2024 American Control Conference (ACC)
July 8-12, 2024. Toronto, Canada

Encouraging Inferable Behavior for Autonomy: Repeated Bimatrix Stackelberg Games with Observations

Mustafa O. Karabag*, Sophia Smith*, David Fridovich-Keil*, and Ufuk Topcu*

***Abstract*— When interacting with other non-competitive decision-making agents, it is critical for an autonomous agent to have inferable behavior: Their actions must convey their intention and strategy. For example, an autonomous car's strategy must be inferable by the pedestrians interacting with the car. We model the inferability problem using a repeated bimatrix Stackelberg game with observations where a leader and a follower repeatedly interact. During the interactions, the leader uses a fixed, potentially mixed strategy. The follower, on the other hand, does not know the leader's strategy and dynamically reacts based on observations that are the leader's previous actions. In the setting with observations, the leader may suffer from an inferability loss, i.e., the performance compared to the setting where the follower has perfect information of the leader's strategy. We show that the inferability loss is upper-bounded by a function of the number of interactions and the stochasticity level of the leader's strategy, encouraging the use of inferable strategies with lower stochasticity levels. As a converse result, we also provide a game where the required number of interactions is lower bounded by a function of the desired inferability loss.**

I. INTRODUCTION

Autonomous agents repeatedly interact with other agents, e.g., humans and other autonomous systems, in their environments during their operations. Often, the intentions and strategies of these autonomous agents are not perfectly known by the other agents, and the other agents rely on inference from the past interactions when they react to the actions of the autonomous agent. For example, an autonomous car interacts with pedestrians who intend to cross the road, and pedestrians do not have a perfect knowledge of the car's strategy. Consequently, acting in an inferable way is consequential for autonomous agents.

We model the interaction between the autonomous agent and the other agent with a bimatrix Stackelberg game. In this game, the autonomous agent is the *leader* that commits to a strategy, and the other agent is the *follower* that does not know the leader's action and reacts to the leader's strategy where the leader's actions are drawn from. The game is repeated between the agents. While the leader follows the same strategy at every interaction, the follower's strategy can change between interactions. For the autonomous car example, the fixed strategy over actions stopping and proceeding represents a version of the car's software.

We consider that the follower does not have perfect information of the leader's strategy and relies on the observations from the previous rounds. In detail, at every interaction, the follower plays optimally against the empirical action distribution from the previous interactions. For example, in the car-pedestrian scenario, the pedestrian would act based on the frequency that the car stopped in the previous interactions.

For a traditional bimatrix Stackelberg game, the leader's optimal strategy may be mixed [1]. However, in the inference setting that we consider, this strategy may not be optimal since the follower reacts to the empirically observed strategy of the leader, not the actual strategy. As a result, the leader might be better off using a less stochastic strategies since such strategies would be more inferable.

The leader's expected return in the inference setting might be lower than its expected return in the perfect information setting. We call the return gap between these settings the leader's *inferability loss*.

We show that when the follower has bounded rationality (modeled by the maximum entropy model), the leader's cumulative inferability loss is bounded above. The upper bound is a function of both the stochasticity level (trace of the covariance matrix) of the leader's strategy and the number of interactions. As the stochasticity level of the leader's strategy decreases, the inferability loss vanishes. In the extreme case where the leader's strategy is deterministic, the leader does not suffer from any inferability loss; the expected return in the inference setting is the same as the expected return in the perfect information setting. The inferability loss at interaction k is at most $\mathcal{O}(1/\sqrt{k})$, implying that $\mathcal{O}(1/\epsilon^2)$ interactions are sufficient to achieve a maximum of ϵ inferability loss.

Motivated by the bound, we use the stochasticity level as a regularization term in the leader's objective function to find optimal strategies for the inference setting. Numerical experiments show that the leader indeed suffers from an inferability loss in the inference setting, and the strategies generated by the regularized objective function lead to improved transient returns compared to the strategies that are optimal for the perfect information case.

Additionally, as a converse result, we provide an example bimatrix Stackelberg game where the inferability loss at interaction k is at least ϵ if k is not at the order of $\mathcal{O}(1/\epsilon^2)$ under the full rationality assumption for the follower.

Related work: Bimatrix Stackelberg games with a commitment to mixed strategies have been extensively studied in the literature under the assumption that the follower has perfect knowledge of the leader's strategy [1], [2], [3]. For these games, an optimal strategy for the leader can be computed in polynomial time via linear programming (assuming that

This work was supported by ARO W911NF2310317 and by the National Science Foundation under Grant No. 1652113, Grant No. 1836900, Grant No. 2211432, and Grant No. 2211548.

