction

Stackelberg games are a simple yet powerful model for sequential interaction among strategic agents. In such games there are two players: a leader and a follower. The leader commits

to a strategy, and the follower acts upon observing the leader's commitment. The simple sequential structure of the game permits modeling a multitude of important scenarios. Indicative applications include the competition between a large and a small firm (von Stackelberg, 2010), the allocation of defensive resources, i.e., Stackelberg security games (Tambe, 2011; Korzhyk et al., 2011; Fang et al., 2013; Delle Fave et al., 2014; Gan et al., 2015), the competition among mining pools in the Bitcoin network (Marmolejo-Cossío et al., 2019; Sun et al., 2020), and the protection against manipulation in elections (Elkind et al., 2021; Yin et al., 2018).

In Stackelberg games, the leader is interested in finding the best commitment she can make, assuming that the follower behaves rationally. The combination of such a commitment by the leader and the follower's rational best response to it leads to a *strong Stackelberg equilibrium* (SSE). In general, the utility that the leader obtains in an SSE is larger than what she would obtain in a Nash equilibrium of the corresponding simultaneous-move game (von Stengel & Zamir, 2004). Hence, the leader prefers to commit than to engage in a simultaneous game with the follower.

In case the leader has access to both her and the follower's payoff parameters (or, payoffs, for short), computing an SSE is a computationally tractable problem (Conitzer & Sandholm, 2006). In practice however, the leader may have limited or no information about the follower's payoffs. In order to determine the optimal commitment, the leader must endeavor to elicit information about the incentives of the follower through indirect means. This avenue of research has led to a plethora of active-learning-based approaches for the computation of SSEs (Balcan et al., 2015; Blum et al., 2014; Letchford et al., 2009; Peng et al., 2019; Roth et al., 2016). Meanwhile, inspired by recent developments in the ML community regarding adversarial examples in classification algorithms (Lowd & Meek, 2005; Barreno et al., 2010), there has been a stream of recent papers exploring the notion of adversarial deception in Stackelberg games, whereby the follower uses fake samples to tamper with the leader's information acquisition.

Specifically, when the leader learns an SSE by querying the follower's best responses, the follower can use fake best responses to change the SSE learned by the algorithm. As recently explored by Gan et al. (2019b), one particular approach the follower can employ, termed *imitative follower deception*, is to imitate best responses implied by payoffs that are different from the actual ones. The key to the success of such a deceptive behavior is thus to pinpoint the fake payoffs that could make the leader learn an SSE in which the actual utility of the follower is maximized. In the scenario studied by Gan et al. (2019b), this task is trivial for the follower as his choices are limited to a finite set of explicitly given payoff matrices, whose size is bounded by the size of the problem representation; thus, to efficiently find out the optimal payoffs, the follower can simply enumerate all possible payoff matrices and see which one of them leads to the best outcome.

In this paper, we consider the general version of the optimal imitative deception problem, where the follower is allowed to imitate *any* payoff matrix, without restrictions on the space of possible values. This general version has been considered in two very recent papers (Gan et al., 2019a; Nguyen & Xu, 2019), but only in the specific context of Stackelberg security games. Besides that, no progress has been made for general Stackelberg games. In this paper, we aim to fill this gap, by completely resolving this computational problem.

1.1 Our Contribution

We explore how a follower can optimally deceive a learning leader in Stackelberg games by misreporting his payoff matrix, and we study the tractability of the corresponding optimization problem. As mentioned above, our objective is to compute an optimal fake payoff matrix, that leads to the best SSE for the follower. However, unlike the related literature, we do not impose any restrictions on the space from which the payoffs are selected or on the type of the game. By exploiting an intuitive characterization of all strategy profiles that can be induced as SSEs in Stackelberg games, we show that it is always possible for the follower to compute an optimal payoff matrix in polynomial time, irrespective of the specific learning algorithm employed by the leader. Furthermore, we strengthen this result to resolve possible equilibrium selection issues, by showing that the follower can construct a payoff matrix that induces a *unique* SSE, in which his utility is maximized up to some arbitrarily small loss.

Our characterization of inducible strategy profiles establishes an interesting connection between the follower's deception and the leader's *maximin* utility: through deception, the follower can induce almost any (fake) SSE if and only if the leader obtains at least her maximin utility in this equilibrium. Given that the maximin utility is always attainable without any additional information about the opponent, this means that the follower can exploit the leader's lack of information to the maximum degree, despite the leader's attempts to learn additional information. This connection to the maximin utility also reflects the findings of Gan et al. (2019a) on Stackelberg security games, who showed that the optimal deception in such games is to use fake payoffs that make the game zero-sum, whereby the leader obtains exactly the maximin utility. In the generic Stackelberg games we study in this paper, the optimal deception may, but need not always, lead to zero-sum games; hence, we fully characterize the space of inducible strategy profiles, which requires completely different techniques.

1.2 Other Related Work

Our paper is related to an emerging line of work at the intersection of machine learning and algorithmic game theory, dealing with scenarios where the samples used for training learning algorithms are controlled by strategic agents, who aim to optimize their personal benefit. Indicatively, there has been recent interest in the analysis of the effect of strategic behavior on the efficiency of existing algorithms, as well as the design of algorithms resilient to strategic manipulation for linear regression (Ben-Porat & Tennenholtz, 2019; Chen et al., 2018; Dekel et al., 2010; Hossain & Shah, 2020; Perote & Perote-Peña, 2004; Waggoner et al., 2015) and classification (Chen et al., 2019; Dong et al., 2018; Meir et al., 2012; Zhang et al., 2019).

Beyond the strategic considerations above, our work is also related to the study of query protocols for learning game-theoretic equilibria. In this setting, as in ours, algorithms for computing equilibria via utility and best response queries are a natural starting point. For utility queries, there has been much work in proving exponential lower bounds for randomized computation of exact, approximate and well-supported Nash equilibria (Babichenko & Rubinstein, 2017; Babichenko, 2016; Chen et al., 2017; Goldberg & Roth, 2016; Hart & Mansour, 2010; Hart & Nisan, 2018), as well as providing query-efficient protocols for

approximate Nash equilibrium computation in bimatrix games, congestion games (Fearnley et al., 2015), anonymous games (Goldberg & Turchetta, 2017), and large games (Goldberg et al., 2019). Best response queries are weaker than utility queries, but they arise naturally in practice, and are also expressive enough to implement fictitious play, a dynamic first proposed by Brown (1949), and proven to converge by Robinson (1951) for two-player zero-sum games to an approximate Nash equilibrium. In terms of equilibrium computation, Goldberg and Marmolejo-Cossío (2018) also provide query-efficient algorithms for computing approximate Nash equilibria for bimatrix games via best response queries provided one agent has a constant number of strategies.

Finally, learning via incentive queries in games is directly related to the theory of preference elicitation, where the goal is to mine information about the private parameters of the agents by interacting with them (Blum et al., 2004; Lahaie & Parkes, 2004; Zinkevich et al., 2003; Goldberg et al., 2020). This has many applications, most notably combinatorial auctions, where access to the valuation functions of the agents is achieved via value or demand queries (Blumrosen & Nisan, 2007; Conen & Sandholm, 2001; Nisan & Segal, 2006).

2. Preliminaries

A Stackelberg game is a sequential game between a *leader* and a *follower*.[1] The leader commits to a strategy, and the follower then acts upon observing this commitment. We consider finite games, in which the leader and the follower have m and n *pure strategies* at their disposal, respectively, and their utilities for all possible outcomes are given by the matrices $u^L, u^F \in \mathbb{R}^{m\times n}$. The entries $u^L(i,j)$ and $u^F(i,j)$ denote the utilities of the leader and the follower under *pure strategy profile* $(i,j) \in [m] \times [n]$, where we use the notation $[k] = \{1, \dots, k\}$ for any positive integer k. We use $\mathcal{G} = (u^L, u^F)$ to denote the Stackelberg game with payoff matrices u^L and u^F; we omit m and n in the tuple as they are clear from context. The games we consider are general-sum games, with no restriction on the matrices $u^L, u^F \in \mathbb{R}^{m\times n}$.

The players are allowed to employ mixed strategies, whereby they randomize over actions in their strategy set. A mixed strategy of the leader is a probability distribution over $[m]$, denoted by $\mathbf{x} \in \Delta^{m-1} = \{\mathbf{x} \geq 0 : \sum_{i\in[m]} x_i = 1\}$. By slightly abusing notation, we let $u^L(\mathbf{x}, j) = \sum_{i\in[m]} x_i \cdot u^L(i,j)$ be the *expected utility* of the leader when she plays the mixed strategy $\mathbf{x}$ and the follower plays a pure strategy j. Similarly, we define $u^F(\mathbf{x}, j) = \sum_{i\in[m]} x_i \cdot u^F(i,j)$ for the follower. For a given mixed strategy $\mathbf{x} \in \Delta^{m-1}$ of the leader, we say that a pure strategy $j \in [n]$ is a *follower best response* if $u^F(\mathbf{x}, j) = \max_{\ell\in[n]} u^F(\mathbf{x}, \ell)$; we denote the set of all follower best responses to $\mathbf{x}$ by $BR(\mathbf{x}) \subseteq [n]$ and refer to the function BR as the *best response correspondence* of the follower, or the *BR correspondence*. We can further generalize the follower's response to be a mixed strategy $\mathbf{y} \in \Delta^{n-1}$, but this is unnecessary as will become clear below.

An SSE is the standard solution concept in Stackelberg games. It captures the situation where the leader commits to a mixed strategy that maximizes her expected utility, while taking into account the follower's best response to her commitment. It is assumed that

1. By convention, we will refer to the leader as a female and the follower as a male.

the follower breaks ties in favor of the leader when he has multiple best responses, and without loss of generality the best response chosen is always a pure strategy. This optimistic assumption is justified by the fact that such tie-breaking behavior can often be enforced by an infinitesimal perturbation in the leader's strategy (von Stengel & Zamir, 2004). Hence, in the definition below, we only consider pure strategies of the follower.[2]

Definition 2.1 (**Strong Stackelberg Equilibrium (SSE)**). A strategy profile $(\mathbf{x}, j)$ is an SSE of the game $\mathcal{G} = (u^L, u^F)$ if

$$(\mathbf{x}, j) \in \underset{\mathbf{y} \in \Delta^{m-1},\, \ell \in BR(\mathbf{y})}{\arg\max} u^L(\mathbf{y}, \ell).$$

Note that the set of SSEs of the game $\mathcal{G} = (u^L, u^F)$ actually only depends on (u^L, BR). Thus, with a slight abuse of notation, we will sometimes also denote the game as $\mathcal{G} = (u^L, BR)$, where the preferences of the follower are directly given by a best response correspondence BR (which may or may not correspond to a payoff matrix $u^F \in \mathbb{R}^{m \times n}$).

2.1 Learning SSEs and Deceptive Follower Behavior

We consider the scenario where the leader has full knowledge of her utility matrix u^L, and aims to compute an SSE by interacting with the follower and gleaning information about u^F. For example, the leader could observe the follower's best responses in play (akin to having query access to BR), or observe the follower's payoffs at pure strategy profiles during play (akin to having query access to u^F as a function). Hence, this can be cast as the problem of learning an SSE with a specified notion of query access to information about the follower's incentives.

