When Should a Leader Act Suboptimally? The Role of Inferability in Repeated Stackelberg Games

Mustafa O. Karabag[1], Sophia Smith[1], Negar Mehr[2], David Fridovich-Keil[1], and Ufuk Topcu[1]

***Abstract*— When interacting with other decision-making agents in non-adversarial scenarios, it is critical for an autonomous agent to have inferable behavior: The agent's actions must convey their intention and strategy. We model the inferability problem using Stackelberg games with observations where a leader and a follower repeatedly interact. During the interactions, the leader uses a fixed mixed strategy. The follower does not know the leader's strategy and dynamically reacts to the statistically inferred strategy based on the leader's previous actions. In the inference setting, the leader may have a lower performance compared to the setting where the follower has full information on the leader's strategy. We refer to the performance gap between these settings as the inferability gap. For a variety of game settings, we show that the inferability gap 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. We also analyze bimatrix Stackelberg games and identify a set of games where the leader's near-optimal strategy may potentially suffer from a large inferability gap.**

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 an autonomous agent are not fully known by the other agents, and the other agents rely on statistical inference from 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 full knowledge of the car's strategy. Consequently, acting in an inferable way is essential for autonomous agents.

We model the interaction between the autonomous agent and the other agent with a Stackelberg game. In this game, the autonomous agent is the *leader* that commits to a strategy (e.g., a software), and the other agent is the *follower* that reacts to the leader's strategy. 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, a version of the car's software defines a fixed strategy over actions, stopping and proceeding.

In a conventional Stackelberg game with mixed strategies, the follower does not know the leader's action but knows the leader's strategy from which the leader's actions are drawn [1]. We, on the other hand, consider that the follower does not have full information of the leader's strategy and relies on the observations from the previous interactions. More specifically, at every interaction, the follower reacts to 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.

The leader's optimal strategy may be mixed for a conventional Stackelberg game [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 strategy since such strategies would be more inferable. As such, the leader's expected return in the inference setting might be lower than its expected return in the full information setting. We call the return gap between these settings the leader's *inferability gap*.

We show that for a repeated static Stackelberg game with parametric action spaces, the leader's cumulative inferability gap is bounded under some Lipschitz continuity assumptions. As a corollary of this result, we show that for a repeated static bimatrix Stackelberg game, if the follower has bounded rationality (modeled by the maximum entropy response), the leader's cumulative inferability gap is bounded. We also show that the inferability gap is bounded for repeated static bimatrix Stackelberg games with fully rational followers and dynamic bimatrix Stackelberg games with myopic, boundedly rational followers. These upper bounds are functions of both the stochasticity level of the leader's strategy (i.e., the total variance of the leader's strategy assuming that the follower's estimator is efficient) and the number of interactions. As the stochasticity level of the leader's strategy decreases, the inferability gap vanishes. In the extreme case where the leader's strategy is deterministic, the leader does not suffer from any inferability gap after a single interaction; the expected return in the inference setting is the same as the expected return in the full information setting. The inferability gap at interaction k is at most $\mathcal{O}(1/\sqrt{k})$ (ignoring the other terms), implying that $\mathcal{O}(1/\epsilon^2)$ interactions are sufficient to achieve ϵ inferability gap.

Motivated by this 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 gap, and the strategies generated by the regularized objective function lead to improved transient returns compared to the strategies optimal for the full information setting. We also conduct a human subject study in a simulated driving environment where the participants interact with an autonomous car. Experiment results show that despite being suboptimal for the full information setting, strategies with lower stochasticity

[1]The University of Texas at Austin.
[2]University of California, Berkeley.
Correspondence to karabag@utexas.edu.

levels lead to improved return in interactions with humans.

Additionally, as a converse result, we provide an example bimatrix Stackelberg game where the inferability gap 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.

We also analyze a spectrum of bimatrix Stackelberg games and identify a set of games where the (near)-optimal strategy of the leader may suffer from a large inferability gap in the inference setting. More precisely, we show that there are strategies for the leader that limit the inferability gap if the game is almost cooperative or competitive, i.e., respectively, the difference or the sum of the leader's and follower's returns are bounded.

Related work: The closest work is the preliminary conference version [2] of this paper. Building on the results of the conference version, we provide analyses of the inferability gap in repeated static Stackelberg games with parametric policies, repeated static bimatrix Stackelberg games with fully rational followers, repeated static bimatrix Stackelberg games with classifying followers, and repeated dynamic discrete Stackelberg games with myopic boundedly rational followers. We show that for almost cooperative or competitive bimatrix Stackelberg games, in the inference setting, there exist strategies that are nearly as good as the optimal strategies in the full information setting. We also provide an additional example to emphasize the importance of inferable strategies in a semi-cooperative setting modeled as a Stackelberg game with parametric strategies.

Bimatrix Stackelberg games with a commitment to mixed strategies have been extensively studied in the literature under the assumption that the follower has full knowledge of the leader's strategy [1], [3], [4]. For these games, an optimal strategy for the leader can be computed in polynomial time via linear programming (assuming that the follower breaks ties in favor of the leader) [3]. The paper [5] 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 [6], [7] also consider this observation setting. To account for the follower's partial information, [6], [7], [8] consider a robust set that represents the possible realizations of the leader's strategy and maximizes 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 full information assumption.

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 [9], [10], [11] 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 [9], [10], [11]. 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 [12], [13]. These works are often concerned with creating a single trajectory that is distant from the trajectories under other intents. Different from the legible behavior literature that considers inferability during a single interaction, we consider statistical inferability over repeated interactions.

The leader's problem in our setting is a bilevel optimization problem under data uncertainty [14]. Works [15], [16], [17] 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 a decision with known data. In our problem, the data distribuiton depends on the leader's decision[1], whereas [15], [16], [17] consider a fixed distribution.

We represent the boundedly rational follower using the maximum entropy model (also known as Boltzmann rationality model or quantal response) [19], [20]. Alternatively, [6], [21] consider boundedly rational followers using the anchoring theory [22] or ϵ-optimal follower models.

II. Notation and Preliminaries on Stackelberg Games

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. $\mathcal{N}(\mu, s^2)$ denotes the normal distribution with mean μ and variance s^2. A function $f:\mathbb{R}^N \to \mathbb{R}^M$ is L-Lipschitz continuous if it satisfies $\|f(z)-f(q)\| \leq L\|z-q\|$ for any z and q. The matrix of ones with appropriate dimensions is denoted by J.

For $z, q \in \Delta^N$, 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 $z, q \in \mathbb{R}^N$ [23].

We define the *stochasticity level* of a discrete 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.

We also define a term $\varphi(z)$ from [24] to measure how concentrated z is. Let $C \subseteq \{1,\dots,N\}$, and $p(C)$ be the probability that a sampled element belongs to set C under distribution z. Define $p_z = \max_{C\subseteq\{1,\dots,N\}} \min(z(C), 1-z(C))$. Note that the maximum value of p_z is less than $1/2$. Also, note that the maximum value of p_z is less than one minus the maximum element of p_z. Define $\varphi(p_z) = \frac{1}{1-2p_z}\log\left(\frac{1-p_z}{p_z}\right)$. Note that the minimum value $\varphi(z)$ can take is 2, and $\varphi(z)$ is an decreasing function of p_z. As the strategy becomes more deterministic, $\varphi(z)$ increases.

Remark 1. *In the following sections, we present various bilevel optimization problems. For notational and analytical simplicity, we assume that there exists a unique solution to*

[1]For single-level stochastic optimization problems, this setting is referred to as non-oblivious stochastic optimization [18].

the inner optimization problem, i.e., the follower's problem, and the leader knows this unique response given its strategy.

We define a general dynamic Stackelberg game model and focus on special cases of this model in the later sections.

A. Dynamic Stackelberg Games with Parametric Mixed Strategies

A *dynamic Stackelberg game* is a two-player game between a *leader* and a *follower*. The game has the state space $\mathcal{S}$, the leader has action space $\mathcal{A}$, and the follower has action space $\mathcal{B}$. $f^{\mathrm{l}} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ is the leader's *utility function* and $f^{\mathrm{f}} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ is the follower's utility function. At time t, when the leader takes action a_t and the follower takes action b_t at state s_t, the leader and follower returns are $f^{\mathrm{l}}(s_t, a_t, b_t)$ and $f^{\mathrm{f}}(s_t, a_t, b_t)$ respectively. At time $t+1$, the game's state follows distribution $\mathcal{P}(s_t, a_t, b_t)$.