*The University of Texas at Austin.
Correspondence to karabag@utexas.edu.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on January 20,2025 at 20:09:25 UTC from IEEE Xplore. Restrictions apply.

the follower breaks ties in favor of the leader) [2]. The paper [4] considers Stackelberg games with partial observability where the follower observes the leader's strategy with some probability and does not otherwise. We consider a different observability setting where the follower gets observations from the leader's strategy. Papers [5], [6] also consider this observation setting. To account for the follower's partial information, [5], [6], [7] consider a robust set that represents the possible realizations of the leader's strategy and maximize the leader's worst-case return by solving a robust optimization problem. We follow a different approach and try to maximize the leader's expected return under observations by relating it to the return under the perfect information setting.

We provide a lower bound on the leader's return that involves the stochasticity level (inferability) of the leader's strategy. To our knowledge, a bound in this spirit does not exist for Stackelberg games with observations. Works [8], [9], [10] increase the stochasticity level of the control policy (the leader's strategy in our context) to improve the non-inferability in different contexts. We consider a stochasticity metric that coincides with the Fisher information metric considered in [8], [9], [10]. However, unlike these works, which focus on minimizing information and providing unachievability results, we provide an achievability result.

Human-robot interactions are more efficient if the human knows the robot's intent. Conveying intent information via movement is explored to create legible behavior [11], [12]. These works are often concerned with creating trajectories that are distant from the trajectories under other intents.

The leader's optimization problem in our setting is a bilevel optimization problem under data uncertainty [13]. Works [14], [15], [16] consider stochastic bilevel optimization problems where first the leader commits to a strategy before the data uncertainty is resolved, then the data uncertainty is resolved, and finally, the follower makes its decision with known data. In our problem, the distribution of data depends on the leader's decision[1], whereas [14], [15], [16] consider a fixed distribution of data.

We represent the boundedly rational follower using the maximum entropy model (also known as Boltzmann rationality model or quantal response) [18], [19]. Alternatively, [5], [20] consider boundedly rational followers using the anchoring theory [21] or ϵ-optimal follower models.

II. PRELIMINARIES

A. Notation

We use upper-case letters for matrices and bold-face letters for random variables. $\|\cdot\|$ denotes the L2 norm. Δ^N denotes the N-dimensional probability simplex. For $z \in \Delta^N$, the entropy of z, is

$$H(z) = \sum_{i=1}^{N} z_i \log(1/z_i)$$

where z_i is the i-th element of z. The softmax function $\sigma_\lambda : \mathbb{R}^N \to \Delta^N$ is defined as

$$\sigma_\lambda(z)_i := \frac{\exp(\lambda z_i)}{\sum_{j=1}^{N} \exp(\lambda z_j)}$$

where $\sigma_\lambda(z)_i$ is the i-th element of $\sigma_\lambda(z)$. The softmax function σ_λ is λ-Lipschitz continuous, i.e., it satisfies $\|\sigma(z) - \sigma(q)\| \leq \lambda\|z - q\|$for all $z, q \in \mathbb{R}^N$ [22].

We define the *stochasticity level* of a probability distribution $z \in \Delta^N$ as

$$\nu(z) := \sqrt{\sum_{i=1}^{N} z_i(1 - z_i)}$$

that is the square root of the trace of the covariance matrix.

B. Bimatrix Stackelberg Games with Mixed Strategies

A *bimatrix Stackelberg game* is a two-player game between a *leader* and a *follower*. The leader has m (enumerated) actions, and the follower has n (enumerated) actions. We call matrix $A \in \mathbb{R}^{m\times n}$ the leader's *utility matrix* and $B \in \mathbb{R}^{m\times n}$ the follower's utility matrix. When the leader takes action i and the follower takes action j, the leader and follower returns are A_{ij} and B_{ij} respectively.

In bimatrix Stackelberg games with *mixed strategies*, the leader has a mixed strategy $x \in \Delta^m$, and the follower has mixed strategy $y \in \Delta^n$. Let $\mathbf{i}$ and $\mathbf{j}$ denote the random versions of i and j, respectively. The leader's *expected utility* is $x^\top Ay = \mathbb{E}_{\mathbf{i}\sim x,\mathbf{j}\sim y}\left[A_{\mathbf{i},\mathbf{j}}\right]$, and the follower's expected utility is $x^\top By = \mathbb{E}_{\mathbf{i}\sim x,\mathbf{j}\sim y}\left[B_{\mathbf{i},\mathbf{j}}\right]$. When deciding on strategy y, the follower knows the leader's strategy x. This means the follower knows the probability distribution of the leader's action but does not know the leader's realized action.

The follower's goal is to maximize its expected return given x, the leader's strategy, by solving:

$$\max_{y\in\Delta^n} x^\top By.$$

An optimal solution exists for the follower's problem.

The leader's goal is to maximize its (conservative[2]) expected return, i.e., solve the bilevel optimization problem:

$$\begin{aligned} \sup_{x\in\Delta^m} \quad & \min_{y^*} x^\top Ay^* \\ \text{s.t.} \quad & y^* \in \arg\max_{y\in\Delta^n} x^\top By. \end{aligned}$$

Note that an optimal solution may not exist for this problem.