Consider a game $\mathcal{G} = (u^L, u^F)$. If the follower controls the flow of information to the leader in this paradigm, he may consider perpetually interacting with the leader as if he had a different payoff matrix $\tilde{u}^F$ (or a different BR correspondence $\widetilde{BR}$), which can make the leader believe that they are playing the game $\widetilde{\mathcal{G}} = (u^L, \tilde{u}^F)$ (or $\widetilde{\mathcal{G}} = (u^L, \widetilde{BR})$). This deceiving power provides the follower with an incentive to act according to $\widetilde{\mathcal{G}}$ for a judicious choice of $\tilde{u}^F$ (or $\widetilde{BR}$), because the SSEs in $\widetilde{\mathcal{G}}$ may yield a larger utility (according to u^F) than the SSEs in $\mathcal{G}$. More concretely, the example below shows that the follower can gain an arbitrarily large benefit by deceiving the leader into playing a different game. The example also shows that the leader's utility loss can be arbitrarily bad.

Example 2.2 (**Beneficial deception**). Let $\alpha \in [0, 1]$ and consider the following matrices:

$$R = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad C_\alpha = \begin{pmatrix} 0 & \alpha \\ 1 & \alpha \end{pmatrix}$$

Now, suppose that $u^L = R$ and $u^F = C_\alpha$, and let $x \in [0, 1]$ represent the probability mass that the leader (row player) places on the first row (her first strategy); thus, $1 - x$ is the probability with which she plays her second strategy. Given this mixed strategy of the leader, the utilities that the follower expects to derive from her two strategies (columns)

2. When mixed strategies of the follower are considered, a strategy profile $(\mathbf{x}, \mathbf{y})$ is an SSE if $(\mathbf{x}, j)$ is an SSE (as in Definition 2.1) for all pure strategies j in the support set of $\mathbf{y}$.

are $u^F(x,1) = 1-x$ and $u^F(x,2) = \alpha$. Consequently, the first strategy is a best response of the follower when $x \in [0, 1-\alpha]$, and the second one is a best response when $x \in (1-\alpha, 1]$ (when $x = 1-\alpha$, the tie is broken in favor of the leader). With this information, it is clear that an SSE of the game occurs when the leader chooses $x = 1-\alpha$ and the follower plays his first strategy; in fact, this is the unique SSE when $\alpha < 1$. As a result, the follower's utility is $u^F(1-\alpha, 1) = \alpha$.

However, for any $\alpha < 1$, the follower has an incentive to deceive the leader into playing a game $\widetilde{\mathcal{G}} = (R, C_\beta)$ with $\beta \in (\alpha, 1)$, which will improve his *true* utility (computed according to the true payoff matrix C_α) in the resulting SSE to $u^F(1-\beta, 1) = \beta$. The follower can achieve this by responding to the learning algorithm's queries optimally according to C_β (or by misreporting C_β as his payoff matrix in scenarios where this information can be conveyed directly). This is an improvement by a multiplicative factor of β/α, which can be arbitrarily large when α approaches 0. Meanwhile, the utility of the leader drops from $1-\alpha$ to $1-\beta$, which amounts to a multiplicative factor of $(1-\alpha)/(1-\beta)$ and can be arbitrarily large when β approaches 1. □

2.2 Inducible Strategy Profiles

The ultimate goal of the follower is to identify the SSE that maximizes his true utility, from the set of SSEs that he can deceive the leader into learning. We will refer to such SSEs as *inducible strategy profiles*. At a high level, the follower's problem can now be expressed as the following optimization problem:

$$\begin{aligned} \max_{\mathbf{x}\in\Delta^{m-1},\ j\in[n]} \quad & u^F(\mathbf{x}, j), \\ \text{subject to} \quad & (\mathbf{x}, j) \text{ is inducible} \end{aligned} \tag{1}$$

This maximum utility for the follower is called the *optimal inducible utility*. If the maximum value is never achieved, then for every $\varepsilon > 0$, we would like to be able to find an inducible SSE that achieves a value ε-close to the supremum value. We assume that the follower has full information about the leader's payoff matrix u^L throughout.

As discussed previously, the leader can learn an SSE by querying the best responses of the follower to particular leader strategies, or by querying more refined information about the follower's payoff matrix. Depending on the type of information queried, we can define various notions of inducible strategy profiles.

In more detail, suppose the leader can only query the best responses of the follower, who behaves according to some best response correspondence $\widetilde{BR} : \Delta^{m-1} \to 2^{[n]} \setminus \{\varnothing\}$. This interaction between the leader and the follower leads to a game denoted as $\widetilde{\mathcal{G}} = (u^L, \widetilde{BR})$, where the only information known is $\widetilde{BR}$ (instead of a payoff matrix implying $\widetilde{BR}$). The definition of $\widetilde{BR}$ enforces a best response answer to any possible query. Consequently, the leader learns an SSE $(\mathbf{x}, j) \in \arg\max_{\mathbf{y}\in\Delta^{m-1},\, \ell\in\widetilde{BR}(\mathbf{y})} u^L(\mathbf{y}, \ell)$, which yields the following notion of *BR-inducible* strategy profiles.

Definition 2.3 (**BR inducibility**)**.** A strategy profile $(\mathbf{x}, j)$ is *BR-inducible* with respect to u^L if there exists a best response correspondence $\widetilde{BR} : \Delta^{m-1} \to 2^{[n]} \setminus \{\varnothing\}$ such that $(\mathbf{x}, j)$ is an SSE of the game $\widetilde{\mathcal{G}} = (u^L, \widetilde{BR})$, in which case we say that $(\mathbf{x}, j)$ is *induced* by $\widetilde{BR}$.

Next, consider the case where the leader can query information about the payoffs of the follower, who can now behave according to a fake payoff matrix $\tilde{u}^F$. We refer to the SSEs of the resulting game $\widetilde{\mathcal{G}} = (u^L, \tilde{u}^F)$ as *payoff-inducible* strategy profiles.

Definition 2.4 (**Payoff inducibility**)**.** A strategy profile $(\mathbf{x}, j)$ is said to be *payoff-inducible* with respect to u^L if there exists $\tilde{u}^F \in \mathbb{R}^{m \times n}$ such that $(\mathbf{x}, j)$ is an SSE in the game $\widetilde{\mathcal{G}} = (u^L, \tilde{u}^F)$, in which case we say that $(\mathbf{x}, j)$ is *induced* by $\tilde{u}^F$.

Clearly, payoff inducibility is stricter than BR inducibility: for every choice of $\tilde{u}^F$, the corresponding best response correspondence $\widetilde{BR}(\mathbf{y}) = \arg\max_{\ell \in [n]} \tilde{u}^F(\mathbf{y}, \ell)$ induces the same SSEs as $\tilde{u}^F$ does.

Note that the above definitions only require an inducible strategy profile to be a verifiable SSE, with respect to the information about the follower's incentives (either $\widetilde{BR}$ or $\tilde{u}^F$). It may happen that the resulting game $\widetilde{\mathcal{G}}$ has multiple SSEs (e.g., the game in Example 2.2 when $\alpha = 1$), which gives rise to an equilibrium selection issue. Indeed, in practice, it is not realistic to assume that the follower has any control over which SSE is chosen by the leader (who moves first in the game), especially when there are SSEs involving different leader strategies. To address this, and thus completely resolve the optimal deception problem for the follower, we introduce an even stricter notion of inducibility on top of payoff inducibility, which requires $\widetilde{\mathcal{G}}$ to have a unique SSE.

Definition 2.5 (**Strong inducibility**)**.** A strategy profile $(\mathbf{x}, j)$ is said to be *strongly inducible* with respect to u^L, if there exists a matrix $\tilde{u}^F \in \mathbb{R}^{m \times n}$ such that $(\mathbf{x}, j)$ is the *unique* SSE of the game $\widetilde{\mathcal{G}} = (u^L, \tilde{u}^F)$, in which case we say that $(\mathbf{x}, j)$ is *strongly* induced by $\tilde{u}^F$.

In the next sections, we will investigate solutions to (1) under the inducibility notions above, from the weakest to the strongest. Our general approach is to decompose (1) into n sub-problems by enumerating all possible follower responses $j \in [n]$. For each response j, we solve the corresponding optimization problem presented below, and pick the response that yields the maximum utility for the follower.

$$\begin{aligned} \max_{\mathbf{x} \in \Delta^{m-1}} \quad & u^F(\mathbf{x}, j), \\ \text{subject to} \quad & (\mathbf{x}, j) \text{ is inducible} \end{aligned} \tag{2}$$

Hence, our problem reduces to solving (2). As a final remark, one may wonder whether relaxing the follower's response in (1) (and in the SSE definition) to a mixed strategy $\mathbf{y} \in \Delta^{n-1}$ could improve the optimal inducible utility. This is not the case: If a strategy profile $(\mathbf{x}, \mathbf{y})$ is an SSE in $\widetilde{\mathcal{G}}$, then $(\mathbf{x}, j)$ should also be an SSE for each pure strategy j in the support set of $\mathbf{y}$, while $u^F(\mathbf{x}, j) \geq u^F(\mathbf{x}, \mathbf{y}) := \sum_{\ell \in [n]} y_\ell \cdot u^F(\mathbf{x}, \ell)$ should hold for at least one j in the support set of $\mathbf{y}$.

3. Best Response Inducibility

Let us start our analysis by considering the case in which the leader queries the best responses of the follower. The aim of the follower is to deceive the leader towards a strategy

profile that is BR-inducible as defined in Definition 2.3. Indeed, if the follower is allowed to use an arbitrary $\widetilde{BR}$ to induce a strategy profile $(\mathbf{x}, j)$, he can simply define $\widetilde{BR}$ as follows:

$$\widetilde{BR}(\mathbf{y}) = \begin{cases} \{j\} & \text{if } \mathbf{y} = \mathbf{x} \\ \arg\min_{\ell \in [n]} u^L(\mathbf{y}, \ell) & \text{if } \mathbf{y} \neq \mathbf{x}. \end{cases}$$

Namely, the follower threatens to choose the worst possible response against any leader strategy $\mathbf{y} \neq \mathbf{x}$, so as to minimize the leader's incentive to commit to these strategies. This $\widetilde{BR}$ will successfully convince the leader that $(\mathbf{x}, j)$ is an SSE of $\widetilde{\mathcal{G}} = (u^L, \widetilde{BR})$ — hence inducing $(\mathbf{x}, j)$ — if the threat is powerful enough, that is, if

$$u^L(\mathbf{x}, j) \geq \min_{\ell \in [n]} u^L(\mathbf{y}, \ell) \quad \text{for all } \mathbf{y} \in \Delta^{m-1},$$

where the left hand side is what the leader obtains by committing to $\mathbf{x}$, and the right hand side is what she obtains by committing to $\mathbf{y} \neq \mathbf{x}$. Equivalently, this means that

$$u^L(\mathbf{x}, j) \geq M := \max_{\mathbf{y} \in \Delta^{m-1}} \min_{\ell \in [n]} u^L(\mathbf{y}, \ell), \tag{3}$$

where M is exactly the leader's maximin utility. Indeed, (3) is necessary for $(\mathbf{x}, j)$ to be BR-inducible: if on the contrary $u^L(\mathbf{x}, j) < M$, then by committing to

$$\mathbf{y}^* \in \arg\max_{\mathbf{y} \in \Delta^{m-1}} \min_{\ell \in [n]} u^L(\mathbf{y}, \ell),$$

the leader can obtain (at least) her maximin utility, which will be strictly larger than $u^L(\mathbf{x}, j)$.

Thus, condition (3) gives a simple criterion for BR inducibility, as well as the following LP (linear program) formulation for (2).

$$\begin{aligned} \max_{\mathbf{x} \in \Delta^{m-1}} \quad & u^F(\mathbf{x}, j) \\ \text{subject to} \quad & u^L(\mathbf{x}, j) \geq M \end{aligned} \tag{4}$$

Although simple, the $\widetilde{BR}$ defined above may be far from being one that arises from a choice of $\tilde{u}^F$. Hence, we will next move one step closer to our next goal of studying payoff inducibility, by imposing a stricter condition on $\widetilde{BR}$, which is necessary (but not sufficient) for a strategy profile induced by $\widetilde{BR}$ to also be payoff-inducible. We will show that, in fact, the extra condition imposed on $\widetilde{BR}$ does not compromise its power, and (4) still applies as a formulation under this stricter notion of BR inducibility. This result will be useful for analyzing payoff inducibility in Section 4.