In a dynamic Stackelberg game with *parametric mixed strategies*, at state s, the leader and follower have *(stationary) mixed strategies* $\pi^{\mathrm{l}}(s; x)$ and $\pi^{\mathrm{f}}(s; y)$ parametrized by $x \in X$ and $y \in Y$, respectively. Here, $\pi^{\mathrm{l}}(s; x)$ and $\pi^{\mathrm{f}}(s; y)$ are probability distibutions over $\mathcal{A}$ and $\mathcal{B}$, respectively. For example, in a parametric game with the same Gaussian action distribution for every state s, one may have $\pi^{\mathrm{l}}(s; x) = \mathcal{N}(\theta, 1)$ where $X = \mathbb{R}$ and $\pi^{\mathrm{f}}(s; y) = \mathcal{N}(0, \phi)$ where $Y = \mathbb{R}_{\geq 0}$. When the follower's strategy is fixed, the game is a Markov decision process (MDP) for the leader, and vice versa.

Let $\boldsymbol{s}_t$, $\boldsymbol{a}_t$, and $\boldsymbol{b}_t$ denote the random versions of s_t, a_t, and b_t, respectively. At time t, the leader's *expected utility* is

$$\mathbb{E}_{\boldsymbol{a}_t \sim \pi^{\mathrm{l}}(s_t; x), \boldsymbol{b}_t \sim \pi^{\mathrm{f}}(s_t; y)} \left[f^{\mathrm{l}}(s_t, \boldsymbol{a}_t, \boldsymbol{b}_t) \right],$$

and the follower's expected utility is

$$\mathbb{E}_{\boldsymbol{a}_t \sim \pi^{\mathrm{l}}(s_t; x), \boldsymbol{b}_t \sim \pi^{\mathrm{f}}(s_t; y)} \left[f^{\mathrm{f}}(s_t, \boldsymbol{a}_t, \boldsymbol{b}_t) \right].$$

Let τ be the random, first hitting time to a return-free, invariant set $\mathcal{S}_{end}$ of states. The players reach $\mathcal{S}_{end}$ after τ steps and cannot leave this set under any sequence of actions. We consider that the interaction between the players ends upon reaching $\mathcal{S}_{end}$. For example, in a navigation scenario $\mathcal{S}_{end}$ could be the target locations. In a cyber system, $\mathcal{S}_{end}$ could represent the state where users are disconnected. The time τ (potentially) depends on the players' strategies. For example, in the navigation scenario, the time that it takes for an agent to reach its target may depend on its strategy and the other player's strategy. The leader's *expected cumulative utility* is

$$\mathbb{E}\left[\sum_{t=0}^{\tau} f^{\mathrm{l}}(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{b}_t) \right],$$

and the follower's expected cumulative utility is

$$\mathbb{E}\left[\sum_{t=0}^{\tau} f^{\mathrm{f}}(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{b}_t) \right],$$

where the expectations are over the randomness of states $\boldsymbol{s}_0 \dots \boldsymbol{s}_\tau$ and actions $\boldsymbol{a}_0 \dots \boldsymbol{a}_\tau$, $\boldsymbol{b}_0 \dots \boldsymbol{b}_\tau$.

In the full information setting, when deciding on its strategy parameter y, the follower knows the leader's parameter x. This means the follower knows the probability distribution of the leader's action for every state but does not know the leader's realized actions. The follower's goal is to maximize its expected return given the leader's strategy x, by solving:

$$\max_{y \in Y} \ \mathbb{E}\left[\sum_{t=0}^{\tau} f^{\mathrm{f}}(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{b}_t) \right]$$

where the expectation is over the randomness of states $\boldsymbol{s}_0 \dots \boldsymbol{s}_\tau$ and actions $\boldsymbol{a}_0 \dots \boldsymbol{a}_\tau$, $\boldsymbol{b}_0 \dots \boldsymbol{b}_\tau$

The leader's goal is to maximize its expected return, i.e., solve the bilevel optimization problem:

$$\begin{aligned} \sup_{x \in X} \ & \mathbb{E}\left[\sum_{t=0}^{\tau} f^{\mathrm{l}}(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{b}_t) \right] \\ \text{s.t.} \quad & y^* = \arg\max_{y \in Y} \mathbb{E}\left[\sum_{t=0}^{\tau} \mathbb{E}\left[f^{\mathrm{f}}(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{b}_t) \right] \right]. \end{aligned}$$

We define

$$\begin{aligned} SR(x) := \ & \mathbb{E}\left[\sum_{t=0}^{\tau} f^{\mathrm{l}}(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{b}_t) \right] \\ \text{s.t.} \quad & y^* = \arg\max_{y \in Y} \mathbb{E}\left[\sum_{t=0}^{\tau} \mathbb{E}\left[f^{\mathrm{f}}(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{b}_t) \right] \right]. \end{aligned}$$

We refer to $SR(x)$ as the Stackelberg return of strategy (parameter) x under full information and full follower rationality.

If random hitting time τ is deterministically 0, i.e., the game has a single state, the game is *static* with parametric mixed strategies. In this case, we drop s from our notation and use notation $\pi^{\mathrm{l}}(x)$, $\pi^{\mathrm{f}}(y)$, $f^{\mathrm{l}}(a, b)$, and $f^{\mathrm{f}}(a, b)$.

If the game has finite state and action spaces, the strategies are distributions over actions. In this case, we drop π and use notation $x(s)$ and $y(s)$ to denote the action distributions of the leader and the follower at state s, respectively.

B. Special Case: Static Bimatrix Stackelberg Games

In a *static bimatrix Stackelberg game*, 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 and the follower take actions i and j, the leader and follower returns are $f^{\mathrm{l}}(i, j) = A_{ij}$ and $f^{\mathrm{f}}(i, j) = B_{ij}$ respectively.

In bimatrix Stackelberg games with *mixed strategies*, the leader has a mixed strategy $x \in \Delta^m$, and the follower has mixed strategy $y \in \Delta^n$. We drop the notation π^{l} and π^{f} when considering bimatrix Stackelberg games as the parameters directly correspond to the action probabilities. Let $\boldsymbol{i}$ and $\boldsymbol{j}$ denote the random versions of i and j, respectively. The leader's *expected utility* is $x^\top A y = \mathbb{E}_{\boldsymbol{i} \sim x, \boldsymbol{j} \sim y}\left[A_{\boldsymbol{i},\boldsymbol{j}} \right]$, and the follower's expected utility is $x^\top B y = \mathbb{E}_{\boldsymbol{i} \sim x, \boldsymbol{j} \sim y}\left[B_{\boldsymbol{i},\boldsymbol{j}} \right]$. When deciding on strategy y, the follower knows the leader's strategy x. The follower knows the leader's action probabilities but does not know the leader's realized action.

1) Fully Rational Followers: The follower's goal is to maximize its expected return given x by solving:

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

The leader's goal is to maximize its expected return, i.e., solve the bilevel optimization problem:

$$\sup_{x\in\Delta^m} \quad x^\top A y^* \\ \text{s.t.} \quad y^* = \arg\max_{y\in\Delta^n} x^\top B y.$$

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

We define

$$SR(x) := x^\top A y^* \\ \text{s.t.} \quad y^* = \arg\max_{y\in\Delta^n} x^\top B y.$$

We refer to SR as the Stackelberg return under full information and full follower rationality. Note that $SR(x)$ is a piecewise linear function of x.

2) Boundedly Rational Followers with Maximum Entropy Response: 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 [25]. We consider the maximum entropy model (Boltzmann rationality) to represent boundedly rational followers [26].

Given the leader's strategy x, the 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 $\sigma_\lambda(B^\top x)$ [23]. 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 $y^* = \sigma_\lambda(B^\top x)$. In this setting, we define

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

as the Stackelberg return under full information and bounded follower rationality with maximum entropy response.

3) Boundedly Rational Followers with Maximum Likelihood Classificatrion: A classification function $c : \Delta^m \to \mathcal{T}$ maps the leader's strategy to discrete set $\mathcal{T} = \{x^1, \dots, x^{|\mathcal{T}|}\} \subset \Delta^m$ of predefined strategies. For example, a pedestrian considers drivers to be either aggressive or defensive. In detail, $c(x)$ is the maximum likelihood estimate given data x and the parameter space $\mathcal{T}$, i.e., $c(x) = \arg\max_{x^i\in\mathcal{T}} \Pr(x|x^i)$. A boundedly rational follower that classifies solves the following optimization problem

$$\max_{y\in\Delta^n} c(x)^\top B y$$

where $c(x)$ denotes the follower's classification of the leader's type. Given the follower's classification function c, the leader's goal is to maximize its expected utility, i.e., solve

$$\sup_{x\in\Delta^m} \quad x^\top A y^* \\ \text{s.t.} \quad y^* = \arg\max_{y\in\Delta^n} c(x)^\top B y.$$

In this setting, we define

$$SR(x) := x^\top A y^* \\ \text{s.t.} \quad y^* = \arg\max_{y\in\Delta^n} c(x)^\top B y,$$

as the Stackelberg return under full information and bounded follower rationality with maximum likelihood classification.