We define

$$\begin{aligned} SR(x) := & \min_{y^*} x^\top Ay^* \\ \text{s.t.} \quad & y^* \in \arg\max_{y\in\Delta^n} x^\top By. \end{aligned}$$

We refer to SR as the Stackelberg return under perfect information and full follower rationality.

[1]For single-level stochastic optimization problems, this setting is referred to as non-oblivious stochastic optimization [17].

[2]Here conservative refers to the fact that if there are multiple optimal follower strategies, the follower chooses the worst strategy for the leader.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on January 20,2025 at 20:09:25 UTC from IEEE Xplore. Restrictions apply.

C. Boundedly Rational Follower

Bounded rationality models represent the decision-making process of an agent with limited information or information processing capabilities and are often used to model the decision-making process of humans [23]. We consider the maximum entropy model (Boltzmann rationality) to represent boundedly rational followers [18].

Given the leader's strategy x, a boundedly rational follower solves the following optimization problem

$$\max_{y \in \Delta^n} x^\top B y + \frac{1}{\lambda} H(y)$$

where λ denotes the follower's rationality level. Note that for $\lambda \in (0, \infty)$, the optimal solution for the above problem is unique since the objective function is strictly concave and is given by $\sigma_\lambda(B^\top x)$ [22]. In words, the action probabilities are weighted exponentially according to their expected returns. As $\lambda \to 0$, the follower does not take its expected utility $x^\top B y$ into account and takes all available actions uniformly randomly. As $\lambda \to \infty$, the follower becomes fully rational. Given that the follower is boundedly rational with level $\lambda \in (0, \infty)$, the leader's goal is to maximize its expected utility, i.e., solve

$$\max_{x \in \Delta^m} x^\top A y^*$$

such that $y^* = \sigma_\lambda(B^\top x)$. We drop the inner optimization problem since the optimal solution to the follower's optimization problem is unique due to strict convexity. We define

$$SR_\lambda(x) := x^\top A y^* \text{ where } y^* = \sigma_\lambda(B^\top x).$$

We refer to SR_λ as the Stackelberg return under perfect information and bounded follower rationality.

III. PROBLEM FORMULATION

A. Repeated Bimatrix Stackelberg Games with Inference

Consider a bimatrix Stackelberg game with mixed strategies that is repeated K times. However, assume the follower does not know the leader's fixed mixed strategy x. Instead, the follower infers the leader's strategy from observations of the previous interactions. At interaction k, let $\hat{x}_k$ be the *sample mean estimation* of the leader's strategy. Specifically, if the leader takes actions $i_1, \ldots, i_{k-1}$ at the previous $k-1$ timesteps,

$$(\hat{x}_k)_l = \frac{\#_{t=1}^{k-1} \ (i_t = l)}{k-1},$$

where $(\hat{x}_k)_l$ is the l^{th} element of vector $\hat{x}_k$ and $\#(\cdot)$ counts the number of times the input is true.

Under these assumptions, the follower's strategy y_k at time k depends on the leader's actions in the previous $k-1$ timesteps. For this reason, y_k changes over time. For example, for a fully rational follower, $y_k^* \in \arg\max_{y \in \Delta^n} \hat{x}_k^\top B y$.

Next, we consider the following formulations of the leader's problem for different levels of follower rationality.

Fully rational follower: The leader's decision-making problem is to a priori select a strategy x^* that maximizes its expected cumulative return under inference, i.e., assuming that the follower rationally responds to the plug-in sample mean estimator of x^* at each time k. Let $\mathbf{i}_k$, $\hat{\mathbf{x}}_k$, and $\mathbf{y}_k$ be random variables denoting the unrealized versions of i_k, $\hat{x}_k$ and y_k, respectively. The leader's optimal strategy is

$$x^* = \arg\max_{x \in \Delta^m} \mathbb{E}\left[\sum_{k=1}^{K} \min_{\mathbf{y}_k^*} x^\top A \mathbf{y}_k^*\right]$$
$$\text{s.t. } \mathbf{y}_k^* \in \arg\max_{y \in \Delta^n} \hat{\mathbf{x}}_k^\top B y.$$

Here, the expectation is over the randomness in the leader's actions $\mathbf{i}_1, \ldots, \mathbf{i}_K$. The leader solves this decision problem prior to taking any action, meaning their future actions $\mathbf{i}_1, \ldots, \mathbf{i}_K$ are random variables. Since the follower's estimation $\hat{\mathbf{x}}_k$ is a function of these future actions and the follower's strategy $\mathbf{y}_k^*$ is a function of $\hat{\mathbf{x}}_k$, they are both random variables as well and therefore bolded.