3.1 Polytopal BR Correspondence

In a similar vein to Goldberg and Marmolejo-Cossío (2018), we require that, for every $\ell \in [n]$, the set of leader strategies to which ℓ is a best response is a closed convex polytope, and the union of all these sets forms a partition of the entire strategy space Δ^{m-1} (for example, see the polytope partition of Δ^2 in Figure 1). Any best response correspondence $\widetilde{BR}$ satisfying this assumption is called *polytopal* as is formally defined below.

Definition 3.1 (**Polytopal BR correspondence** (Goldberg & Marmolejo-Cossío, 2018))**.** A best response correspondence $\widetilde{BR} : \Delta^{m-1} \to 2^{[n]} \setminus \{\varnothing\}$ is *polytopal* if it also satisfies the following:

- $\widetilde{BR}^{-1}(\ell)$ is a closed convex polytope for each $\ell \in [n]$, and
- For each $k \neq \ell$, either $\text{relint}(\widetilde{BR}^{-1}(k)) \cap \text{relint}(\widetilde{BR}^{-1}(\ell)) = \varnothing$ or $\widetilde{BR}^{-1}(k) = \widetilde{BR}^{-1}(\ell)$, where $\text{relint}(H)$ denotes the relative interior of a set H.

Being polytopal is necessary for $\widetilde{BR}$ to arise from some payoff matrix. Indeed, the *true* best response correspondence BR, which arises from u^F, is polytopal: Each $BR^{-1}(\ell)$ is a closed convex polytope defined by the hyperplanes $u^F(\mathbf{y}, \ell) \geq u^F(\mathbf{y}, k)$ for all $k \in [n]$ and the borders of Δ^{m-1}; in addition, $\cup_{\ell=1}^{n} BR^{-1}(\ell) = \Delta^{m-1}$, and for any $\ell \neq k$, the polytopes $BR^{-1}(\ell)$ and $BR^{-1}(k)$ only intersect at their borders unless $u^F(\cdot, \ell) = u^F(\cdot, k)$. Thus, if the follower attempts to deceive the leader via a fake $\widetilde{BR}$, the leader might spot the deception in case $\widetilde{BR}$ is not polytopal.

It turns out that the following correspondence, which we denote as $\widetilde{BR}_{\text{P}}$, is polytopal and, as we will shortly show, it is in fact as powerful as any best response correspondence:

$$\widetilde{BR}_{\text{P}}(\mathbf{y}) = \begin{cases} \{j\} & \text{if } \mathbf{y} \in \Delta^{m-1} \setminus \overline{U_j(\mathbf{x})} \\ \{j\} \cup \arg\min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}, \ell) & \text{if } \mathbf{y} \in \overline{U_j(\mathbf{x})} \setminus U_j(\mathbf{x}) \\ \arg\min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}, \ell) & \text{if } \mathbf{y} \in U_j(\mathbf{x}) \end{cases}$$

where $\overline{U_j(\mathbf{x})}$ is the closure of

$$U_j(\mathbf{x}) := \left\{ \mathbf{y} \in \Delta^{m-1} : u^L(\mathbf{y}, j) > u^L(\mathbf{x}, j) \right\},$$

and $\mathbf{x}$ is the leader's strategy that we want to induce.[3] Intuitively, it is safe for the follower to respond by playing j against any leader strategy $\mathbf{y}$ if $u^L(\mathbf{y}, j) \leq u^L(\mathbf{x}, j)$, in which case the leader does not strictly prefer commitment $\mathbf{y}$ to commitment $\mathbf{x}$. In response to the other strategies, however, the follower needs to respond differently in order to minimize the leader's incentive to commit to these strategies. Therefore, $\widetilde{BR}_{\text{P}}$ will successfully induce $(\mathbf{x}, j)$ if and only if the following holds:

$$u^L(\mathbf{x}, j) \geq \max_{\mathbf{y} \in \overline{U_j(\mathbf{x})}} \min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}, \ell), \tag{5}$$

where we use the convention that $\max \varnothing = -\infty$. It is easy to see that $\widetilde{BR}_{\text{P}}$ is indeed polytopal: $\widetilde{BR}_{\text{P}}^{-1}(j) = \Delta^{m-1} \setminus U_j(\mathbf{x})$ is a closed convex polytope, and the same holds for the sets $\widetilde{BR}_{\text{P}}^{-1}(\ell)$ defined by the hyperplanes $u^L(\mathbf{y}, \ell) \leq u^L(\mathbf{y}, k)$, $k \in [n] \setminus \{j\}$ and the borders of $\overline{U_j(\mathbf{x})}$, which further form a partition of $\overline{U_j(\mathbf{x})}$.

The next lemma shows that (5) is in fact equivalent to (3), meaning that $\widetilde{BR}_{\text{P}}$ is as powerful as any $\widetilde{BR}$: if $(\mathbf{x}, j)$ can be induced by an arbitrary $\widetilde{BR}$ then it can also be induced by $\widetilde{BR}_{\text{P}}$.

3. Note that the use of $\overline{U_j(\mathbf{x})}$, instead of the set $\left\{ \mathbf{y} \in \Delta^{m-1} : u^L(\mathbf{y}, j) \geq u^L(\mathbf{x}, j) \right\}$, is important: when $u^L(\mathbf{y}, j) = u^L(\mathbf{x}, j)$ for all $\mathbf{y} \in \Delta^{m-1}$, these two sets define different behaviors.

$$u^L = \begin{pmatrix} 0 & 1 & 1 \\ 1 & -1/2 & 1/2 \\ 1 & 1/2 & -1/2 \end{pmatrix}$$

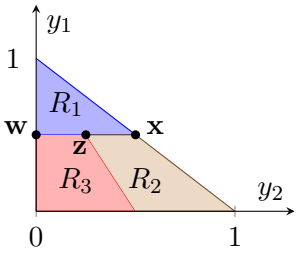

Figure 1: No payoff matrix $\tilde{u}^F$ realizes the polytopal BR correspondence $\widetilde{BR}_{\mathrm{P}}$, such that $\ell \in \widetilde{BR}_{\mathrm{P}}$ if and only if $\mathbf{y} \in R_\ell$, where $R_1 = \{\mathbf{y} \in \Delta^2 : y_1 \geq y_2 + y_3\}$, $R_2 = \{\mathbf{y} \in \Delta^2 : y_1 \leq y_2 + y_3 \text{ and } y_2 \geq y_3\}$, and $R_3 = \{\mathbf{y} \in \Delta^2 : y_1 \leq y_2 + y_3 \text{ and } y_2 \leq y_3\}$.

Lemma 3.2. *$u^L(\mathbf{x}, j) \geq M$ if and only if $u^L(\mathbf{x}, j) \geq \max_{\mathbf{y} \in \overline{U_j(\mathbf{x})}} \min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}, \ell)$.*

Proof. Recall that we want to show that $u^L(\mathbf{x}, j) \geq M$ if and only if

$$u^L(\mathbf{x}, j) \geq \max_{\mathbf{y} \in \overline{U_j(\mathbf{x})}} \min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}, \ell) \tag{6}$$

where M is the maximin utility of the leader. We show that (6) does not hold if and only if $u^L(\mathbf{x}, j) < M$.

Suppose that (6) does not hold. Then $u^L(\mathbf{x}, j) < \max_{\mathbf{y} \in \overline{U_j(\mathbf{x})}} \min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}, \ell)$ by definition, which implies that $U_j(\mathbf{x}) \neq \varnothing$. By the continuity of $\min_{\ell \in [n] \setminus \{j\}} u^L(\cdot, \ell)$, there exists $\mathbf{y}^* \in U_j(\mathbf{x})$ such that

$$u^L(\mathbf{x}, j) < \min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}^*, \ell).$$

By the definition of $U_j(\mathbf{x})$, we also have $u^L(\mathbf{x}, j) < u^L(\mathbf{y}^*, j)$. Thus,

$$u^L(\mathbf{x}, j) < \min_{\ell \in [n]} u^L(\mathbf{y}^*, \ell) \leq \max_{\mathbf{y} \in \Delta^{m-1}} \min_{\ell \in [n]} u^L(\mathbf{y}, \ell) = M.$$

Conversely, suppose that $u^L(\mathbf{x}, j) < M$. Let $\mathbf{y}^* \in \arg\max_{\mathbf{y} \in \Delta^{m-1}} \min_{\ell \in [n]} u^L(\mathbf{y}, \ell)$. Thus, $M = \min_{\ell \in [n]} u^L(\mathbf{y}^*, \ell)$, and we have

$$u^L(\mathbf{x}, j) < M = \min_{\ell \in [n]} u^L(\mathbf{y}^*, \ell) \leq u^L(\mathbf{y}^*, j)$$

which implies that $\mathbf{y}^* \in U_j(\mathbf{x})$. It follows that $M = \max_{\mathbf{y} \in \overline{U_j(\mathbf{x})}} \min_{\ell \in [n]} u^L(\mathbf{y}, \ell)$ and thus

$$u^L(\mathbf{x}, j) < \max_{\mathbf{y} \in \overline{U_j(\mathbf{x})}} \min_{\ell \in [n]} u^L(\mathbf{y}, \ell) \leq \max_{\mathbf{y} \in \overline{U_j(\mathbf{x})}} \min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}, \ell),$$

so (6) does not hold. □

We summarize our results from this section in Theorem 3.3. At this point, it might be tempting to think that with the polytopal constraint imposed, we would also be able to construct an explicit payoff matrix $\tilde{u}^F$ to implement $\widetilde{BR}_{\mathrm{P}}$. Unfortunately, this is not true

as Example 3.4 illustrates. Surprisingly though, in the next section we will show that, even though we cannot construct a payoff matrix that implements $\widetilde{BR}_{\mathrm{P}}$ exactly, every strategy profile $(\mathbf{x}, j)$ that is $\widetilde{BR}_{\mathrm{P}}$-inducible is in fact also payoff-inducible. We also present an efficient algorithm for computing a payoff matrix $\tilde{u}^F$ to induce such $(\mathbf{x}, j)$.

Theorem 3.3. *A strategy profile $(\mathbf{x}, j)$ is BR-inducible if and only if $u^L(\mathbf{x}, j) \geq M$. The result holds even if we require the best response correspondence to be polytopal.*

Example 3.4. Consider a 3×3 game with the leader payoff matrix given in Figure 1. Let $\widetilde{BR}_{\mathrm{P}}$ be a polytopal BR correspondence defined by the regions R_1, R_2, and R_3 in Figure 1, such that $\ell \in \widetilde{BR}_{\mathrm{P}}$ if and only if $\mathbf{y} \in R_\ell$. This best response behavior cannot be realized by any payoff matrix. To see this, suppose $\widetilde{BR}_{\mathrm{P}}$ is realized by some $\tilde{u}^F \in \mathbb{R}^{3\times 3}$. Let $\mathbf{x} = (\frac{1}{2}, \frac{1}{2}, 0)$, $\mathbf{w} = (\frac{1}{2}, 0, \frac{1}{2})$, and $\mathbf{z} = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$. We have $\widetilde{BR}_{\mathrm{P}}(\mathbf{z}) = \{1, 2, 3\}$ and $\widetilde{BR}_{\mathrm{P}}(\mathbf{w}) = \{1, 3\}$. This means that

$$u^L(\mathbf{z},\, 1) = u^L(\mathbf{z},\, 3) = u^L(\mathbf{z},\, 2), \quad \text{and} \quad u^L(\mathbf{w}, 1) = u^L(\mathbf{w}, 3) > u^L(\mathbf{w}, 2).$$

Since $\mathbf{x} = 2\mathbf{z} - \mathbf{w}$, by the linearity of the utility function, we then have $u^L(\mathbf{x}, 1) = u^L(\mathbf{x}, 3) < u^L(\mathbf{x}, 2)$, which contradicts the fact that $\widetilde{BR}_{\mathrm{P}}(\mathbf{x}) = \{1, 2\}$. □

4. Payoff Inducibility

In this section, we will show that every strategy profile that can be induced by $\widetilde{BR}_{\mathrm{P}}$ is also payoff-inducible, and a corresponding payoff matrix can be efficiently constructed. Recall that $M = \max_{\mathbf{y}\in\Delta^{m-1}} \min_{\ell\in[n]} u^L(\mathbf{y}, \ell)$ is the maximin utility of the leader. We will show the following characterization as one of our key results, which enables us to use the LP in (4) to efficiently compute a payoff matrix that achieves the optimal inducible utility.