TABLE I: Overview of the problem settings and results

Achievability Results (Upper bounds on $SR(x) - IR_k(x)$)				
Problem section	Result section	Horizon	Strategy class	Follower type
III-A	IV-A	Static	Parametric	Rational
III-B.1	IV-C	Static	Discrete	Rational
III-B.2	IV-B	Static	Discrete	Max. Ent.
III-B.3	IV-D	Static	Discrete	Classifying
III-C	IV-E	Dynamic	Discrete	Myopic Max. Ent.
Converse Result (Lower bound on $SR(x) - IR_k(x)$)				
Problem section	Result section	Horizon	Strategy class	Follower type
III-B.1	IV-F	Static	Discrete	Rational
Achievability Result (Upper bound on $\max_x SR(x) - \max_x IR_k(x)$)				
Setting section	Result section	Horizon	Strategy class	Follower type
II-B.1	IV-G	Static	Discrete	Rational

III. Problem Formulation: Repeated Stackelberg Games With Inference

In this section, we formulate decision-making problems for the leader where the follower statistically infers the strategy of the leader through repeated interactions. Let $IR_k(x)$ be the expected return of the leader at interaction k in the inference setting. $SR(x)$ is the expected return of the leader in the full information setting. Since $IR_k(x)$ depends on the statistical inference of x by the follower, it is not necessarily the same with $SR(x)$. We refer the difference $SR(x) - IR_k(x)$ as the *inferability gap* of the leader at interaction k. We are interested in analyzing the inferability gap to find strategies that remain performant in the inference setting.

In detail, the leader's strategy affects the follower's optimal strategy in two ways in Stackelberg games with inference. First, as in the conventional Stackelberg game formulation, the leader's strategy determines the expected return for different follower actions, i.e., $x^\top B$ in the bimatrix Stackelberg games. This affects the optimal strategies for the follower. Second, unlike in the full information Stackelberg game, the leader's strategy x modifies the distribution of its empirical action distribution $\hat{\boldsymbol{x}}_k$ and, consequently, the follower's strategy $\boldsymbol{y}_k$.

A strategy with a high Stackelberg return under full information may be highly suboptimal in a Stackelberg game with inference. Different realizations of $\hat{\boldsymbol{x}}_k$ lead to different solutions for $\boldsymbol{y}_k$. If x is poorly inferred by the follower, the follower's strategy $\boldsymbol{y}_k$ 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 full information and efficiently conveying information about itself to the follower.

We analyze inferability gap for different problem settings. We first consider a static game with parametric action spaces. Then, we consider a static bimatrix game over discrete action spaces with fully rational and boundedly rational followers. Finally, we consider a dynamic game over discrete state and action spaces. Table I summarizes of the settings and results.

Remark 2. *In some settings, the leader's strategy could be deterministic in a higher-dimensional state space. However, the follower may perceive the strategy as mixed in a lower-dimensional state space. For example, the car's decision may be a deterministic function of the car's exact distance to the pedestrian. However, a pedestrian with imperfect information about the car's state, i.e., the exact distance, may think the car has a mixed strategy. While we do not explicitly consider games with imperfect information, the problems and results can easily be extended to these settings.*

As mentioned in Remark 1, for notational and analytical convenience, in the remaining sections, we assume that the follower's optimal response exists and is unique. We note that if the follower's response is not unique, one can formulate the leader's problem by considering the worst-case outcome as given in Definitions 3.26 and 3.27 of [27], i.e., among the strategies that maximize its objective function, the follower chooses a strategy that minimizes the leader's return (under full information or inference). The problem formulations and the results derived in the remaining sections easily extend to the worst-case formulation. We also note that for the followers with strongly convex utility functions, the optimal response is unique even without the assumption. For example, the follower's response is unique for the boundedly rational followers with maximum entropy response.

A. Repeated Static Parametric Stackelberg Games with Inference Against Fully Rational Followers

Consider a Stackelberg game with mixed strategies that is repeated K times. The follower knows the leader's parametric strategy space but does not know the leader's fixed mixed strategy (and its associated parameter) $\pi^{\mathrm{l}}(x)$. Instead, the follower infers the leader's parameter x from observations of the previous interactions. At interaction k, let $\hat{x}_k$ be the estimator of the leader's parameter by the follower based on the leader's previous actions $a_1, \dots, a_{k-1}$.

Under these assumptions, the follower's strategy $\pi^{\mathrm{f}}(y_k)$ at interaction k depends on the leader's actions in the previous $k-1$ interactions. For this reason, $\pi^{\mathrm{f}}(y_k)$ changes with k. For a fully rational follower based on the estimate $\hat{x}_k$, we denote the follower's optimal parameter with y_k^* such that $\pi^{\mathrm{f}}(y_k^*) = \arg\max_{y \in Y} f^{\mathrm{f}}(\pi^{\mathrm{l}}(\hat{x}_k), \pi^{\mathrm{f}}(y))$.

The leader's decision-making problem is to a priori select a strategy $\pi^{\mathrm{l}}(x^*)$ that maximizes its expected cumulative return under inference, i.e., assuming that the follower rationally responds to the estimator $\pi^{\mathrm{l}}(\hat{x}_k)$ of $\pi^{\mathrm{l}}(x^*)$ at each interaction k. Let $\boldsymbol{a}_k$, $\hat{\boldsymbol{x}}_k$, and $\boldsymbol{y}_k$ be random variables denoting the unrealized versions of a_k, $\hat{x}_k$ and y_k, respectively. If exists, the leader's optimal strategy $x \in X$ maximizes

$$\mathbb{E}\left[\sum_{k=1}^{K} f^{\mathrm{l}}(\boldsymbol{a}_k, \boldsymbol{b}_k)\right]$$
$$\text{s.t. } \boldsymbol{y}_k^* = \arg\max_{\pi^{\mathrm{f}} \in Y} \mathbb{E}_{\boldsymbol{a} \sim \pi^{\mathrm{l}}(\hat{\boldsymbol{x}}_k), \boldsymbol{b} \sim \pi^{\mathrm{f}}(y)} \left[f^{\mathrm{f}}(\boldsymbol{a}, \boldsymbol{b})\right].$$

Here, $\boldsymbol{a}_k \sim \pi^{\mathrm{l}}(x), \boldsymbol{b}_k \sim \pi^{\mathrm{l}}(\boldsymbol{y}_k^*)$, and the randomness of $\boldsymbol{y}_k^*$ is over the leader's actions $\boldsymbol{a}_1, \dots, \boldsymbol{a}_{k-1}$. The leader solves this decision problem prior to taking any action, meaning their future actions $\boldsymbol{a}_1, \dots, \boldsymbol{a}_K$ are random variables that are all independently sampled from x^*. Since the follower's estimation $\hat{\boldsymbol{x}}_k$ is a function of these future actions and the follower's strategy $\boldsymbol{y}_k^*$ is a function of $\hat{\boldsymbol{x}}_k$, they are both random variables as well and therefore bolded.

B. Repeated Static Bimatrix Stackelberg Games with Inference Against Fully or Boundedly Rational Followers

Consider a bimatrix Stackelberg game with mixed strategies that is repeated K times. The follower infers the leader's strategy from observations of the previous interactions. For example, at interaction k, let $\hat{x}_k$ be the *plug-in sample mean estimation* of the leader's strategy. Specifically, if the leader takes actions $i_1, \dots, i_{k-1}$ at the previous $k-1$ interactions,

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

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.

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

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

1) 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 of x^* at each interaction k. Let $\boldsymbol{i}_k$, $\hat{\boldsymbol{x}}_k$, and $\boldsymbol{y}_k^*$ be random variables denoting the unrealized versions of i_k, $\hat{x}_k$ and y_k^*, respectively. The leader's decision problem is to maximize

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

The expectation is over the randomness in the leader's (random) actions $\boldsymbol{i}_1, \dots, \boldsymbol{i}_K$.