Boundedly rational follower: The leader's decision problem is to find a strategy x^* such that

$$x^* = \arg\max_{x \in \Delta^m} \mathbb{E}\left[\sum_{k=1}^{K} x^\top A \mathbf{y}_k\right]$$
$$\text{s.t. } \mathbf{y}_k = \sigma_\lambda(B^\top \hat{\mathbf{x}}_k).$$

Once again, the expectation is over the randomness in the leader's (random) actions $\mathbf{i}_1, \ldots, \mathbf{i}_K$.

In the Stackelberg game with inference, the leader's strategy affects the follower's optimal strategy in two ways, regardless of the level of follower rationality. First, as in the original Stackelberg game formulation, the leader's strategy determines the expected return for different follower actions, i.e., $x^\top B$. This affects which strategies are optimal for the follower. Second, unlike in the perfect information Stackelberg game, the leader's strategy x modifies the distribution of its empirical action distribution $\hat{\mathbf{x}}_k$ and, consequently, the follower's strategy $\mathbf{y}_k$.

A strategy with a high Stackelberg return under perfect information may be highly suboptimal in a Stackelberg game with inference. Different realizations of $\hat{\mathbf{x}}_k$ lead to different solutions for $\mathbf{y}_k$. If x is poorly inferred by the follower, the follower's strategy $\mathbf{y}_k$ may yield poor returns when simultaneously played with x. In the inference setting, an optimal strategy x^* will strike a balance between having a high Stackelberg return under perfect information and efficiently conveying information about itself to the follower.

Remark 1. *Inferability is important in semi-cooperative games where the objectives of the players are weakly positively correlated. Due to the positive correlation between the utility matrices A and B, it is useful for the leader to be correctly inferred by the follower. On the other hand, since there is only a weak correlation between A and B, i.e., $A \neq B$, the leader's optimal strategy may still be mixed.*

B. Motivating Example: Pedestrian and Autonomous Car Interaction

This section describes a motivating example of the interaction between an autonomous car and a pedestrian. Similar scenarios have been considered in [24], [25].

Authorized licensed use limited to: University of Texas at Austin. Downloaded on January 20,2025 at 20:09:25 UTC from IEEE Xplore. Restrictions apply.

Consider an autonomous car moving in its lane. A pedestrian is dangerously close to the road and aims to cross. The car has the right of way and wants to proceed, as an unnecessary stop is inefficient. However, if the pedestrian decides to cross the car may need to make a dangerous emergency stop. In the event the pedestrian crosses, they may get fined for jaywalking. The pedestrian and the car must make simultaneous decisions that will determine the outcome. Since the autonomous car's software is fixed prior to deployment, the car's decision is drawn from a fixed strategy that does not change over time.

TABLE I

UTILITIES FOR THE AUTONOMOUS CAR AND PEDESTRIAN INTERACTION

		Pedestrian's actions	
		Wait	Cross
Car's actions	Stop	(0,2)	(0,1)
	Proceed	(2,0)	(-8,1)

For the pedestrian, crossing has a value $r^f_{\text{cr}} = 2$, and potentially getting fined for jaywalking has $r^f_{\text{jw}} = -1$ value. For the car, proceeding without a stop has a value $r^l_{\text{pr}} = 2$, and making an emergency stop has $r^l_{\text{em}} = -8$ value.

Scenario 1: The car stops and the pedestrian waits for the car. The pedestrian's return is $r^f_{\text{cr}} = 2$ since the car's stop allows them to cross. The car's return is 0 since it does not proceed and stops unnecessarily.

Scenario 2: The car stops, but the pedestrian crosses before the car yields the right of way. In this case, the pedestrian's return is $r^f_{\text{cr}} + r^f_{\text{jw}} = 1$ since they cross the road but risk being fined for jaywalking. Once again, the car receives a return of 0 since it does not proceed.

Scenario 3: The car proceeds and the pedestrian waits. The car's return is $r^l_{\text{pr}} = 2$ since it makes no unnecessary stops. The pedestrian gets a return of 0 since it can not cross.

Scenario 4: The car proceeds and the pedestrian crosses. The pedestrian's return is $r^f_{\text{cr}} + r^f_{\text{jw}} = 1$. While the car proceeds, it makes an emergency stop due to the crossing pedestrian, resulting in a return of $r^l_{\text{em}} = -8$.

Assume that the car stops with probability p and proceeds with probability $1-p$. If the pedestrian knows the probability p, the pedestrian would wait if $2p+0(1-p) > 1p+1(1-p)$, i.e., $p > 0.5$. Knowing that the pedestrian would wait when $p > 0.5$, the car gets a return of $0p + 2(1-p)$. Knowing that the pedestrian would cross when $p < 0.5$, the car gets a return of $0p - 8(1-p)$. Hence, it is optimal for the car to choose a p such that $p > 0.5$ and $p \approx 0.5$. While such a strategy is optimal and has a return of ≈ 1 for the car, it relies on the fact that the pedestrian has perfect information of the car's strategy. Such a strategy may not be optimal if the pedestrian does not know p and relies on observations.