Theorem 4.1. *A strategy profile $(\mathbf{x}, j)$ is payoff-inducible if and only if $u^L(\mathbf{x}, j) \geq M$. Furthermore, a matrix $\tilde{u}^F$ inducing $(\mathbf{x}, j)$ can be constructed in polynomial time.*

One direction of the characterization is easy to show. Indeed, if $(\mathbf{x}, j)$ is payoff-inducible, then it is also BR-inducible, and as seen in Section 3, it holds that $u^L(\mathbf{x}, j) \geq M$.

Now consider any profile $(\mathbf{x}, j)$ such that $u^L(\mathbf{x}, j) \geq M$. Recall that $U_j(\mathbf{x}) = \{\mathbf{y} \in \Delta^{m-1} : u^L(\mathbf{y}, j) > u^L(\mathbf{x}, j)\}$. Without loss of generality, in what follows, we can also assume that $U_j(\mathbf{x}) \neq \varnothing$: if $U_j(\mathbf{x}) = \varnothing$, then $(\mathbf{x}, j)$ will be an SSE if the follower always responds by playing j; this can easily be achieved by claiming that j strictly dominates all other strategies, i.e., by letting $\tilde{u}^F(i, j) = 1$ and $\tilde{u}^F(i, \ell) = 0$ for all $i \in [m]$ and $\ell \in [n] \setminus \{j\}$.

We begin by analyzing the following payoff function that forms the basis of our approach. Let $\widehat{S} \subseteq [n] \setminus \{j\}$ and pick $k \in \arg\min_{\ell\in\widehat{S}} u^L(\mathbf{x}, \ell)$ arbitrarily. For all $\mathbf{y} \in \Delta^{m-1}$, let

$$\tilde{u}^F(\mathbf{y}, \ell) = \begin{cases} -u^L(\mathbf{y}, \ell) & \text{if } \ell \in \widehat{S} \\ -u^L(\mathbf{y}, k) - 1 & \text{if } \ell \in [n] \setminus (\widehat{S} \cup \{j\}) \\ -u^L(\mathbf{y}, k) + \alpha \left(u^L(\mathbf{x}, j) - u^L(\mathbf{y}, j)\right) & \text{if } \ell = j \end{cases} \tag{7}$$

where $\alpha > 0$ is a constant. In what follows, we will let $\widetilde{BR}$ denote the best response correspondence corresponding to $\tilde{u}^F$, i.e., $\widetilde{BR}(\mathbf{y}) = \arg\max_{\ell\in[n]} \tilde{u}^F(\mathbf{y}, \ell)$. Note that once

$\widehat{S}$ and α are given, we can compute the payoff matrix corresponding to $\tilde{u}^F$ in polynomial time. Then, the hope is that with appropriately chosen $\widehat{S}$ and α, this payoff matrix will induce $(\mathbf{x}, j)$. Indeed, $\tilde{u}^F$ has the following nice properties:

i. Strategy j is indeed a best response to $\mathbf{x}$, since, by the choice of k we have

$$\tilde{u}^F(\mathbf{x}, j) = -u^L(\mathbf{x}, k) \geq -\min_{\ell \in \widehat{S}} u^L(\mathbf{x}, \ell) = \max_{\ell \in \widehat{S}} \tilde{u}^F(\mathbf{x}, \ell).$$

ii. Any $\ell \in [n] \setminus (\widehat{S} \cup \{j\})$ cannot be a best response of the follower as it is strictly dominated by k, i.e., $\tilde{u}^F(\mathbf{y}, \ell) < \tilde{u}^F(\mathbf{y}, k)$ for all $\mathbf{y} \in \Delta^{m-1}$. Thus, $\widetilde{BR}(\mathbf{y}) \subseteq \widehat{S} \cup \{j\}$ for all $\mathbf{y} \in \Delta^{m-1}$.

iii. If j is a best response to some $\mathbf{y} \in \Delta^{m-1}$, then $u^L(\mathbf{y}, j) \leq u^L(\mathbf{x}, j)$. Indeed, $j \in \widetilde{BR}(\mathbf{y})$ implies that

$$\tilde{u}^F(\mathbf{y}, j) = \max_{\ell \in [n]} \tilde{u}^F(\mathbf{y}, \ell) \geq \tilde{u}^F(\mathbf{y}, k).$$

Substituting $\tilde{u}^F(\mathbf{y}, j) = -u^L(\mathbf{y}, k) + \alpha \left(u^L(\mathbf{x}, j) - u^L(\mathbf{y}, j)\right)$ into this inequality and rearranging the terms immediately gives $u^L(\mathbf{y}, j) \leq u^L(\mathbf{x}, j)$.

iv. If any $\ell \in \widehat{S}$ is a best response to some $\mathbf{y} \in \Delta^{m-1}$, then it holds that $\tilde{u}^F(\mathbf{y}, \ell) = \max_{\ell' \in \widehat{S}} \tilde{u}^F(\mathbf{y}, \ell')$, which implies that

$$u^L(\mathbf{y}, \ell) = \min_{\ell' \in \widehat{S}} u^L(\mathbf{y}, \ell'). \tag{8}$$

Therefore, if it also holds for the $\mathbf{y}$ in (iv) that

$$\min_{\ell' \in \widehat{S}} u^L(\mathbf{y}, \ell') \leq u^L(\mathbf{x}, j),$$

then by (8) we will have $u^L(\mathbf{y}, \ell) \leq u^L(\mathbf{x}, j)$ for every $\ell \in \widetilde{BR}(\mathbf{y}) \cap \widehat{S}$. This, together with (ii) and (iii), will imply that $u^L(\mathbf{x}, j) \geq u^L(\mathbf{y}, \ell)$ for every $\ell \in \widetilde{BR}(\mathbf{y})$. Therefore, $(\mathbf{x}, j)$ will indeed form an SSE given that $j \in \widetilde{BR}(\mathbf{x})$ according to (i). We state this observation as the following lemma.

Lemma 4.2. *If $\min_{\ell' \in \widehat{S}} u^L(\mathbf{y}, \ell') \leq u^L(\mathbf{x}, j)$ holds for all $\mathbf{y} \in \Delta^{m-1}$ such that $\widetilde{BR}(\mathbf{y}) \cap \widehat{S} \neq \varnothing$, then the payoff matrix defined by (7) induces $(\mathbf{x}, j)$.*

Theorem 4.1 then follows from the following proposition, which we prove in Section 4.1 below.

Proposition 4.3. *If $u^L(\mathbf{x}, j) \geq M$ and $U_j(x) \neq \varnothing$, then we can construct $\widehat{S} \subseteq [n] \setminus \{j\}$ and $\alpha > 0$ in polynomial time, with which the condition of Lemma 4.2 holds for $\tilde{u}^F$ defined by (7).*

4.1 Proof of Proposition 4.3

The proof relies on Farkas' Lemma presented below.

Lemma 4.4 (Farkas' Lemma (Boyd & Vandenberghe, 2014))**.** *Let* $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ *and* $\mathbf{b} \in \mathbb{R}^{n_1}$*. Then exactly one of the following statements is true:*

1. *there exists* $\mathbf{z} \in \mathbb{R}^{n_2}$ *such that* $\mathbf{A}\mathbf{z} = \mathbf{b}$ *and* $\mathbf{z} \geq 0$*;*
2. *there exists* $\mathbf{z} \in \mathbb{R}^{n_1}$ *such that* $\mathbf{A}^\mathsf{T}\mathbf{z} \geq 0$ *and* $\mathbf{b} \cdot \mathbf{z} < 0$*.*

Consider any strategy profile $(\mathbf{x}, j)$ with $u^L(\mathbf{x}, j) \geq M$ and $U_j(x) \neq \varnothing$. We begin by taking care of a simple case, as an immediate corollary of Lemma 4.2.

Corollary 4.5. *A matrix* $\tilde{u}^F$ *that induces* $(\mathbf{x}, j)$ *can be constructed in polynomial time if it holds that*

$$u^L(\mathbf{x}, j) \geq M_{-j} := \max_{\mathbf{y} \in \Delta^{m-1}} \min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}, \ell). \tag{9}$$

Proof. Let $\widehat{S} = [n] \setminus \{j\}$. Then, for every $\mathbf{y} \in \Delta^{m-1}$, we immediately obtain that

$$u^L(\mathbf{x}, j) \geq \max_{\mathbf{y} \in \Delta^{m-1}} \min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}, \ell) \geq \min_{\ell \in \widehat{S}} u^L(\mathbf{y}, \ell)$$

By Lemma 4.2, the payoff matrix defined by (7) (with, say, $\alpha = 1$) then induces $(\mathbf{x}, j)$, and can clearly be computed in polynomial time. □

The more challenging case is when (9) does not hold (e.g., the case with the profile $(\mathbf{x}, 1)$ in Example 3.4). In what follows, we prove Proposition 4.3 by showing that there is still a choice of $\widehat{S}$ and α that leads to the condition in Lemma 4.2, even when (9) does not hold. Thus, from now on, we assume that

$$u^L(\mathbf{x}, j) < M_{-j}. \tag{10}$$

We define the following useful components. By Lemma 3.2 and the assumption that $u^L(\mathbf{x}, j) \geq M$, we know that

$$u^L(\mathbf{x}, j) \geq V := \max_{\mathbf{y} \in \overline{U_j(\mathbf{x})}} \min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}, \ell). \tag{11}$$

Since $\overline{U_j(\mathbf{x})} \neq \varnothing$, there exists $\mathbf{y}^* \in \overline{U_j(\mathbf{x})}$ such that

$$\min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}^*, \ell) = V, \tag{12}$$

which can be computed efficiently by solving an LP (i.e., maximize μ, subject to $\mu \leq u^L(\mathbf{y}, \ell)$ for all $\ell \in [n] \setminus \{j\}$ and $\mathbf{y} \in \overline{U_j(\mathbf{x})}$). We then let

$$S = \{\ell \in [n] \setminus \{j\} \,:\, u^L(\mathbf{y}^*, \ell) = V\}.$$

Before we proceed, we prove two useful technical results.

Lemma 4.6. $u^L(\mathbf{y}^*, j) = u^L(\mathbf{x}, j)$.

Proof. For the sake of contradiction, suppose that $u^L(\mathbf{y}^*, j) \neq u^L(\mathbf{x}, j)$. Since $\mathbf{y}^* \in \overline{U_j(\mathbf{x})}$, we have that $u^L(\mathbf{y}^*, j) \geq u^L(\mathbf{x}, j)$, so it must be that $u^L(\mathbf{y}^*, j) > u^L(\mathbf{x}, j)$.