2) Boundedly rational follower with maximum entropy response: The leader's decision problem is to maximize

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

3) Boundedly rational follower with maximum likelihood classification: The leader's decision problem is to maximize

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

C. Repeated Discrete Dynamic Parametric Stackelberg Games with Inference Against Boundedly Rational, Myopic Followers

Assume that the state space $\mathcal{S}$, the leader's action space $\mathcal{A}$, and the follower's action space $\mathcal{B}$ are finite. After every interaction $s_0 a_0 b_0 \dots s_\tau a_\tau b_\tau$, the follower updates its estimation of the leader's strategy using the plug-in sample mean estimator for every state using all previous sample actions. At interaction k, we denote the follower's estimation of the leader's strategy for state s with $\hat{x}_k(s)$. At interaction k, a *boundedly rational, miyopic* follower approximately maximizes its one-step return. For every state s, we define utility matrices $A(s)_{ij} := f^{\mathrm{l}}(s,i,j) =$ and $B(s)_{ij} := f^{\mathrm{f}}(s,i,j)$. The leader's decision problem is to maximize

$$\mathbb{E}\left[\sum_{k=1}^{K}\sum_{t=0}^{\tau} x(\boldsymbol{s}_t)^\top A(\boldsymbol{s}_t)\boldsymbol{y}_k^*(\boldsymbol{s}_t)\right]$$
$$\text{s.t. } \boldsymbol{y}_k^*(\boldsymbol{s}_t) = \sigma_\lambda(B(\boldsymbol{s}_t)^\top \hat{\boldsymbol{x}}_k(\boldsymbol{s}_t)).$$

We note that different from the static case, the leader aims to optimize its expected return for the whole interaction.

D. An Example Where Inferable Behavior is Necessary: Autonomous Car and Pedestrian Interaction

This section describes a motivating example of the interaction between an autonomous car and a fully-rational pedestrian. The interaction is modeled as a bimatrix game. Similar scenarios have been considered in [28], [29].

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 II: Utilities for the 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 $r^{\mathrm{f}}_{\mathrm{cr}} = 2$, and potentially getting fined for jaywalking has $r^{\mathrm{f}}_{\mathrm{jw}} = -1$ value. For the car, proceeding without a stop has a value $r^{\mathrm{l}}_{\mathrm{pr}} = 2$, and making an emergency stop has $r^{\mathrm{l}}_{\mathrm{em}} = -8$ value.

Scenario 1: The car stops, and the pedestrian waits for the car. The pedestrian's return is $r^{\mathrm{f}}_{\mathrm{cr}} = 2$ 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 $r^{\mathrm{f}}_{\mathrm{cr}} + r^{\mathrm{f}}_{\mathrm{jw}} = 1$ 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 $r^{\mathrm{l}}_{\mathrm{pr}} = 2$ 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 $r^{\mathrm{f}}_{\mathrm{cr}} + r^{\mathrm{f}}_{\mathrm{jw}} = 1$. While the car proceeds, it makes an emergency stop due to the crossing pedestrian, resulting in a return of $r^{\mathrm{l}}_{\mathrm{em}} = -8$.

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 full 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 choose 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 full 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 gap*. 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 full information case. The difference 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.

A similar inferability gap happens for the followers that classify. Consider a pedestrian that classifies cars into two types: cars that stop with probability 0.99 and cars that proceed with probability 0.99. For these classes, the optimal action for the pedestrian is to cross and wait, respectively. For such a classification, in the full information setting, it is still optimal for the car to choose a p such that $p > 0.5$ and $p \approx 0.5$. However, under such a strategy, the follower's classification of the leader's strategy alternates frequently. Similar to the fully rational setting, the leader's expected return under inference is ≈ -1.5 in the second interaction.

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 full information in different settings. With an abuse of notation, for different settings, we use $IR_k(x)$ to denote the leader's expected return at interaction k under inference and $SR(x)$ to denote the leader's expected return under full information under strategy (parameter) x. 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 $k = 2, \dots, K$. Let $v^{\text{Stck}}(x)$ be the leader's expected cumulative return in the full information setting and $v^{\text{Infr}}(x)$ be the leader's expected cumulative return in the inference setting for interactions $k = 2, \dots, K$. Due to the linearity of expectation, the expected cumulative return can be represented as a sum of expected returns of every interaction, i.e., $v^{\text{Stck}}(x) = (K-1)SR(x)$ and $v^{\text{Infr}}(x) = \sum_{k=2}^{K} IR_k(x)$. Consequently, we are interested in analyzing the cumulative gap $(K-1)SR(x) - \sum_{k=2}^{K} IR_k(x)$.

For the bimatrix games, without loss of generality, we assume that the utility matrices are normalized, i.e., $\max_{i,j} B_{i,j} - \min_{i,j} B_{i,j} = \max_{i,j} A_{i,j} - \min_{i,j} A_{i,j} = 1$.

Summary of the technical results: In Sections IV-A-IV-E, we provide lower bounds on the leader's return under inference $IR_k(x)$ for various settings depending on the action spaces, the horizon length, and the follower type. The bounds share two elements other than the follower's return under full information $SR(x)$: (i) the interaction number[2] and (ii) the stochasticity (concentrability) of the leader's strategy.

Dependency on the interaction number: At interaction $k+1$, the inferability gap scales with $1/\sqrt{k}$ meaning that it decreases with each interaction, and the cumulative inferability gap grows as $\sqrt{\mathcal{O}(K)}$.

Dependency on the stochasticity (or concentrability) of the leader's strategy: As the leader's action distribution becomes less stochastic, the upper bound on the inferability loss decreases, implying that the return under inference gets closer to the return under full information. The bounds are asymptotically vanishing: as the leader's strategy becomes deterministic, the inferability gap approaches zero. We remark that while the inferability gap approaches zero, deterministic strategies do not necessarily maximize the return under inference.

For repeated games where the leader's return may depend on the inferability of its strategy, the strategy maximizing the return under full information is not necessarily optimal if the number of interactions is limited. In these scenarios, one can regulate the stochasticity level of the strategy to improve the return under inference. While such regulated strategies are, in general, suboptimal in the limit (when the follower has full information of the leader's strategy), they may maximize the cumulative return under inference in a limited number of interactions. We demonstrate the effects of regulating the total variance through a numerical example in Section V-C.

[2]Theorem 1 does not directly depend on the interaction number. However, the MSE of the follower's estimate inherently depends on the interaction number. Consider that the follower has an efficient estimator of the leader's parameter. In this case, the MSE of the follower's estimate is inversely proportional to the number of interactions since the leader's actions are i.i.d.. Consequently, one can obtain a bound that grows as $\mathcal{O}(\sqrt{K})$ and depends on the stochasticity level of the leader's action distribution.

A. Achievability Bound for Repeated Parametric Static Stackelberg Games Against Fully Rational Followers

Let $\hat{\boldsymbol{x}}_k$ be the random variable denoting follower's estimation of the leader's parameter x after $k-1$ interactions based on the leader's previous actions $\boldsymbol{a}_1, \dots, \boldsymbol{a}_{k-1}$.

In this setting, at interaction k, we have

$$IR_k(x) = \mathbb{E}_{\boldsymbol{a}_k \sim \pi^{\text{l}}(x), \boldsymbol{b}_k \sim \pi^{\text{f}}(\boldsymbol{y}_k^*)} \left[f^{\text{l}}(\boldsymbol{a}_k, \boldsymbol{b}_k) \right] \quad \text{s.t. } \boldsymbol{y}_k^* = \arg\max_{y \in Y} \mathbb{E}_{\boldsymbol{a} \sim \pi^{\text{l}}(\hat{\boldsymbol{x}}_k), \boldsymbol{b} \sim \pi^{\text{f}}(y)} \left[f^{\text{f}}(\boldsymbol{a}, \boldsymbol{b}) \right],$$

and

$$SR(x) = \mathbb{E}_{\boldsymbol{a}_k \sim \pi^{\text{l}}(x), \boldsymbol{b}_k \sim \pi^{\text{f}}(y^*)} \left[f^{\text{l}}(\boldsymbol{a}_k, \boldsymbol{b}_k) \right] \quad \text{s.t. } y^* = \arg\max_{y \in Y} \mathbb{E}_{\boldsymbol{a} \sim \pi^{\text{l}}(x), \boldsymbol{b} \sim \pi^{\text{f}}(y)} \left[f^{\text{f}}(\boldsymbol{a}, \boldsymbol{b}) \right].$$

If $\hat{\boldsymbol{x}}_k$ converges to x and the follower's optimal response is a continuous mapping, then the follower's response under inference converges to the follower's response under full information. Furthermore, if the leader's objective function is a continuous mapping, then the leader's expected return under inference $IR_k(x)$ converges to the leader's return under full information $SR(x)$. However, with a finite number of interactions, these returns are not necessarily the same, and the leader may suffer from an *inferability gap*.