Consider a scenario where the pedestrian and car will interact a certain number of times. The pedestrian estimates the car's fixed strategy using observations from previous interactions. If in most of the previous interactions the car stopped, the pedestrian would expect the car to stop in the next interaction. Knowing that the pedestrian relies on observations, the car should pick an easily inferable strategy.

If the pedestrian has a good estimate $\hat{p}$ of the car's strategy, the pedestrian will act optimally with respect to the car's actual strategy. On the flip side, the car will get a return that is close to the perfect information case. For example, consider that $p = 1$. In this case, the pedestrian has the correct estimate $\hat{p} = 1$ after a single interaction, and the car will get a return of 0 in the subsequent interactions.

If the car's strategy is not easily inferable, then the car may suffer from an *inferability loss*. For example, if p is such that $p > 0.5$ and $p \approx 0.5$, the car has an expected return of ≈ -1.5 in the second interaction. This is significantly lower compared to the expected return of 1 in the perfect information case. This is because the pedestrian's estimate will be $\hat{p} = 0$ with probability $1 - p \approx 0.5$. In those events, the pedestrian will cross, and the car will get -8 return if it proceeds. Overall, a strategy that maximizes the car's expected return over a finite number of interactions should take the pedestrian's estimation errors into account.

IV. PERFORMANCE BOUNDS UNDER INFERENCE

In this section, we compare the leader's expected utility under repeated interactions with inference with the leader's expected utility under repeated interactions with perfect information. We define

$$IR_k(x) := \mathbb{E}\left[\min_{\mathbf{y}^*_k} x^\top A \mathbf{y}^*_k\right] \quad \text{s.t. } \mathbf{y}^*_k \in \arg\max_{y \in \Delta^n} \hat{\mathbf{x}}_k^\top B y$$

that is the leader's expected (conservative) return under inference against a fully rational follower at interaction k. Similarly, we define

$$IR_{k,\lambda}(x) := \mathbb{E}\left[\sum_{k=1}^{K} x^\top A \mathbf{y}_k\right] \quad \text{s.t. } \mathbf{y}_k = \sigma_\lambda(B^\top \hat{\mathbf{x}}_k)$$

that is the leader's expected return under inference against a boundedly rational follower at the k^{th} interaction. The leader's expected return at the first interaction is arbitrary since the follower does not have any action samples. Hence, we are interested in analyzing the expected cumulative return for interactions $k = 2, \ldots, K$. Due to the linearity of expectation, the expected cumulative return can be represented as a sum of expected returns of every interaction. In the fully rational follower setting

$$\sum_{k=2}^{K} IR_k(x) = \mathbb{E}\left[\sum_{k=2}^{K} \min_{\mathbf{y}^*_k} x^\top A \mathbf{y}^*_k\right]$$

where $\mathbf{y}^*_k \in \arg\max_{y \in \Delta^n} \hat{\mathbf{x}}_k^\top B y$, and in the boundedly rational follower setting

$$\sum_{k=2}^{K} IR_{k,\lambda}(x) = \mathbb{E}\left[\sum_{k=2}^{K} x^\top A \mathbf{y}_k\right]$$

where $\mathbf{y}_k = \sigma_\lambda(B^\top \hat{\mathbf{x}}_k)$.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on January 20,2025 at 20:09:25 UTC from IEEE Xplore. Restrictions apply.

A. Achievability Bound for a Boundedly Rational Follower

The follower uses the sample mean estimator $\hat{x}_k$ to infer the leader's strategy x. As $k \to \infty$, the estimate converges to the true distribution x. Given that $\sigma(x)$ is a continuous mapping, the leader's expected utility under inference, i.e. the observation return, converges to the expected utility in the perfect information setting, i.e., the Stackelberg return. However, with a finite number of interactions these returns are not necessarily the same, and the leader may suffer from an *inferability loss*. The following result shows that the inferability loss is upper bounded by a function of the trace of the covariance matrix of the leader's strategy.

Theorem 1. *Define* $d^f = \max_{i,j} B_{i,j} - \min_{i,j} B_{i,j}$ *and* $d^l = \max_{i,j} A_{i,j} - \min_{i,j} A_{i,j}$. *We have*

$$(K-1)SR_\lambda(x) - \sum_{k=2}^{K} IR_{k,\lambda}(x) \leq \sum_{k=2}^{K} \frac{d^l d^f \lambda n^{3/2}\sqrt{m}\nu(x)}{4\sqrt{(k-1)}}.$$

Remark 2. *There are* $\frac{(k+m-2)!}{(k-1)!(m-1)!} \approx (k-1)^{m-1}$ *(assuming* $k \gg m$*) different values of* $\hat{x}_k$. *Computing the exact value of* IR *may require evaluating the expected return under all possible realizations of* $\hat{x}_k$.