The assumption (10) that $u^L(\mathbf{x}, j) < M_{-j}$ implies that there exists $\hat{\mathbf{y}} \in \Delta^{m-1}$ such that

$$\min_{\ell \in [n] \setminus \{j\}} u^L(\hat{\mathbf{y}}, \ell) > u^L(\mathbf{x}, j) \geq V,$$

where we also use (11). Now that $\min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}^*, \ell) = V$ by (12), by the concavity of $\min_{\ell \in [n] \setminus \{j\}} u^L(\cdot, \ell)$, it follows that $\min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{z}, \ell) > V$ for all $\mathbf{z}$ on the segment $[\hat{\mathbf{y}}, \mathbf{y}^*)$; $\mathbf{z} \in \Delta^{m-1}$ as Δ^{m-1} is convex. Now that we have $u^L(\mathbf{y}^*, j) > u^L(\mathbf{x}, j)$ under our assumption, when $\mathbf{z}$ is sufficiently close to $\mathbf{y}^*$, we can have $u^L(\mathbf{z}, j) \geq u^L(\mathbf{x}, j)$ and hence, $\mathbf{z} \in \overline{U_j(\mathbf{x})}$. This leads to the contradiction that

$$V = \max_{\mathbf{y} \in \overline{U_j(\mathbf{x})}} \min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}, \ell) \geq \min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{z}, \ell) > V. \qquad \square$$

Lemma 4.7. $\min_{\ell \in S} u^L(\mathbf{y}, \ell) < V$ *for all* $\mathbf{y} \in U_j(\mathbf{x})$.

Proof. For the sake of contradiction, assume that there exists $\hat{\mathbf{y}} \in U_j(\mathbf{x})$ such that

$$\min_{\ell \in S} u^L(\hat{\mathbf{y}}, \ell) \geq V.$$

By assumption (10) that $u^L(\mathbf{x}, j) < M_{-j}$, we have that there exists $\mathbf{z} \in \Delta^{m-1}$ such that $\min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{z}, \ell) > u^L(\mathbf{x}, j) \geq V$, which immediately yields the following, given that $S \subseteq [n] \setminus \{j\}$ by definition:

$$\min_{\ell \in S} u^L(\mathbf{z}, \ell) > V.$$

By definition, $u^L(\mathbf{y}^*, \ell) = V$ for all $\ell \in S$, which also implies that $u^L(\mathbf{y}^*, \ell) > V$ for all $\ell \in [n] \setminus (\{j\} \cup S)$ (otherwise, we would have $\min_{\ell \in [n] \setminus \{j\} u^L(\mathbf{y}^*, \ell)} < V$). Thus, we have

$$\min_{\ell \in S} u^L(\mathbf{y}^*, \ell) = V \quad \text{and} \quad \min_{\ell \in [n] \setminus (\{j\} \cup S)} u^L(\mathbf{y}^*, \ell) > V.$$

Now consider a point $\mathbf{w}$ on the segment $(\mathbf{y}^*, \hat{\mathbf{y}}]$. Since $\mathbf{y}^* \in \overline{U_j(\mathbf{x})}$ and $\hat{\mathbf{y}} \in U_j(\mathbf{x})$, i.e., $u^L(\mathbf{y}^*, j) \geq u^L(\mathbf{x}, j)$ and $u^L(\hat{\mathbf{y}}, j) > u^L(\mathbf{x}, j)$, we have $u^L(\mathbf{w}, j) > u^L(\mathbf{x}, j)$ and hence, $\mathbf{w} \in U_j(\mathbf{x})$. In addition, by continuity, when $\mathbf{w}$ is sufficiently close to $\mathbf{y}^*$, we have

$$\min_{\ell \in [n] \setminus (\{j\} \cup S)} u^L(\mathbf{w}, \ell) > V. \tag{13}$$

By concavity of the function $\min_{\ell \in S} u^L(\cdot, \ell)$, since $\min_{\ell \in S} u^L(\mathbf{y}, \ell) \geq V$ for both $\mathbf{y} \in \{\mathbf{y}^*, \hat{\mathbf{y}}\}$, we have

$$\min_{\ell \in S} u^L(\mathbf{w}, \ell) \geq V. \tag{14}$$

Analogously, we can find a point $\mathbf{w}' \in U_j(\mathbf{x})$ on the segment $(\mathbf{w}, \mathbf{z}]$, such that (13) and (14) hold for $\mathbf{w}'$ while (14) is strict, in particular. Thus, we have

$$\min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{w}', \ell) > V = \max_{\mathbf{y} \in \overline{U_j(\mathbf{x})}} \min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{y}, \ell),$$

which is a contradiction as $\mathbf{w}' \in U_j(\mathbf{x})$. $\square$

In what follows, we use the coordinates $(y_1, \ldots, y_{m-1})$ for every point $\mathbf{y} \in \Delta^{m-1}$, i.e., we have

$$\Delta^{m-1} = \left\{ (y_1, \ldots, y_{m-1}) \in \mathbb{R}_{\geq 0} \,:\, \sum_{i=1}^{m-1} y_i \leq 1 \right\}.$$

Accordingly, we can write the utility function as

$$u^L(\mathbf{y}, \ell) = \mathbf{g}_\ell \cdot \mathbf{y} + u^L(m, \ell),$$

where $\mathbf{g}_\ell \in \mathbb{R}^{m-1}$ and its i-th component is $g_{\ell,i} = u^L(i, \ell) - u^L(m, \ell)$; "$\cdot$" denotes the inner product. Hence, we have

$$u^L(\mathbf{y}, \ell) = \mathbf{g}_\ell \cdot (\mathbf{y} - \mathbf{y}^*) + u^L(\mathbf{y}^*, \ell) = \begin{cases} \mathbf{g}_\ell \cdot (\mathbf{y} - \mathbf{y}^*) + V & \text{if } \ell \in S \\ \mathbf{g}_j \cdot (\mathbf{y} - \mathbf{y}^*) + u^L(\mathbf{x}, j) & \text{if } \ell = j \end{cases} \tag{15}$$

where $u^L(\mathbf{y}^*, \ell) = V$ for all $\ell \in S$ by the definition of S, and $u^L(\mathbf{y}^*, j) = u^L(\mathbf{x}, j)$ by Lemma 4.6. Note that since $U_j(\mathbf{x}) \neq \varnothing$, it must be that $\mathbf{g}_j \neq 0$.

We also write the m boundary conditions that define Δ^{m-1} as $\mathbf{e}_i \cdot \mathbf{y} \geq \beta_i$. Namely, for each $i \in [m-1]$, let $\mathbf{e}_i \in \mathbb{R}^{m-1}$ be the i-th unit vector and $\beta_i = 0$, while $\mathbf{e}_m = (-1, \ldots, -1) \in \mathbb{R}^{m-1}$ and $\beta_m = -1$. Thus, $\Delta^{m-1} = \{\mathbf{y} \in \mathbb{R}^{m-1} \,:\, \mathbf{e}_i \cdot \mathbf{y} \geq \beta_i \text{ for all } i \in [m]\}$. Let

$$B = \{i \in [m] \,:\, \mathbf{e}_i \cdot \mathbf{y}^* = \beta_i\}$$

be the set of boundary conditions that are tight for $\mathbf{y}^*$. Note that for any $\mathbf{y} \in \Delta^{m-1}$ we have

$$\mathbf{e}_i \cdot (\mathbf{y} - \mathbf{y}^*) \geq 0 \quad \text{for all } i \in B. \tag{16}$$

We can now prove the following result using Farkas' Lemma (Lemma 4.4), which allows us to express $-\mathbf{g}_j$ as a non-negative linear combination of $\mathbf{g}_\ell$'s and $\mathbf{e}_i$'s.

Lemma 4.8. *$-\mathbf{g}_j$ can be expressed as a non-negative linear combination of $\{\mathbf{g}_\ell : \ell \in S\} \cup \{\mathbf{e}_i : i \in B\}$, i.e. $-\mathbf{g}_j = \sum_{\ell \in S} \lambda_\ell \mathbf{g}_\ell + \sum_{i \in B} \mu_i \mathbf{e}_i$, where $\lambda_\ell \geq 0$ and $\mu_i \geq 0$.*

Proof. We use Farkas' Lemma (Lemma 4.4) and let $n_1 = m - 1$ and $n_2 = |S| + |B|$. The columns of $\mathbf{A}$ are exactly the vectors $\{\mathbf{g}_\ell : \ell \in S\} \cup \{\mathbf{e}_i : i \in B\}$. We set $\mathbf{b} = -\mathbf{g}_j$. Note that the first alternative of Farkas' Lemma immediately yields the statement we want to prove. Thus, we set out to prove that the second alternative cannot hold.

Assume, for the sake of contradiction, that there exists $\mathbf{z} \in \mathbb{R}^{m-1}$ such that $\mathbf{A}^\mathsf{T}\mathbf{z} \geq 0$ and $\mathbf{b} \cdot \mathbf{z} < 0$, i.e., $\mathbf{g}_\ell \cdot \mathbf{z} \geq 0$ for all $\ell \in S$, $\mathbf{e}_i \cdot \mathbf{z} \geq 0$ for all $i \in B$, and $\mathbf{g}_j \cdot \mathbf{z} > 0$.

Then, by picking $\delta > 0$ sufficiently small, it holds for $\mathbf{y} = \mathbf{y}^* + \delta\mathbf{z}$ that:

- By (15), we have the following for all $\ell \in S$:

$$u^L(\mathbf{y}, \ell) = \mathbf{g}_\ell \cdot (\mathbf{y} - \mathbf{y}^*) + V = \delta \mathbf{g}_\ell \cdot \mathbf{z} + V \geq V.$$

  In addition,

$$u^L(\mathbf{y}, j) = \mathbf{g}_j \cdot (\mathbf{y} - \mathbf{y}^*) + u^L(\mathbf{y}^*, j) = \delta \mathbf{g}_j \cdot \mathbf{z} + u^L(\mathbf{x}, j) > u^L(\mathbf{x}, j).$$

- $\mathbf{y} \in \Delta^{m-1}$: For $i \in B$, we immediately obtain that $\mathbf{e}_i \cdot \mathbf{y} = \mathbf{e}_i \cdot (\mathbf{y}^* + \delta \mathbf{z}) \geq \mathbf{e}_i \cdot \mathbf{y}^* = \beta_i$, which means that these boundary conditions are satisfied. For $i \in [m] \setminus B$, we know that $\mathbf{e}_i \cdot \mathbf{y}^* > \beta_i$ and thus by picking $\delta > 0$ small enough, we can ensure that $\mathbf{e}_i \cdot \mathbf{y} = \mathbf{e}_i \cdot \mathbf{y}^* + \delta(\mathbf{e}_i \cdot \mathbf{z}) \geq \beta_i$.

Thus, it follows that $\mathbf{y} \in U_j(\mathbf{x})$ and $\min_{\ell \in S} u^L(\mathbf{y}, \ell) \geq V$. But this cannot hold according to Lemma 4.7. □

To complete the proof of Proposition 4.3, we express $-\mathbf{g}_j$ as a non-negative linear combination of the vectors $\{\mathbf{g}_\ell : \ell \in S\} \cup \{\mathbf{e}_i : i \in B\}$. By Lemma 4.8 we know that this is feasible and it is easy to see that we can find the coefficients in polynomial time (e.g. by solving an LP). We thus obtain $-\mathbf{g}_j = \sum_{\ell \in S} \lambda_\ell \mathbf{g}_\ell + \sum_{i \in B} \mu_i \mathbf{e}_i$, where $\lambda_\ell \geq 0$ for every $\ell \in S$ and $\mu_i \geq 0$ for every $i \in B$. Let $\widehat{S} = \{\ell \in S : \lambda_\ell > 0\}$. We will argue that $\widehat{S} \neq \varnothing$.