The following result shows that the inferability gap is upper bounded if the follower's estimation of the leader's parameter has a bounded mean squared error (MSE), the follower's response is Lipschitz continuous, and the leader's objective function is Lipschitz continuous.

Theorem 1. *Let $MSE_k \geq 0$ be constants for $k = 2, 3, \dots, K$. For a repeated parametric static Stackelberg game, if*

1) $\mathbb{E}[\|\hat{\boldsymbol{x}}_k - x\|^2] \leq MSE_k$,
2) $\mathbb{E}_{\boldsymbol{a} \sim \pi^{l}(x), \boldsymbol{b} \sim \pi^{f}(y)} \left[f^{l}(\boldsymbol{a}, \boldsymbol{b}) \right]$ *is an L^l-Lipschitz function of y, and*
3) $\arg\max_{y \in Y} \mathbb{E}_{\boldsymbol{a} \sim \pi^{l}(x'), \boldsymbol{b} \sim \pi^{f}(y)} \left[f^{f}(\boldsymbol{a}, \boldsymbol{b}) \right]$ *is an L^f-Lipschitz function of x' for all $x' \in X$,*

then $(K-1)SR(x) - \sum_{k=2}^{K} IR_k(x) \leq \sum_{k=2}^{K} L^l L^f \sqrt{MSE_k}$.

Proof of Theorem 1. Condition 3 implies

$$\|y^* - \boldsymbol{y}_k^*\| \leq L^{\text{f}} \|x - \hat{\boldsymbol{x}}_k\|.$$

Combining this inequality with condition 2, we get

$$\left| \mathbb{E}_{\boldsymbol{a} \sim \pi^{\text{l}}(x), \boldsymbol{b} \sim \pi^{\text{l}}(y)} \left[f^{\text{l}}(\boldsymbol{a}, \boldsymbol{b}) \right] - \mathbb{E}_{\boldsymbol{a} \sim \pi^{\text{l}}(x), \boldsymbol{b} \sim \pi^{\text{l}}(\boldsymbol{y}_k^*)} \left[f^{\text{l}}(\boldsymbol{a}, \boldsymbol{b}) \right] \right| \leq L^{\text{l}} L^{\text{f}} \|x - \hat{\boldsymbol{x}}_k\|$$

which implies

$$\mathbb{E}_{\boldsymbol{a} \sim \pi^{\text{l}}(x), \boldsymbol{b} \sim \pi^{\text{l}}(\boldsymbol{y}_k^*)} \left[f^{\text{l}}(\boldsymbol{a}, \boldsymbol{b}) \right] \geq \mathbb{E}_{\boldsymbol{a} \sim \pi^{\text{l}}(x), \boldsymbol{b} \sim \pi^{\text{l}}(y)} \left[f^{\text{l}}(\boldsymbol{a}, \boldsymbol{b}) \right] - L^{\text{l}} L^{\text{f}} \|x - \hat{\boldsymbol{x}}_k\|.$$

Let $h(t)$ be the p.d.f. of $\|\hat{\boldsymbol{x}}_k - x\|$. We have

$$IR_k(x) = \mathbb{E}\left[\mathbb{E}_{\boldsymbol{a}\sim\pi^{\mathrm{l}}(x),\boldsymbol{b}\sim\pi^{\mathrm{l}}(\boldsymbol{y}_k^*)}\left[f^{\mathrm{l}}(\boldsymbol{a},\boldsymbol{b})\right]\right] \tag{1a}$$

$$= \int_0^\infty \mathbb{E}\left[\mathbb{E}_{\boldsymbol{a}\sim\pi^{\mathrm{l}}(x),\boldsymbol{b}\sim\pi^{\mathrm{l}}(\boldsymbol{y}_k^*)}\left[f^{\mathrm{l}}(\boldsymbol{a},\boldsymbol{b})\right] \middle| \|\hat{\boldsymbol{x}}_k - x\| = t\right] h(t)dt \tag{1b}$$

$$\geq \int_0^\infty \left(\mathbb{E}_{\boldsymbol{a}\sim\pi^{\mathrm{l}}(x),\boldsymbol{b}\sim\pi^{\mathrm{l}}(y)}\left[f^{\mathrm{l}}(\boldsymbol{a},\boldsymbol{b})\right] - L^{\mathrm{l}}L^{\mathrm{f}}\|x - \hat{\boldsymbol{x}}_k\|\right) h(t)dt \tag{1c}$$

$$= SR(x) - L^{\mathrm{l}}L^{\mathrm{f}}\int_0^\infty t h(t)dt \tag{1d}$$

where (1b) is due to the definition of expectation, (1c) is due to the above bound, and (1d) is due to the definition of SR and $\int_{t=0}^\infty h(t) = 1$. Note that

$$\mathbb{E}[\|\hat{\boldsymbol{x}}_k - x\|^2] = \int_0^\infty t^2 h(t)dt \geq \left(\int_0^\infty t h(t)dt\right)^2 \tag{2}$$

due to Jensen's inequality. Condition 1 in (1d) and (2) imply

$$\begin{aligned} IR_k(x) &\geq SR(x) - L^{\mathrm{l}}L^{\mathrm{f}}\sqrt{\left(\int_0^\infty t^2 h(t)dt\right)} \\ &= SR(x) - L^{\mathrm{l}}L^{\mathrm{f}}\sqrt{MSE_k} \end{aligned}$$

Summation from $k = 2$ to K yields the desired result. □

B. Achievability Bound for Repeated Bimatrix Stackelberg Games Against Boundedly Rational Followers with Maximum Entropy Response

In this setting, we define

$$IR_k(x) := \mathbb{E}\left[x^\top A \boldsymbol{y}_k\right] \quad \text{s.t. } \boldsymbol{y}_k = \sigma_\lambda(B^\top \hat{\boldsymbol{x}}_k),$$

and

$$SR(x) := x^\top A y^* \text{ s.t. } y^* = \sigma_\lambda(B^\top x).$$

The inferability gap in this setting can be bounded directly using Theorem 1 due to three facts: (i) Since the leader's actions $i_1, \ldots, i_{k-1}$ are i.i.d., the plug-in estimator satisfies $\mathbb{E}\left[\|\hat{\boldsymbol{x}}_{\boldsymbol{k}} - x\|^2\right] = {}^{\nu(x)^2}/_{k-1}$. (ii) The leader's expected return is a ${}^{\sqrt{n}}/_2$-Lipschitz continuous function of the follower's strategy. (iii) The follower's maximum entropy optimal response is a ${}^{\lambda\sqrt{nm}}/_2$-Lipschitz continuous function of the leader's empirical action distribution.

Corollary 1. *For a repeated bimatrix Stackelberg games with a boundedly rational follower with maximum entropy response,*

$$(K-1)SR(x) - \sum_{k=2}^K IR_k(x) \leq \sum_{k=2}^K \frac{\lambda n\sqrt{m}\nu(x)}{4\sqrt{(k-1)}}.$$

We use the following lemmas to prove the corollary. We give the proofs to the lemmas in the appendix.

Lemma 1.

$$\mathbb{E}[\|\hat{\boldsymbol{x}}_k - x\|^2] = \frac{\nu(x)^2}{k-1}.$$

Lemma 2. *Let $y_k = \sigma_\lambda(B^\top \hat{x}_k)$ and $y = \sigma_\lambda(B^\top x)$.*

$$|x^\top A y - x^\top A y_k| \leq \frac{\sqrt{n}}{2}\|y - y_k\|.$$

Lemma 3.

$$\|\sigma_\lambda(B^\top x) - \sigma_\lambda(B^\top \hat{x}_k)\| \leq \frac{\lambda\sqrt{nm}}{2}\|\hat{x}_k - x\|. \tag{3}$$

Proof of Corollary 1. The result directly follows from Theorem 1 since Lemmas 1, 2, and 3, satisfy conditions 1, 2, and 3 in Theorem 1, respectively. □

Remark 3. *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 gap grows sublinearly, i.e,

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

We note that the MSE of the follower's estimator depends on the stochasticity level of the leader's strategy. As the leader's strategy becomes deterministic, i.e., $\nu(x) \to 0$, the inferability gap 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 gap. As the follower becomes irrational, i.e., $\lambda \to 0$, the inferability gap vanishes to 0, and when the follower is fully irrational, $\lambda = 0$, the leader does not suffer from an inferability gap 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 full information and having a minimal inferability gap, i.e., efficiently conveying information about itself to the follower.