The cumulative inferability loss grows sublinearly, i.e,

$$\sum_{k=2}^{K} \frac{d^l d^f \lambda n^{3/2}\sqrt{m}\nu(x)}{4\sqrt{(k-1)}} = \mathcal{O}\left(\sqrt{K}\lambda\nu(x)\right).$$

As the leader's strategy becomes deterministic, i.e., $\nu(x) \to 0$, the inferability loss vanishes to 0. In the extreme case where the leader's strategy is deterministic $\nu(x) = 0$, the leader does not suffer from an inferability loss. As the follower becomes irrational, i.e., $\lambda \to 0$, the inferability loss vanishes to 0, and when the follower is fully irrational, $\lambda = 0$, the leader does not suffer from an inferability loss since the follower's strategy is uniformly random and does not depend on observations.

The leader's optimal strategy under inference depends on various factors. Such a strategy should have a balance between having a high Stackelberg return under perfect information and having a minimal inferability loss, i.e., efficiently conveying information about itself to the follower.

The proof[3] of Theorem 1 follows from the Lipschitz continuity of the follower's response $\sigma_\lambda(\cdot)$, the Lipschitz continuity of the leader's return for different values of y, and the concentration of $\hat{\mathbf{x}}_k$ around $x(y_k)$.

B. Converse Bound for a Fully Rational Follower

Theorem 1 shows that with a boundedly rational follower, the gap between the leader's expected return in the perfect information setting and in the inference setting, $SR_\lambda(x) - IR_{k,\lambda}(x)$, is at most at the order of $\mathcal{O}(1/\sqrt{k})$ at interaction k. In other words, after $\mathcal{O}(1/\epsilon^2)$ interactions, we have $SR_\lambda(x) - IR_{k,\lambda}(x) \leq \epsilon$. In this section, we give an example for the fully rational follower setting that matches the upper bound: $\mathcal{O}(1/\epsilon^2)$ interactions are required to achieve $SR(x) - IR_k(x) \leq \epsilon$. We consider

$$A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}. \tag{1}$$

For these choices of A and B, we have

$$v^* = \sup_{x \in \Delta^m} \min_{y^*} x^\top A y \;\; \text{s.t.} \;\; y^* \in \arg\max_{y \in \Delta^n} x^\top B y = \frac{1}{2}$$

Proposition 1. *Let* A *and* B *be as defined in* (1). *For every* $\epsilon \in (0, 1/2)$ *and* $x \in \Delta^2$ *such that* $SR(x) \geq (1/2) - \epsilon$, *if*

$$k \leq \frac{1 - 20\epsilon + 132\epsilon^2}{32\epsilon^2},$$

then

$$SR(x) - IR_k(x) \geq \epsilon.$$

For small enough ϵ, the term $1/(32\epsilon^2)$ dominates the other terms. If there are $o(1/\epsilon^2)$ interactions, then the leader's expected return under inference is at least ϵ worse than its return under perfect information.

The strategies with near-optimal Stackelberg returns, i.e., $SR(x) \geq (1/2) - \epsilon$, will have poor returns under inference since they are close to the decision boundary where the follower abruptly changes its strategy and the empirical distribution may be on the other side of the decision boundary.

V. NUMERICAL EXAMPLES

In this section, we evaluate the effect of inference on repeated bimatrix Stackelberg games with boundedly rational followers. As an example, we consider the aforementioned car-pedestrian interaction and randomly generated bimatrix games. For clarity of presentation, we plot the average return $\frac{1}{K}\sum_{k=2}^{K} IR_{k,\lambda}(x)$, which is the expected cumulative return up to interaction K divided by K. We approximate the expectation with repeated simulations.

a) Car-Pedestrian Interactions: We consider the bimatrix game presented in Table I. We simulate the game play under inference for 100 interactions with a rationality constant $\lambda = 100$. The car's strategy is determined by p, i.e., the probability the car stops, and for $\lambda = 100$, the optimal p is 0.53 in the perfect information setting. We repeat the simulation $10,000$ times, and the leader's average expected return for different values of p. Results are shown in Fig. 1.

Each strategy converges to its perfect information case as the number of interactions increases. In the long run, the optimal strategy for the car in the perfect information setting, i.e., $p = 0.53$, would achieve the highest return. However, after 100 interactions, this strategy is still underperforming compared to more deterministic strategies. This is because a small error in the pedestrian's estimation $\hat{\mathbf{p}}_k$ results in large changes in the pedestrian's strategy, demonstrating the impact inference has on the leader's return. On the other hand, as we expected, the strategies with higher stopping probabilities achieve higher transient returns: More deterministic strategies are easier for the pedestrian to infer at small time horizons K, and any error in $\hat{\mathbf{p}}_k$ results in only small changes to the pedestrian's strategy.