Now that $-\mathbf{g}_j = \sum_{\ell \in S} \lambda_\ell \mathbf{g}_\ell + \sum_{i \in B} \mu_i \mathbf{e}_i$, and we have $\mathbf{e}_i \cdot (\mathbf{y} - \mathbf{y}^*) \geq 0$ for all $\mathbf{y} \in \Delta^{m-1}$ and $i \in B$ by (16), it follows that, for all $\mathbf{y} \in \Delta^{m-1}$, we have

$$\begin{aligned} -\mathbf{g}_j \cdot (\mathbf{y} - \mathbf{y}^*) &= \sum_{\ell \in S} \lambda_\ell \mathbf{g}_\ell \cdot (\mathbf{y} - \mathbf{y}^*) + \sum_{i \in B} \mu_i \mathbf{e}_i \cdot (\mathbf{y} - \mathbf{y}^*) \\ &\geq \sum_{\ell \in S} \lambda_\ell \mathbf{g}_\ell \cdot (\mathbf{y} - \mathbf{y}^*) \\ &= \sum_{\ell \in \widehat{S}} \lambda_\ell \mathbf{g}_\ell \cdot (\mathbf{y} - \mathbf{y}^*), \end{aligned} \tag{17}$$

where the last transition is due to the fact that $\lambda_\ell = 0$ for all $\ell \in S \setminus \widehat{S}$, as implied by the definition of $\widehat{S}$.

Since $U_j(\mathbf{x}) \neq \varnothing$, consider any $\mathbf{y} \in U_j(\mathbf{x})$. By definition, this means that $u^L(\mathbf{y}, j) > u^L(\mathbf{x}, j)$, which further implies that $\mathbf{g}_j \cdot (\mathbf{y} - \mathbf{y}^*) > 0$ since $u^L(\mathbf{y}, j) = \mathbf{g}_j \cdot (\mathbf{y} - \mathbf{y}^*) + u^L(\mathbf{x}, j)$ by (15). By (17), we then have

$$\sum_{\ell \in \widehat{S}} \lambda_\ell \mathbf{g}_\ell \cdot (\mathbf{y} - \mathbf{y}^*) < 0.$$

Hence, $\widehat{S} \neq \varnothing$.

It remains to show that with the above $\widehat{S}$ and, in particular, $\alpha = 1/\lambda_k$ (recall that $k \in \arg\min_{\ell \in \widehat{S}} u^L(\mathbf{x}, \ell)$), the condition in Lemma 4.2 holds, i.e., we prove that $\min_{\ell \in \widehat{S}} u^L(\mathbf{y}, \ell) \leq u^L(\mathbf{x}, j)$ holds for all $\mathbf{y} \in \Delta^{m-1}$ such that $\widetilde{BR}(\mathbf{y}) \cap \widehat{S} \neq \varnothing$.

For the sake of contradiction, suppose that there exists $\mathbf{y} \in \Delta^{m-1}$ such that $\widetilde{BR}(\mathbf{y}) \cap \widehat{S} \neq \varnothing$, but $u^L(\mathbf{y}, \ell) > u^L(\mathbf{x}, j)$ for all $\ell \in \widehat{S}$. By (11), we have $u^L(\mathbf{x}, j) \geq V$, and thus $u^L(\mathbf{y}, \ell) > V$ for all $\ell \in \widehat{S}$. Since we have $u^L(\mathbf{y}, \ell) = \mathbf{g}_\ell \cdot (\mathbf{y} - \mathbf{y}^*) + V$ according to (15), it follows that

$$\mathbf{g}_\ell \cdot (\mathbf{y} - \mathbf{y}^*) > 0$$

for all $\ell \in \widehat{S}$. Using (17) and the fact that $k \in \widehat{S}$ by our choice, we then obtain

$$-\mathbf{g}_j \cdot (\mathbf{y} - \mathbf{y}^*) \geq \sum_{\ell \in \widehat{S}} \lambda_\ell \mathbf{g}_\ell \cdot (\mathbf{y} - \mathbf{y}^*) \geq \lambda_k \mathbf{g}_k \cdot (\mathbf{y} - \mathbf{y}^*).$$

By (15), we have

$$u^L(\mathbf{x}, j) - u^L(\mathbf{y}, j) = -\mathbf{g}_j \cdot (\mathbf{y} - \mathbf{y}^*).$$

Recall that it is defined that $\tilde{u}^F(\mathbf{y}, j) = -u^L(\mathbf{y}, k) + \alpha \left(u^L(\mathbf{x}, j) - u^L(\mathbf{y}, j)\right)$ as in (7). Using the above two equations and (15), we then obtain:

$$\begin{aligned}
\tilde{u}^F(\mathbf{y}, j) &= -u^L(\mathbf{y}, k) + \alpha \left(u^L(\mathbf{x}, j) - u^L(\mathbf{y}, j)\right) \\
&= -\mathbf{g}_k \cdot (\mathbf{y} - \mathbf{y}^*) - V - \alpha \mathbf{g}_j \cdot (\mathbf{y} - \mathbf{y}^*) \\
&\geq -V + (\alpha \lambda_k - 1)\mathbf{g}_k \cdot (\mathbf{y} - \mathbf{y}^*) \\
&= -V.
\end{aligned}$$

However, by (7) we also have $\tilde{u}^F(\mathbf{y}, \ell) = -u^L(\mathbf{y}, \ell)$ if $\ell \in \widehat{S}$, which implies that for all $\ell \in \widehat{S}$ it holds that

$$\tilde{u}^F(\mathbf{y}, j) \geq -V > -u^L(\mathbf{y}, \ell) = \tilde{u}^F(\mathbf{y}, \ell).$$

Hence, $\widetilde{BR}(\mathbf{y}) \cap \widehat{S} = \varnothing$, which contradicts our assumption. This concludes the proof of Proposition 4.3.

5. Robustness with Respect to Equilibrium Selection

As discussed in Section 2, a weakness of BR- and payoff-inducible strategy profiles is that the resulting games may have multiple SSEs, in which case the follower depends on the leader to choose the SSE that maximizes his utility. To avoid this, in this section, we turn our attention to strong inducibility (see Definition 2.5) and attempt to find a payoff matrix $\tilde{u}^F$ such that $\widetilde{\mathcal{G}}$ has a unique SSE.

We begin with an example showcasing that, in general, the best strongly inducible profile can be much worse than the best payoff-inducible profile.

Example 5.1. Consider a 3×2 game $\mathcal{G} = (u^L, u^F)$ with the payoff matrices given in Figure 2. Note that the follower obtains positive utility only by playing his strategy 1. Now, observe that the SSE $(\mathbf{x}^*, 1)$, $\mathbf{x}^* = (0, 0, 1) \in \Delta^2$, is payoff-inducible and yields utility 1 for the follower: it can be induced by any payoff matrix in which strategy 1 of the follower strictly dominates all other strategies. However, such a payoff matrix will also induce other SSEs, e.g., $(\mathbf{y}^*, 1)$ with $\mathbf{y}^* = (1, 0, 0) \in \Delta^2$. Indeed, it holds that no profile of the form $(\mathbf{y}, 1)$ can be *strongly* induced, and thus the optimal utility the follower can obtain at a strongly inducible profile is 0. To see this, first note that, as seen above, if the follower claims that strategy 1 is his unique best response for all points in Δ^2, then the SSE is not unique. On the other hand, if strategy 2 is a best response at some point $\mathbf{z} \in \Delta^2$, then $(\mathbf{y}, 1)$ will not be an SSE, since for the leader $u^L(\mathbf{y}, 1) < u^L(\mathbf{z}, 2)$ for any $\mathbf{y}, \mathbf{z} \in \Delta^2$. □

The problem in Example 5.1 stems from the following observation: if the follower reports a payoff matrix such that strategy 1 is the unique best response for all points in the domain, then there are multiple SSEs. This can be thought of as a "degenerate" case, since it would occur with probability 0, if the payoffs of the leader were drawn uniformly at random in $[0, 1]$. We formalize this as follows.

$$u^L = \begin{pmatrix} 1/4 & 1 \\ 0 & 1/3 \\ 1/4 & 2/3 \end{pmatrix} \qquad u^F = \begin{pmatrix} 1/2 & 0 \\ 1/2 & 0 \\ 1 & 0 \end{pmatrix}$$

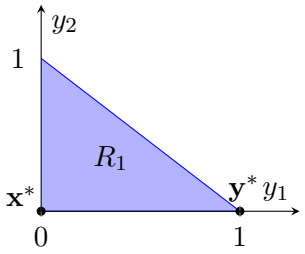

Figure 2: A game where the optimal inducible utility is 1, but the optimal *strongly* inducible utility is 0.

$$u^L = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1/2 & 1/4 \end{pmatrix} \qquad u^F = \begin{pmatrix} 1/4 & 1/3 \\ 1/4 & 1/2 \\ 1 & 1/6 \end{pmatrix}$$

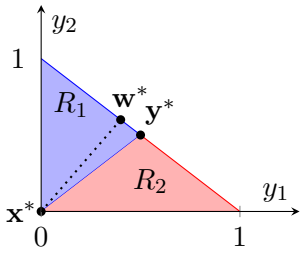

Figure 3: A non-max-degenerate game for which the optimal inducible utility cannot be achieved by any strongly inducible profile.

Definition 5.2 (Max-degeneracy). A leader payoff matrix u^L is said to be *max-degenerate*, if there exists $j \in [n]$ such that $|\arg\max_{i \in [m]} u^L(i,j)| > 1$.

We next provide an example showing that even when u^L is *not* max-degenerate, we cannot hope to *exactly* achieve the optimal inducible utility via a strongly inducible profile.

Example 5.3. Consider a 3×2 game with the leader and follower payoff matrices given in Figure 3. It is easy to check that u^L is not max-degenerate. Now, observe that the maximin utility of the leader is $M = 1/2$ and is achieved at the point $\mathbf{y}^* = (\frac{1}{2}, \frac{1}{2}, 0) \in \Delta^2$. Let $\mathbf{x}^* = (0, 0, 1) \in \Delta^2$. Since $u^L(\mathbf{x}^*, 1) = 1/2 \geq M$, it follows that $(\mathbf{x}^*, 1)$ is payoff-inducible by Theorem 4.1. Indeed, the partition (R_1, R_2) of Δ^2 in Figure 3 shows how $(\mathbf{x}^*, 1)$ can be induced. Note that $u^F(\mathbf{x}^*, 1) = 1$, while any profile different from $(\mathbf{x}^*, 1)$ yields utility strictly less than 1 for the follower. We will now show that $(\mathbf{x}^*, 1)$ cannot be strongly induced, which implies that any strongly inducible profile gives utility strictly less than 1 to the follower. Indeed, suppose that $(\mathbf{x}^*, 1)$ is induced by some $\tilde{u}^F$. If by $\tilde{u}^F$ strategy 1 is a best response to $\mathbf{y}^*$, then $(\mathbf{x}^*, 1)$ cannot be the unique SSE, since $u^L(\mathbf{x}^*, 1) = u^L(\mathbf{y}^*, 1)$. On the other hand, if strategy 2 is the only best response to $\mathbf{y}^*$, then there exists some sufficiently small $\delta > 0$ such that strategy 2 is also a best response to $\mathbf{w}^* = (\frac{1}{2} - \delta, \frac{1}{2} + \delta, 0)$ (see Figure 3). However, this means that $(\mathbf{x}^*, 1)$ cannot be an SSE, since $u^L(\mathbf{x}^*, 1) = 1/2$ and $u^L(\mathbf{w}^*, 2) = 1/2 + \delta$. □

As a result, unlike in the previous section, here we cannot hope to solve the problem exactly. However, the next theorem shows that we can approximate the optimal utility with arbitrarily good precision.