C. Achievability Bound for Repeated Bimatrix Stackelberg Games against Fully Rational Followers

In this setting, we define

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

and

$$SR(x) := x^\top A y^* \qquad \text{s.t.} \quad y^* = \arg\max_{y\in\Delta^n} x^\top B y.$$

Unlike the boundedly rational followers, a fully rational follower's optimal strategy is not a continuous function of the leader's strategy. For example, the pedestrian's optimal strategy as a function of the leader's strategy has a discontinuity in the motivating example. Hence, Theorem 1 is not directly applicable to this setting.

While the follower's optimal strategy is not a continuous function of the leader's strategy, it is a piecewise constant function of it. In words, the follower's optimal strategy is the same within subdomains of the leader's strategy simplex. Let e_i denote a probability vector such that the i-th entry is 1 and the others are 0. Strategy profile e_i is optimal if and only if $x^\top A e_i \geq x^\top A e_j$ for all $j \in [n]$. Let $C_i = \{x \mid \forall j \in [n], x^\top A e_i \geq x^\top A e_j\}$, i.e., the set of leader strategies such that strategy profile e_i is optimal for the follower.

In the inference setting, if the leader has a strategy profile x such that e_i is uniquely optimal, i.e., $x \in C_i$ and $x \notin C_j$

for every $j \in [n] \setminus \{i\}$, then $\boldsymbol{y}_k^*$ almost surely converges to e_i as $\hat{\boldsymbol{x}}_k$ falls in only C_i with probability 1.

Inspired by the piecewise constant property and the convergence of $\hat{\boldsymbol{x}}_k$ to x, we show that the inferability gap is bounded for a finite number of actions. We first note that the boundary between C_i and C_j is a hyperplane for every $i \neq j \in [n]$. Let e_i be uniquely optimal for x and $d(x)$ denote the L2 distance to the closest boundary, i.e., $\min_{j\in[n]\setminus\{i\}} \min_{x'\in C_j} \|x - x'\|$. We note that $d(x)$ has an analytical expression. We have $\mathbb{E}\left[\|\hat{\boldsymbol{x}}_{\boldsymbol{k}} - x\|^2\right] = {}^{\nu(x)^2}/_{k-1}$, which implies $\Pr(\|\hat{\boldsymbol{x}}_{\boldsymbol{k}} - x\| \leq d(x)) = {}^{\nu(x)}/_{\sqrt{k-1}d(x)}$ due to the Markov bound. Using this, we get the following bound on the inferability gap.

Theorem 2. *For a repeated bimatrix Stackelberg game with a fully rational follower, we have*

$$(K-1)SR(x) - \sum_{k=2}^{K} IR_k(x) \leq \sum_{k=2}^{K} \frac{\nu(x)}{d(x)\sqrt{(k-1)}}.$$

Proof of Theorem 2. Let i be the optimal action for the follower given the leader's strategy x. We have

$$\begin{aligned} IR_k(x) = (1 - \Pr(\hat{\boldsymbol{x}}_k \notin C_i))\mathbb{E}\left[x^\top A \boldsymbol{y}_k^* | \hat{\boldsymbol{x}}_k \in C_i\right] \\ + \Pr(\hat{\boldsymbol{x}}_k \notin C_i)\mathbb{E}\left[x^\top A \boldsymbol{y}_k^* | \hat{\boldsymbol{x}}_k \notin C_i\right] \end{aligned}$$

We note that $\hat{\boldsymbol{x}}_k \in C_i$ implies that $\boldsymbol{y}_k^* = e_i = y$ and hence $\mathbb{E}\left[x^\top A \boldsymbol{y}_k^* | \hat{\boldsymbol{x}}_k \in C_i\right] = x^\top A y = SR(x)$. We have

$$\begin{aligned} &SR(x) - IR_k(x) \\ &= \Pr(\hat{\boldsymbol{x}}_k \notin C_i)(\mathbb{E}\left[x^\top A \boldsymbol{y}_k^* | \hat{\boldsymbol{x}}_k \in C_i\right] - \mathbb{E}\left[x^\top A \boldsymbol{y}_k^* | \hat{\boldsymbol{x}}_k \notin C_i\right]) \end{aligned}$$

Note that

$$\begin{aligned} &\max_{x\in\Delta^m} \mathbb{E}\left[x^\top A \boldsymbol{y}_k^* | \hat{\boldsymbol{x}}_k \in C_i\right] - \mathbb{E}\left[x^\top A \boldsymbol{y}_k^* | \hat{\boldsymbol{x}}_k \notin C_i\right] \\ &\leq \max_{i,j} A_{ij} - \min_{i,j} A_{i,j} = 1. \end{aligned}$$

Therefore, $SR(x) - IR_k(x) \leq \Pr(\hat{\boldsymbol{x}}_k \notin C_i)$. The event $\hat{\boldsymbol{x}}_k \notin C_i$ happens only if $\|\hat{\boldsymbol{x}}_{\boldsymbol{k}} - x\| \geq d(x)$ which happens with prob. at most $\frac{\mathbb{E}[\|\hat{\boldsymbol{x}}_{\boldsymbol{k}} - x\|]}{d(x)}$ due to the Markov bound. We have

$$SR(x) - IR_k(x) \leq \frac{\mathbb{E}\left[\|\hat{\boldsymbol{x}}_{\boldsymbol{k}} - x\|\right]}{d(x)}.$$

Due to Jensen's inequality we get

$$SR(x) - IR_k(x) \leq \frac{\mathbb{E}\left[\|\hat{\boldsymbol{x}}_{\boldsymbol{k}} - x\|\right]}{d(x)} \leq \frac{\sqrt{\mathbb{E}\left[\|\hat{\boldsymbol{x}}_{\boldsymbol{k}} - x\|^2\right]}}{d(x)}.$$

Using Lemma 1 and summation from $k = 2$ to K yields the desired result. □

We note that we can derive an asymptotically tighter bound using the Chernoff bound for the convergence of $\hat{x}_k$ to x instead of the Markov bound. Using $\|\hat{\boldsymbol{x}}_{\boldsymbol{k}} - x\| \leq \|\hat{\boldsymbol{x}}_{\boldsymbol{k}} - x\|_1$ and Theorem 2.1 of [24], we can get the result

$$\sum_{k=2}^{K} (SR(x) - IR_k(x)) \leq \sum_{k=2}^{K} 2^m \exp\left(-\frac{(k-1)\varphi(x)d(x)^2}{4}\right),$$

which implies the cumulative inferability gap is bounded by a constant depending on m, $\varphi(x)$, and $d(x)$ regardless of K.

D. Achievability Bound for Repeated Bimatrix Stackelberg Games Against Classifying Boundedly Rational Followers

In this setting, we define

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

and

$$SR(x) := x^\top A y^* \qquad \text{s.t.} \quad y^* = \arg\max_{y\in\Delta^n} c(x)^\top B y$$

where c is the maximum likelihood classification function of the follower.

The classification function c assigns data x to $x^i \in \mathcal{T}$ if $\Pr(x|x^i) > \Pr(x|x^j)$ for all $j \in [l] \setminus \{i\}$, i.e., the likelihood of data x is maximized under x^i among strategies in $\mathcal{T} = \{x^1, \ldots, x^l\}$. We note that $\log \Pr(x|x^i) = \langle x, \log(x^i)\rangle$, and $\Pr(x|x^i) > \Pr(x|x^j)$ implies $\log \Pr(x|x^i) - \log \Pr(x|x^j) > 0$, i.e., $\langle x, \log(x^i) - \log(x^j)\rangle \geq 0$. In words, similar to the fully rational setting, the classification boundaries are hyperplanes, and the optimal response of the follower is a piecewise linear function of the leader's strategy. Let $C_i = \{x \mid \forall j \in [n], x^\top \log(x^i) \geq x^\top \log(x^j)\}$, i.e., the set of leader's strategies that is classified into x^i.

In the inference setting, if the leader has a strategy profile x such that $\arg\max_{x^i\in\mathcal{T}} \Pr(x|x^i)$ is unique, then $\boldsymbol{y}_k^*$ almost surely converges to $y^* = \arg\max_{y\in\Delta^n} c(x)^\top B y$ as $\hat{\boldsymbol{x}}_k$ falls in only C_i with probability 1. Let $d(x)$ denote the L2 distance to the closest boundary, i.e., $\min_{j\in[l]\setminus\{i\}} \min_{x'\in C_j} \|x - x'\|$. Theorem 2 applies to this case with the new definition of $d(x)$.