[3]The proof for all results can be found at https://arxiv.org/abs/2310.00468.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on January 20,2025 at 20:09:25 UTC from IEEE Xplore. Restrictions apply.

b) Randomly Generated Bimatrix Games: We evaluate the performance under inference for randomly generated bimatrix games when the follower has bounded rationality. From the achievability bound given in Theorem 1,

$$(K-1)SR_\lambda(x) - c\,\nu(x) \leq \sum_{k=2}^{K} IR_{k,\lambda}(x)$$

for some constant c depending on K, A and B. We use ν as a regularizer and optimize the bound for fixed values of c:

$$x^*(c) = \arg\max_{x\in\Delta^m} SR_\lambda(x) - c\,(\nu(x))^2.$$

and compare the performance of leader strategies for different values of c. We replace ν with ν^2 in the optimization problem since the gradients of ν^2 are Lipschitz continuous. We note that even when $c = 0$, this is a nonconvex optimization problem. To find a maximum, we use gradient descent with decaying stepsize. We use the leader's optimal strategy from the Stackelberg game with a fully rational follower as the starting point for the gradient descent.

In this example, we randomly generate bimatrix games. For each bimatrix game, the entries of the leader's utility matrix A are uniformly randomly distributed between 0 and 1. The follower's utility matrix $B = {}^{A}/_{2} + {}^{C}/_{2}$, where C is a uniformly randomly distributed matrix between 0 and 1. This construction makes A and B weakly positively correlated highlighting the importance of mixed strategies and inferability as explained in Remark 1.

We randomly generate $10,000$ 4×4 bimatrix games. For each random bimatrix game, we find the leader's strategy $x^*(c)$ for $c = 0, 1, 10$, and 100. For each bimatrix game, we simulate play for 100 interactions with rationality constant $\lambda = 100$. We repeat the simulations 100 times, and the leader's return is averaged at each interaction over these simulations. Then, the leader's average return until interaction k for each bimatrix is averaged at each interaction k over all bimatrix games. Results are shown in Fig. 2.

In these simulations, higher regularization constants correspond to more inferable (less stochastic) strategies, as more weight is given to the stochasticity level of a strategy. The optimal strategies for the perfect information setting $x^*(c = 0)$ (the optimal strategy with no stochasticity regularization) achieves a higher average expected return in the long run (after 45 interactions) since the follower's estimation accuracy improves with more interactions. However, these strategies still suffer inferability loss after 100 interactions. For the first 45 interactions, the regularization constant $c = 100$ yields higher average returns, and the average return reaches its final value even after the first interaction since the generated strategies are deterministic and estimated by the follower perfectly even with a single sample.

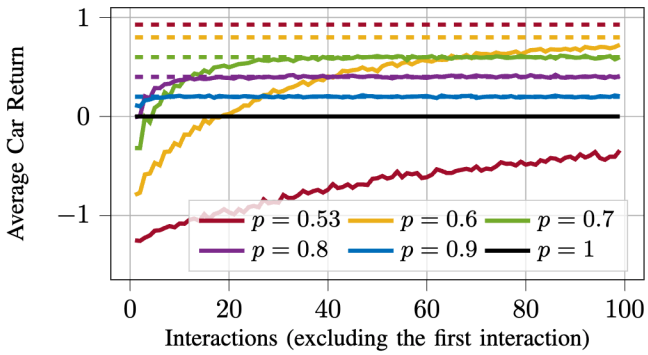

Fig. 1. The car's average return in the pedestrian-car example. Solid lines represent the average return for different strategies where p is the probability of the car stopping. Dashed lines represent the average return per interaction under perfect information, i.e., $x^\top A\sigma_\lambda(B^\top x)$ for $x = [p, 1-p]$.

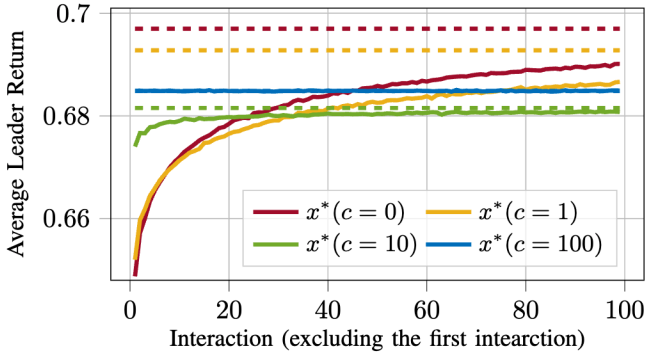

Fig. 2. The leader's average return for the randomly generated 4x4 bimatrix games. Solid lines represent the average return for the bound's local maxima for different values of the regularization constant c. Dashed lines represent the average return per interaction under perfect information, i.e., $(x^*(c))^\top A\sigma_\lambda(B^\top x^*(c))$.