Theorem 5.4. *If u^L is not max-degenerate, then for any $\varepsilon > 0$, the follower can strongly induce a profile $(\mathbf{x}, j)$ that yields the optimal inducible utility up to an additive loss of at most ε. Furthermore, a matrix $\tilde{u}^F$ strongly inducing $(\mathbf{x}, j)$ can be constructed in time polynomial in $\log(1/\varepsilon)$ (and the size of the representation of the game).*

Proof. Let $(\mathbf{x}^*, j)$ be a payoff-inducible profile that yields the optimal inducible payoff for the follower. By Theorem 4.1, such a profile can be computed in polynomial time.

We begin by solving the following LP.

$$\begin{aligned} \max_{\delta, \mathbf{x}} \quad & \delta \\ \text{subject to} \quad & \mathbf{x} \in \Delta^{m-1} \\ & u^F(\mathbf{x}, j) \geq u^F(\mathbf{x}^*, j) - \varepsilon \\ & u^L(\mathbf{x}, j) = u^L(\mathbf{x}^*, j) + \delta \end{aligned} \tag{18}$$

Note that this LP can be solved in time polynomial in $\log(1/\varepsilon)$. Furthermore, note that the polytope of feasible points is not empty since $\delta = 0$ and $\mathbf{x} = \mathbf{x}^*$ satisfy all the constraints. Finally, the LP is not unbounded since δ can be at most $\max_{i \in [m]} u^L(i, j) - u^L(\mathbf{x}^*, j)$.

In the rest of this proof let δ and $\mathbf{x}$ denote an optimal solution to this LP. Note that we can in particular assume that $\mathbf{x}$ is a vertex of the convex polytope $P_\delta = \{\mathbf{y} \in \Delta^{m-1} : u^L(\mathbf{y}, j) = u^L(\mathbf{x}^*, j) + \delta\}$. Indeed, given a solution $\delta, \mathbf{x}$ to LP (18), if $\mathbf{x}$ is not a vertex of P_δ, then we consider the LP

$$\begin{aligned} \max_{\mathbf{y}} \quad & u^F(\mathbf{y}, j) \\ \text{subject to} \quad & \mathbf{y} \in \Delta^{m-1} \\ & u^L(\mathbf{y}, j) = u^L(\mathbf{x}^*, j) + \delta \end{aligned}$$

It is known that a solution of an LP that is also a vertex of the feasible polytope can be computed in polynomial time (Grötschel et al., 1981). Note that in this case the feasible polytope is exactly P_δ. Let $\mathbf{y}$ be an optimal solution that is a vertex of P_δ. We know that $\mathbf{x} \in P_\delta$ and $u^F(\mathbf{x}, j) \geq u^F(\mathbf{x}^*, j) - \varepsilon$, which implies that $u^F(\mathbf{y}, j) \geq u^F(\mathbf{x}^*, j) - \varepsilon$. But this means that $\delta, \mathbf{y}$ is also an optimal solution to the original LP (18). Thus, by letting $\mathbf{x} := \mathbf{y}$, we indeed have that $\mathbf{x}$ is a vertex of the convex polytope P_δ.

Let us first handle the case where $\delta = 0$ by showing that $(\mathbf{x}^*, j)$ itself can be strongly induced. Since $\delta = 0$, it follows that $U_j(\mathbf{x}^*) = \varnothing$. Indeed, if there exists $\hat{\mathbf{y}} \in \Delta^{m-1}$ with $u^L(\hat{\mathbf{y}}, j) > u^L(\mathbf{x}^*, j)$, then there exists $\mathbf{y}$ on the segment $(\mathbf{x}^*, \hat{\mathbf{y}}]$ such that $u^F(\mathbf{y}, j) \geq u^F(\mathbf{x}^*, j) - \varepsilon$ (when $\mathbf{y}$ is sufficiently close to $\mathbf{x}^*$) and $u^L(\mathbf{y}, j) > u^L(\mathbf{x}^*, j)$, a contradiction to the optimality of $\delta = 0$. Now, given that $U_j(\mathbf{x}^*) = \varnothing$, we have that $u^L(\mathbf{y}, j) \leq u^L(\mathbf{x}^*, j)$ for all $\mathbf{y} \in \Delta^{m-1}$. But since u^L is not max-degenerate (in the sense of Definition 5.2), it follows that in fact $u^L(\mathbf{y}, j) < u^L(\mathbf{x}^*, j)$ for all $\mathbf{y} \in \Delta^{m-1} \setminus \{\mathbf{x}^*\}$. Thus, if the follower always best responds with strategy j, then $(\mathbf{x}^*, j)$ will be the unique SSE. As seen before, it is easy to implement this behavior by reporting $\tilde{u}^F(i, j) = 1$ and $\tilde{u}^F(i, \ell) = 0$ for all $i \in [m]$ and $\ell \in [n] \setminus \{j\}$.

In the rest of this proof, we consider the case $\delta > 0$ and show that $(\mathbf{x}, j)$ can be strongly induced. Since $u^F(\mathbf{x}, j) \geq u^F(\mathbf{x}^*, j) - \varepsilon$, this means that at $(\mathbf{x}, j)$ the follower achieves the

optimal inducible utility up to an additive error of ε. Using the same notation as in the proof of Proposition 4.3, we let

$$B = \{i \in [m] : \mathbf{e}_i \cdot \mathbf{x} = \beta_i\}$$

denote the set of boundary conditions of Δ^{m-1} that are tight for $\mathbf{x}$. Note that since $\mathbf{x}$ is a vertex of the polytope P_δ, it follows that $B \neq \varnothing$. We let $\mathbf{h} = \sum_{i \in B} \mathbf{e}_i$. As in the proof of Proposition 4.3, we have that for all $\mathbf{y} \in \Delta^{m-1}$ it holds that

$$\mathbf{h} \cdot (\mathbf{y} - \mathbf{x}) = \sum_{i \in B} \mathbf{e}_i \cdot (\mathbf{y} - \mathbf{x}) \geq 0. \tag{19}$$

Furthermore, since $\mathbf{x}$ is a vertex of P_δ, it follows that for all $\mathbf{y} \in P_\delta \setminus \{\mathbf{x}\}$ there exists $i \in B$ such that $\mathbf{e}_i \cdot (\mathbf{y} - \mathbf{x}) > 0$, and thus

$$\mathbf{h} \cdot (\mathbf{y} - \mathbf{x}) > 0. \tag{20}$$

Indeed, if $\mathbf{e}_i \cdot (\mathbf{y} - \mathbf{x}) = 0$ for all $i \in B$ for some $\mathbf{y} \in P_\delta \setminus \{\mathbf{x}\}$, this would contradict the fact that $\mathbf{x}$ is a vertex of P_δ (i.e. the unique point in P_δ for which the boundary conditions in B are tight).

We are now ready to construct the payoff matrix reported by the follower. Pick an arbitrary $k \in \arg\min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{x}, \ell)$. For all $\mathbf{y} \in \Delta^{m-1}$ let

$$\tilde{u}^F(\mathbf{y}, \ell) = \begin{cases} -u^L(\mathbf{y}, \ell) & \text{if } \ell \in [n] \setminus \{j\} \\ -u^L(\mathbf{y}, k) + \alpha \left(u^L(\mathbf{x}, j) - u^L(\mathbf{y}, j)\right) - \mathbf{h} \cdot (\mathbf{y} - \mathbf{x}) & \text{if } \ell = j \end{cases} \tag{21}$$

where $\alpha = \left(2 \max_{i \in [m]} \max_{\ell \in [n]} \left|u^L(i, \ell)\right| + m\right) / \delta > 0$. Note that we can compute the payoff matrix corresponding to this utility function in polynomial time. In the remainder of this proof, we show that $(\mathbf{x}, j)$ is the unique SSE of the game $(u^L, \tilde{u}^F)$.

Clearly, j is a best response at $\mathbf{x}$, since

$$\tilde{u}^F(\mathbf{x}, j) = -u^L(\mathbf{x}, k) = -\min_{\ell \in [n] \setminus \{j\}} u^L(\mathbf{x}, \ell) = \max_{\ell \in [n] \setminus \{j\}} \tilde{u}^F(\mathbf{x}, \ell),$$

by the choice of k.

Next, let us show that if j is a best response at some $\mathbf{y} \in \Delta^{m-1} \setminus \{\mathbf{x}\}$, then $u^L(\mathbf{y}, j) < u^L(\mathbf{x}, j)$. Indeed, if j is a best response at $\mathbf{y}$, then in particular $\tilde{u}^F(\mathbf{y}, j) \geq \tilde{u}^F(\mathbf{y}, k)$, which implies that

$$\alpha \left(u^L(\mathbf{x}, j) - u^L(\mathbf{y}, j)\right) \geq \mathbf{h} \cdot (\mathbf{y} - \mathbf{x}). \tag{22}$$

Since $\mathbf{h} \cdot (\mathbf{y} - \mathbf{x}) \geq 0$ by (19), and $\alpha > 0$, it follows that $u^L(\mathbf{x}, j) \geq u^L(\mathbf{y}, j)$. It remains to show that $u^L(\mathbf{x}, j) \neq u^L(\mathbf{y}, j)$. But if $u^L(\mathbf{x}, j) = u^L(\mathbf{y}, j)$, then $\mathbf{y} \in P_\delta \setminus \{\mathbf{x}\}$ and so by (20) we have $\mathbf{h} \cdot (\mathbf{y} - \mathbf{x}) > 0$, which contradicts (22).

Finally, it remains to show that if $\ell \in [n] \setminus \{j\}$ is a best response at some $\mathbf{y} \in \Delta^{m-1}$, then it must be that $u^L(\mathbf{y}, \ell) < u^L(\mathbf{x}, j)$: Indeed, if $\ell \in [n] \setminus \{j\}$ is a best response at $\mathbf{y}$, then in particular $\tilde{u}^F(\mathbf{y}, j) \leq \tilde{u}^F(\mathbf{y}, \ell)$, which by (21) means that

$$\begin{aligned} \alpha \left(u^L(\mathbf{x}, j) - u^L(\mathbf{y}, j)\right) &\leq -u^L(\mathbf{y}, \ell) + u^L(\mathbf{y}, k) + \mathbf{h} \cdot (\mathbf{y} - \mathbf{x}) \\ &\leq -u^L(\mathbf{y}, \ell) + u^L(\mathbf{y}, k) + \|\mathbf{h}\|_2 \|\mathbf{y} - \mathbf{x}\|_2 \\ &\leq 2 \max_{i \in [m]} \max_{\ell' \in [n]} |u^L(i, \ell')| + \sqrt{m-1}\sqrt{m-1} \\ &\leq \alpha\delta \end{aligned}$$

by the choice of α. Thus, we obtain that $u^L(\mathbf{x}, j) - u^L(\mathbf{y}, j) \leq \delta$, which implies that $u^L(\mathbf{y}, j) \geq u^L(\mathbf{x}^*, j)$, i.e. $\mathbf{y} \in \overline{U_j(\mathbf{x}^*)}$ (since $U_j(\mathbf{x}^*) \neq \varnothing$). Since $(\mathbf{x}^*, j)$ is payoff-inducible, which means that $u^L(\mathbf{x}^*, j) \geq M$, we can use Lemma 3.2 to obtain

$$u^L(\mathbf{x}, j) = u^L(\mathbf{x}^*, j) + \delta > u^L(\mathbf{x}^*, j) \geq \min_{\ell' \in [n] \setminus \{j\}} u^L(\mathbf{y}, \ell') = u^L(\mathbf{y}, \ell)$$

where the last equality comes from the fact that ℓ is a best response at $\mathbf{y}$, i.e., in particular $\tilde{u}^F(\mathbf{y}, \ell) = \max_{\ell' \in [n] \setminus \{j\}} \tilde{u}^F(\mathbf{y}, \ell')$. □

6. Conclusion and Future Work

Our work — essentially establishing that the follower can always optimally and efficiently deceive the leader — demonstrates the power of exploiting information asymmetry in Stackelberg games. Indeed, we have shown that the follower can exploit such an asymmetry to the fullest: any strategy profile can be induced as an SSE as long as it provides the leader her maximin utility. In particular, our results indicate that there are inherent risks when the leader uses a learning algorithm to decide on how to commit in a Stackelberg game. An interesting question that emerges is thus how to design countermeasures to mitigate the potential loss of a learning leader due to possible deceptive behavior of the follower. This problem has been studied by Gan et al. (2019b) by using a mechanism design approach. They proposed that the leader could counteract by committing to a mechanism, which prescribes a strategy to play for each possible payoff matrix the follower uses. Such a mechanism admits an easy representation in their model because there is only a finite number of payoff matrices the follower can use, but this is not the case for our model.