E. Achievability Bound for a Repeated Discrete Dynamic Stackelberg Games Against Myopic Boundedly Rational Followers with Maximum Entropy Response

In this setting, we define

$$\begin{aligned} IR_k(x) :=\ & \mathbb{E}\left[\sum_{t=0}^{\tau} f^1(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{b}_t) \middle| b_t \sim \boldsymbol{y}_k^*(\boldsymbol{s}_t)\right] \\ \text{s.t. } & \boldsymbol{y}_k^*(\boldsymbol{s}_t) = \sigma_\lambda(B(\boldsymbol{s}_t)^\top \hat{\boldsymbol{x}}_k(\boldsymbol{s}_t)) \end{aligned}$$

and

$$\begin{aligned} SR(x) :=\ & \mathbb{E}\left[\sum_{t=0}^{\tau} f^1(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{b}_t) \middle| b_t \sim y^*(\boldsymbol{s}_t)\right] \\ \text{s.t. } & y^*(\boldsymbol{s}_t) = \sigma_\lambda(B(\boldsymbol{s}_t)^\top x(\boldsymbol{s}_t)). \end{aligned}$$

To show the closeness of $IR_k(x)$ and $SR(x)$, we follow a similar approach model-based off-policy evaluation for MDPs [30]. We note that for a myopic follower, the dynamic game is an MDP for the leader. With the increasing number of sample actions from a state s, the follower's estimate $\hat{\boldsymbol{x}}_k(s)$ converges to the leader's strategy $x(s)$. Since the follower has the maximum entropy response and is myopic, the follower's response under inference $\boldsymbol{y}_k^*(s)$ converges to the follower's response under full information $y^*(s)$. Using the closeness of $\boldsymbol{y}_k^*(s)$ and $y^*(s)$, we can show the the closeness of $IR_k(x)$ and $SR(x)$ if the game has the contraction property. We use

a series of lemmas to prove this result and give the proof sketches in the appendix.

To ensure the boundedness of $IR(x)$ and $SR(x)$, we make the following contraction assumption.

Assumption 1. *For all y, $\sum_{q\in\mathcal{S}_{end}} \mathcal{P}(s,a,b,q) \geq 1-\gamma$, where $\mathcal{S}_{end}$ is a return-free, invariant set of states.*

This contraction assumption means that at every time step, the interaction ends with probability at least $1-\gamma$ and is similar to having a discount factor of γ. For example, in a navigation example, it can mean that the target is reached or the battery runs out with a certain probability at every time step. Under the contraction property, the well-known simulation lemma [31], [30] shows that if two strategies are close for every state, then the returns of these strategies are close.

Lemma 4 (Lemma 1 from [31])**.** *Under Assumption 1, if $\|\boldsymbol{y}_k^*(s) - y^*(s)\|_1 \leq \epsilon$ for all $s \in \mathcal{S}$, then*

$$|SR(x) - IR(x)| \leq \frac{\gamma\epsilon}{(1-\gamma)^2}.$$

To invoke Lemma 4, we need to show the closeness of $\boldsymbol{y}_k^*(s)$ and $y^*(s)$. We note that Lemma 3 immediately applies to bound the closeness between $\boldsymbol{y}_k^*(s)$ and $y^*(s)$ as a function of $\|\hat{\boldsymbol{x}}_k(s) - x(s)\|$. However, different from Section IV-B, the sample actions of the leader are not independently sampled (since they come from interactions of a dynamic game), and Lemma 1 does not apply. To bound $\|\hat{\boldsymbol{x}}_k(s) - x(s)\|$, we use a result from [30] that can handle random stopping times.

Lemma 5 (Lemma 3 of [30], Theorem 2.1 of [24])**.** *Let w be the number of sample leader actions from state s in the first $k-1$ interactions. With probability at least $1-\delta$,*

$$\|\hat{\boldsymbol{x}}_k(s) - x(s)\|_1 \leq \epsilon\sqrt{\frac{2}{\varphi(x(s))}}$$

if $w \geq \frac{40|\mathcal{A}|}{\epsilon^2}\log\left(\frac{1}{\epsilon}\right)\log\left(\frac{3}{5\delta}\right)$.

To ensure that each state s has enough samples, we have the following assumption.

Assumption 2. *Let $\rho > 0$ and $\mu > 0$ be constants. Under any y, for all $s \in \mathcal{S} \setminus \mathcal{S}_{end}$,*

$$\Pr(\boldsymbol{s}_0 = s \vee \boldsymbol{s}_1 = s \ldots) \geq \rho$$

and

$$\Pr(\boldsymbol{s}_{t+1} = s \vee \boldsymbol{s}_{t+2} = s \ldots | \boldsymbol{s}_t = s) \geq \mu.$$

Assumption 2 ensures that every non-terminal state is reached and resampled with high probability, and the occupancy measure of each state is at least $\rho/(1-\mu)$. The following result is due to the concentration of Bernoulli and geometric random variables and shows that the number of samples from each state is lower bounded under Assumption 2.

Lemma 6 (Unification of Lemmas 4 and 5 from [30])**.** *Under Assumption 2, if*

$$k \geq \frac{6\max(8w(1-\mu), \log(2/\delta))}{\rho}$$

the number of sample leader actions from state s in the first $k-1$ interactions is at least w with probability at least $1-\delta$.

Combining the above results, we get the following result on the gap between $SR_k(x)$ and $IR(x)$.

Theorem 3. *Under Assumptions 1 and 2, if*

$$k \geq \frac{6\max(\frac{320|\mathcal{A}|}{\epsilon^2}\log(1/\epsilon)\log(9|\mathcal{S}|/5\delta)(1-\mu), \log(3|\mathcal{S}|/\delta))}{\rho},$$

then with probability $1-\delta$,

$$SR(x) - IR_k(x) \leq \frac{\lambda|\mathcal{B}|\sqrt{|\mathcal{A}|}\gamma\epsilon}{\sqrt{2}(1-\gamma)^2\min_s\sqrt{\varphi(x(s))}}.$$

Proof of Theorem 3. Setting the failure probability to $\delta/3|\mathcal{S}|$ in Lemma 5 and the failure probability to $2\delta/3|\mathcal{S}|$ in Lemma 6, and using the union bound, we get that

$$\|x_k(\boldsymbol{s}_t) - \hat{\boldsymbol{x}}_k(\boldsymbol{s}_t)\|_1 \leq \epsilon\sqrt{\frac{2}{\varphi(x(s))}}$$

for all $s \in \mathcal{S} \setminus \mathcal{S}_{end}$ with probability at least $1-\delta$. Using $\|z\|_2 \leq \|z\|_1 \leq \sqrt{d}\|z\|_2$ and Lemma 3, we have

$$\|y^*(\boldsymbol{s}_t) - \hat{\boldsymbol{y}}_k(\boldsymbol{s}_t)\|_1 \leq \frac{\lambda|\mathcal{B}|\sqrt{|\mathcal{A}|}\epsilon}{\sqrt{2\varphi(x(s))}}$$

for all $s \in \mathcal{S} \setminus \mathcal{S}_{end}$ with probability at least $1-\delta$. Using this result in Lemma 4 gives the desired result. □

The inferability gap has the term $\sqrt{\varphi(x(s))}$ that measures the concentration of the leader's strategy for state s. The term $\frac{1}{\min_s\varphi(x(s))}$ in the bound is analogous to the stochasticity level ν: As the strategies for every state get more deterministic, $\frac{1}{\min_s\varphi(x(s))}$ decreases making the inferability gap diminish. In the extreme case where the leader's strategy is deterministic for every state, we have $1/\min_s\sqrt{\varphi(x(s))} = 0$, which implies that the inferability gap is 0 with high probability after a certain number of interactions.

Similar to the static setting $\tilde{\mathcal{O}}(1/\epsilon^2)$ interactions are sufficient to achieve an inferability gap of ϵ with high probability.

We also note that the result for the fully rational followers can also be extended to this setting by using the closeness between $\boldsymbol{y}_k^*(s)$ and $y^*(s)$ as described in Section IV-C.