VI. CONCLUSIONS

When interacting with other non-competitive agents, an agent should have an inferable behavior to inform others about intentions effectively. We model the inferability problem using a repeated bimatrix Stackelberg game where the follower infers the leader's strategy via observation from previous interactions. We show that in the inference setting, the leader may suffer from an inferability loss compared to the perfect information setting. However, this loss is upper bounded by a function that depends on the stochasticity level of the leader's strategy. The bound and experimental results show that to maximize the transient returns, the leader may be better off using a less stochastic strategy compared to the strategy that is optimal in the perfect information setting.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on January 20,2025 at 20:09:25 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] V. Conitzer, "On Stackelberg mixed strategies," *Synthese*, vol. 193, no. 3, pp. 689–703, 2016.

[2] V. Conitzer and T. Sandholm, "Computing the optimal strategy to commit to," in *Proceedings of the 7th ACM conference on Electronic commerce*, 2006, pp. 82–90.

[3] B. Von Stengel and S. Zamir, "Leadership games with convex strategy sets," *Games and Economic Behavior*, vol. 69, no. 2, pp. 446–457, 2010.

[4] D. Korzhyk, V. Conitzer, and R. Parr, "Solving Stackelberg games with uncertain observability." in *AAMAS*, 2011, pp. 1013–1020.

[5] J. Pita, M. Jain, M. Tambe, F. Ordónez, and S. Kraus, "Robust solutions to Stackelberg games: Addressing bounded rationality and limited observations in human cognition," *Artificial Intelligence*, vol. 174, no. 15, pp. 1142–1171, 2010.

[6] J. Karwowski, J. Mańdziuk, and A. Żychowski, "Sequential Stackelberg games with bounded rationality," *Applied Soft Computing*, vol. 132, p. 109846, 2023.

[7] Z. Yin, M. Jain, M. Tambe, and F. Ordonez, "Risk-averse strategies for security games with execution and observational uncertainty," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, 2011, pp. 758–763.

[8] F. Farokhi and H. Sandberg, "Fisher information as a measure of privacy: Preserving privacy of households with smart meters using batteries," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4726–4734, 2017.

[9] ——, "Ensuring privacy with constrained additive noise by minimizing Fisher information," *Automatica*, vol. 99, pp. 275–288, 2019.

[10] M. O. Karabag, M. Ornik, and U. Topcu, "Least inferable policies for Markov decision processes," in *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 1224–1231.

[11] C. Bodden, D. Rakita, B. Mutlu, and M. Gleicher, "A flexible optimization-based method for synthesizing intent-expressive robot arm motion," *The International Journal of Robotics Research*, vol. 37, no. 11, pp. 1376–1394, 2018.

[12] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 301–308.

[13] Y. Beck, I. Ljubić, and M. Schmidt, "A survey on bilevel optimization under uncertainty," *European Journal of Operational Research*, 2023.

[14] M. Patriksson and L. Wynter, "Stochastic nonlinear bilevel programming," in *Technical report. PRISM*. Citeseer, 1997.

[15] ——, "Stochastic mathematical programs with equilibrium constraints," *Operations research letters*, vol. 25, no. 4, pp. 159–167, 1999.

[16] G.-H. Lin and M. Fukushima, "Stochastic equilibrium problems and stochastic mathematical programs with equilibrium constraints: A survey," *Pacific Journal of Optimization*, vol. 6, no. 3, pp. 455–482, 2010.

[17] H. Hassani, A. Karbasi, A. Mokhtari, and Z. Shen, "Stochastic conditional gradient++," *arXiv preprint arXiv:1902.06992*, 2019.

[18] D. A. Braun and P. A. Ortega, "Information-theoretic bounded rationality and ε-optimality," *Entropy*, vol. 16, no. 8, pp. 4662–4676, 2014.

[19] R. D. McKelvey and T. R. Palfrey, "Quantal response equilibria for normal form games," *Games and economic behavior*, vol. 10, no. 1, pp. 6–38, 1995.

[20] M. H. Zare, O. A. Prokopyev, and D. Sauré, "On bilevel optimization with inexact follower," *Decision Analysis*, vol. 17, no. 1, pp. 74–95, 2020.

[21] C. R. Fox and R. T. Clemen, "Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior," *Management Science*, vol. 51, no. 9, pp. 1417–1432, 2005.

[22] B. Gao and L. Pavel, "On the properties of the softmax function with application in game theory and reinforcement learning," *CoRR*, vol. abs/1704.00805, 2017.

[23] A. Rubinstein, *Modeling bounded rationality*. MIT press, 1998.

[24] X. Di, X. Chen, and E. Talley, "Liability design for autonomous vehicles and human-driven vehicles: A hierarchical game-theoretic approach," *Transportation research part C: emerging technologies*, vol. 118, p. 102710, 2020.

[25] A. Millard-Ball, "Pedestrians, autonomous vehicles, and cities," *Journal of planning education and research*, vol. 38, no. 1, pp. 6–12, 2018.

Authorized licensed use limited to: University of Texas at Austin. Downloaded on January 20,2025 at 20:09:25 UTC from IEEE Xplore. Restrictions apply.