It would also be interesting to explore whether the optimal follower payoff matrix (or a good approximation thereof) can still be computed efficiently, when additional constraints are imposed on how much the follower can deviate from his true payoff matrix. Furthermore, our results rely on having access to the leader's payoff matrix u^L, so it would be interesting to see whether deception is still possible when partial access to u^L is given, or the values of u^L are revealed in an online manner via querying. Finally, another interesting direction would be to perform empirical analyses to study the *average* utility gain of the follower, as well as the average loss of the leader, using both synthetic and real world data.

Acknowledgments

Georgios Birmpas was supported by the ERC Starting grant number 639945 (ACCORD), the ERC Advanced Grant 788893 AMDROMA "Algorithmic and Mechanism Design Research in Online Markets", and the MIUR PRIN project ALGADIMAR "Algorithms, Games, and Digital Markets". Jiarui Gan was supported by the EPSRC International Doctoral Scholars Grant EP/N509711/1 and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 945719). Alexandros Hollender was supported by an EPSRC doctoral studentship (Reference 1892947). Ninad Rajgopal was supported by the Future Leaders Fellowship MR/S031545/1. A preliminary version appeared in Proceedings of the 34th Conference

on Neural Information Processing Systems (NeurIPS 2020). We would like to thank the anonymous reviewers of NeurIPS and JAIR for their valuable comments.

References

Babichenko, Y. (2016). Query complexity of approximate Nash equilibria. *Journal of the ACM*, *63*(4), 36:1–36:24.

Babichenko, Y., & Rubinstein, A. (2017). Communication complexity of approximate Nash equilibria. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 878–889.

Balcan, M., Blum, A., Haghtalab, N., & Procaccia, A. D. (2015). Commitment without regrets: Online learning in Stackelberg security games. In *Proceedings of the 16th ACM Conference on Economics and Computation (EC)*, pp. 61–78.

Barreno, M., Nelson, B., Joseph, A. D., & Tygar, J. D. (2010). The security of machine learning. *Machine Learning*, *81*(2), 121–148.

Ben-Porat, O., & Tennenholtz, M. (2019). Regression equilibrium. In *Proceedings of the 2019 ACM Conference on Economics and Computation (EC)*, pp. 173–191.

Blum, A., Haghtalab, N., & Procaccia, A. D. (2014). Learning optimal commitment to overcome insecurity. In *Proceedings of the 28th Conference on Neural Information Processing Systems (NIPS)*, pp. 1826–1834.

Blum, A., Jackson, J. C., Sandholm, T., & Zinkevich, M. (2004). Preference elicitation and query learning. *Journal of Machine Learning Research*, *5*, 649–667.

Blumrosen, L., & Nisan, N. (2007). Combinatorial auctions. In *Algorithmic Game Theory*, chap. 11, pp. 267–299. Cambridge University Press.

Boyd, S. P., & Vandenberghe, L. (2014). *Convex Optimization*. Cambridge University Press.

Brown, G. W. (1949). Some notes on computation of game solutions. *RAND corporation report*, *P-78*.

Chen, X., Cheng, Y., & Tang, B. (2017). Well-supported versus approximate Nash equilibria: Query complexity of large games. In *Proceedings of the 8th ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, pp. 57:1–57:9.

Chen, Y., Liu, Y., & Podimata, C. (2019). Grinding the space: Learning to classify against strategic agents. *CoRR*, *abs/1911.04004*.

Chen, Y., Podimata, C., Procaccia, A. D., & Shah, N. (2018). Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC)*, pp. 9–26.

Conen, W., & Sandholm, T. (2001). Preference elicitation in combinatorial auctions. In *Proceedings of the 3rd ACM conference on Electronic Commerce (EC)*, pp. 256–259.

Conitzer, V., & Sandholm, T. (2006). Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM Conference on Electronic Commerce (EC)*, pp. 82–90.

Dekel, O., Fischer, F. A., & Procaccia, A. D. (2010). Incentive compatible regression learning. *Journal of Computer and System Sciences*, *76*(8), 759–777.

Delle Fave, F. M., Jiang, A. X., Yin, Z., Zhang, C., Tambe, M., Kraus, S., & Sullivan, J. P. (2014). Game-theoretic patrolling with dynamic execution uncertainty and a case study on a real transit system. *Journal of Artificial Intelligence Research*, *50*, 321–367.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., & Wu, Z. S. (2018). Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC)*, pp. 55–70.

Elkind, E., Gan, J., Obraztsova, S., Rabinovich, Z., & Voudouris, A. A. (2021). Protecting elections by recounting ballots. *Artificial Intelligence*, *290*, article 103401.

Fang, F., Jiang, A. X., & Tambe, M. (2013). Protecting moving targets with multiple mobile resources. *Journal of Artificial Intelligence Research*, *48*, 583–634.

Fearnley, J., Gairing, M., Goldberg, P. W., & Savani, R. (2015). Learning equilibria of games via payoff queries. *Journal of Machine Learning Research*, *16*, 1305–1344.

Gan, J., An, B., & Vorobeychik, Y. (2015). Security games with protection externalities. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15)*, p. 914–920.

Gan, J., Guo, Q., Tran-Thanh, L., An, B., & Wooldridge, M. (2019a). Manipulating a learning defender and ways to counteract. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8274–8283.

Gan, J., Xu, H., Guo, Q., Tran-Thanh, L., Rabinovich, Z., & Wooldridge, M. (2019b). Imitative follower deception in Stackelberg games. In *Proceedings of the 2019 ACM Conference on Economics and Computation (EC)*, p. 639–657.

Goldberg, P. W., Lock, E., & Marmolejo-Cossío, F. (2020). Learning strong substitutes demand via queries. In *Proceedings of The 16th Conference on Web and Internet Economics (WINE)*, p. to appear.

Goldberg, P. W., & Marmolejo-Cossío, F. J. (2018). Learning convex partitions and computing game-theoretic equilibria from best response queries. In *International Conference on Web and Internet Economics (WINE)*, pp. 168–187.

Goldberg, P. W., Marmolejo-Cossío, F. J., & Wu, Z. S. (2019). Logarithmic query complexity for approximate Nash computation in large games. *Theory of Computing Systems*, *63*(1), 26–53.

Goldberg, P. W., & Roth, A. (2016). Bounds for the query complexity of approximate equilibria. *ACM Transactions on Economics and Computation*, *4*(4), 24:1–24:25.

Goldberg, P. W., & Turchetta, S. (2017). Query complexity of approximate equilibria in anonymous games. *Journal of Computer and System Sciences*, *90*, 80–98.

Grötschel, M., Lovász, L., & Schrijver, A. (1981). The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, *1*(2), 169–197.

Hart, S., & Mansour, Y. (2010). How long to equilibrium? the communication complexity of uncoupled equilibrium procedures. *Games and Economic Behavior*, *69*(1), 107–126.

Hart, S., & Nisan, N. (2018). The query complexity of correlated equilibria. *Games and Economic Behavior*, *108*, 401–410.

Hossain, S., & Shah, N. (2020). The effect of strategic noise in linear regression. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 511–519.

Korzhyk, D., Yin, Z., Kiekintveld, C., Conitzer, V., & Tambe, M. (2011). Stackelberg vs. Nash in security games: An extended investigation of interchangeability, equivalence, and uniqueness. *Journal of Artificial Intelligence Research*, *41*, 297–327.

Lahaie, S. M., & Parkes, D. C. (2004). Applying learning algorithms to preference elicitation. In *Proceedings of the 5th ACM conference on Electronic commerce (EC)*, pp. 180–188.

Letchford, J., Conitzer, V., & Munagala, K. (2009). Learning and approximating the optimal strategy to commit to. In *International Symposium on Algorithmic Game Theory (SAGT)*, pp. 250–262.

Lowd, D., & Meek, C. (2005). Adversarial learning. In *Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining (KDD)*, pp. 641–647.

Marmolejo-Cossío, F. J., Brigham, E., Sela, B., & Katz, J. (2019). Competing (semi-) selfish miners in Bitcoin. In *Proceedings of the 1st ACM Conference on Advances in Financial Technologies*, pp. 89–109.

Meir, R., Procaccia, A. D., & Rosenschein, J. S. (2012). Algorithms for strategyproof classification. *Artificial Intelligence*, *186*, 123–156.

Nguyen, T. H., & Xu, H. (2019). Imitative attacker deception in Stackelberg security games. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 528–534.

Nisan, N., & Segal, I. (2006). The communication requirements of efficient allocations and supporting prices. *Journal of Economic Theory*, *129*(1), 192–224.

Peng, B., Shen, W., Tang, P., & Zuo, S. (2019). Learning optimal strategies to commit to. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, pp. 2149–2156.

Perote, J., & Perote-Peña, J. (2004). Strategy-proof estimators for simple regression. *Mathematical Social Sciences*, *47*(2), 153–176.

Robinson, J. (1951). An iterative method of solving a game. *The Annals of Mathematics*, *54*(2), 296–301.

Roth, A., Ullman, J., & Wu, Z. S. (2016). Watch and learn: Optimizing from revealed preferences feedback. In *Proceedings of the 48th annual ACM symposium on Theory of Computing (STOC)*, pp. 949–962.

Sun, J., Tang, P., & Zeng, Y. (2020). Games of miners. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 1323–1331.

Tambe, M. (2011). *Security and Game theory: Algorithms, Deployed Systems, Lessons Learned*. Cambridge University Press.

von Stackelberg, H. (2010). *Market structure and equilibrium*. Springer Science & Business Media.

von Stengel, B., & Zamir, S. (2004). Leadership with commitment to mixed strategies. *CDAM Research Report*, *LSE-CDAM-2004-01*.

Waggoner, B., Frongillo, R., & Abernethy, J. D. (2015). A market framework for eliciting private data. In *Advances in Neural Information Processing Systems*, pp. 3510–3518.

Yin, Y., Vorobeychik, Y., An, B., & Hazon, N. (2018). Optimal defense against election control by deleting voter groups. *Artificial Intelligence*, *259*, 32–51.

Zhang, H., Cheng, Y., & Conitzer, V. (2019). When samples are strategically selected. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Vol. 97.

Zinkevich, M. A., Blum, A., & Sandholm, T. (2003). On polynomial-time preference elicitation with value queries. In *Proceedings of the 4th ACM Conference on Electronic Commerce (EC)*, pp. 176–185.