F. Converse Bound for Repeated Static Bimatrix Stackelberg Games with Fully Rational Followers

Theorem 2 shows that with a fully rational follower, the gap between the leader's expected return in the full information setting and in the inference setting, $SR(x) - IR_k(x)$, 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 $SR_(x) - IR_k(x) \leq \epsilon$. 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

$$A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}. \tag{4}$$

For these choices of A and B, we have

$$\left(\sup_{x\in\Delta^m} x^\top Ay^* \quad \text{s.t.} \quad y^* = \arg\max_{y\in\Delta^n} x^\top By\right) = \frac{1}{2}$$

Proposition 1. *Let A and B be as defined in* (4)*. For every $\epsilon \in (0, 1/2)$ and $x \in \Delta^2$ such that $SR(x) \geq (1/2) - \epsilon$, if*

$$k \leq \frac{1 - 20\epsilon + 128\epsilon^2 + 80\epsilon^3 - 400\epsilon^4}{32\epsilon^2},$$

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

Proof of Proposition 1. We follow a proof similar to the proofs for bandit lower bounds. For a strategy with high return under full information, we choose an alternative, close strategy. We show that the alternative strategy has a low return in the inference setting, and these strategies have similar returns in the inference setting since they are not distinguishable. Therefore, the return of the strategy with high return under full information is low in the inference setting.

We first define some notation. Let $E^1 = \{x | x \in \Delta^2, x^\top [1, 1] > 0\}$ and $E^2 = \Delta^2 \setminus E^1$. Note that E^1 is the set of leader strategies for which action 1 is optimal for the follower, and E^2 is the set of leader strategies for which action 2 is optimal for the follower. We consider two strategies:

$$x = \left[\frac{1}{2} + \epsilon,\ \frac{1}{2} - \epsilon\right] \text{ and } z = \left[\frac{1}{2} - \epsilon,\ \frac{1}{2} + \epsilon\right].$$

These strategies are both ϵ away from the follower's decision boundary at $[1/2, 1/2]$. We prove the statement in four steps:

1) Show that the expected return of x under full information $SR(x)$ is lower bounded.
2) Show that the expected return of z under inference $IR_k(z)$ is upper bounded.
3) Show that the expected returns of z and x under inference, $IR_k(z)$ and $IR_k(x)$ are close.
4) Combine the above results.

Step 1: Note that $x \in E^2$. Consequently, $SR(x) = (1/2) - \epsilon$ which trivially implies

$$SR(x) \geq \frac{1}{2} - \epsilon. \tag{5}$$

Step 2: Next, we upper bound $IR_k(z)$. We have

$$\begin{aligned} IR_k(z) &= \sum_{i=1}^{2} \sum_{j=1}^{2} \Pr(\boldsymbol{i}_k = i|z) \Pr(\boldsymbol{j}_k = j|z) A_{i,j} \\ &= \left(\frac{1}{2} + \epsilon\right) \Pr(\boldsymbol{j}_k = 1|z) \end{aligned}$$

since $A_{1,1} = A_{1,2} = A_{2,2} = 0$, $A_{2,1} = 1$, and $\Pr(\boldsymbol{i}_k = 2|z) = (1/2) + \epsilon$. We note that $\boldsymbol{j}_k = 1$ if and only if $\hat{\boldsymbol{z}}_k \in E^1$, i.e., the empirical distribution of the leader's actions belongs to E^1. It implies that $\Pr(\hat{\boldsymbol{z}}_k \in E^1|z) = \Pr(\boldsymbol{j}_k = 1|z)$. $\Pr(\hat{\boldsymbol{z}}_k \in E^1|z) \leq 1/2$ since z has more bias towards action 2. We have

$$IR_k(z) \leq \frac{1}{2}\left(\frac{1}{2} + \epsilon\right) = \frac{1}{4} + \frac{\epsilon}{2}. \tag{6}$$

Step 3: Finally, we upper bound $|IR_k(x) - IR_k(z)|$. Note that

$$\begin{aligned} &IR_k(x) - IR_k(z) \\ &= \Pr(\boldsymbol{i}_k = 2|x) \Pr(\boldsymbol{j}_k = 2|x) - \Pr(\boldsymbol{i}_k = 2|z) \Pr(\boldsymbol{j}_k = 2|z) \\ &= \left(\frac{1}{2} - \epsilon\right) \Pr(\boldsymbol{j}_k = 2|x) - \left(\frac{1}{2} + \epsilon\right) \Pr(\boldsymbol{j}_k = 2|z) \\ &\leq \frac{1}{2}(\Pr(\boldsymbol{j}_k = 1|x) - \Pr(\boldsymbol{j}_k = 1|z)). \end{aligned} \tag{7a}$$

Let $\mathcal{D}_{\hat{\boldsymbol{z}}_k}$ and $\mathcal{D}_{\hat{\boldsymbol{x}}_k}$ be the distributions of $\hat{\boldsymbol{z}}_k$ and $\hat{\boldsymbol{x}}_k$, respectively. Also, let $KL(\mathcal{D}^1||\mathcal{D}^2)$ denote the KL divergence between distributions $\mathcal{D}^1$ and $\mathcal{D}^2$, and $Be(p)$ denote the Bernoulli random variable with parameter p. We have

$$KL(\mathcal{D}_{\hat{\boldsymbol{z}}_k}||\mathcal{D}_{\hat{\boldsymbol{x}}_k}) = (k-1) KL\left(Be\left(\frac{1}{2} - \epsilon\right)||Be\left(\frac{1}{2} + \epsilon\right)\right)$$

since the leader's actions are withdrawn from z (and from x) at every interaction independently. Using the bound given in Theorem 1 of [32], we get $KL(\mathcal{D}_{\hat{\boldsymbol{z}}_k}||\mathcal{D}_{\hat{\boldsymbol{x}}_k}) \leq \frac{16\epsilon^2(k-1)}{1-4\epsilon^2}$.

Since $\boldsymbol{j}_k$ is a function of the emprical distribution $\hat{\boldsymbol{z}}_k$ ($\hat{\boldsymbol{x}}_k$), using the data processing inequality [33], we get

$$KL(Be(\Pr(\boldsymbol{j}_k = 1|z))||Be(\Pr(\boldsymbol{j}_k = 1|x))) \leq KL(\mathcal{D}_{\hat{\boldsymbol{z}}_k}||\mathcal{D}_{\hat{\boldsymbol{x}}_k})$$

By Pinsker's inequality and the above inequalities,

$$|\Pr(\boldsymbol{j}_k = 1|z) - \Pr(\boldsymbol{j}_k = 1|x)| \leq \sqrt{\frac{8\epsilon^2(k-1)}{1-4\epsilon^2}}.$$

Combining this with (7), we get

$$IR_k(x) - IR_k(z) \leq \sqrt{\frac{2\epsilon^2(k-1)}{1-4\epsilon^2}}. \tag{8}$$

Step 4: Combining (5), (6), and (8) yields

$$SR(x) - IR_k(x) \geq \frac{1}{4} - \frac{3\epsilon}{2} - \sqrt{\frac{2\epsilon^2(k-1)}{1-4\epsilon^2}}.$$

The bound is a monotone function of k. Setting the right-hand side to ϵ and solving for k yields the desired result. □

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 full 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. The returns for the strategies close to the boundary will be poor since the empirical distribution may be on the other side of the decision boundary.

The alternative bound from in Section IV-C implies that for an inferability gap of ϵ, $\frac{m \log 2 + \log(1/\epsilon)}{d(x)^2 \varphi(x)}$ interactions are sufficient. Using $\varphi(x) \geq 2$, we get that if $k \geq \frac{m \log 2 + \log(1/\epsilon)}{2d(x)^2} + 1$ then $SR(x) - IR_k(x) \leq \epsilon$. For the example used in Proposition 1, we have $d(x) = \epsilon$, which implies if $k \geq \frac{m \log 2 + \log(1/\epsilon)}{2\epsilon^2} + 1 = \tilde{\mathcal{O}}(1/\epsilon^2)$ then $SR(x) - IR_k(x) \leq \epsilon$. On the other hand, Proposition 1 shows that if $k \leq \frac{1 - 20\epsilon + 128\epsilon^2 + 80\epsilon^3 - 400\epsilon^4}{32\epsilon^2} = \tilde{\mathcal{O}}(1/\epsilon^2)$, then $SR(x) - IR_k(x) \geq \epsilon$. Therefore, the bound derived using the concentrability metric $\varphi(x)$ is optimal up to logarithmic factors.

G. Static Bimatrix Games with Limited Inferability Gaps

The inferability of mixed strategies is particularly important in general-sum games where the objectives of the players are weakly positively correlated. Consider the static bimatrix game setting. 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.

Proposition 2 shows that for bimatrix Stackelberg games, there exist strategies with a limited inferability gap (regardless of the interaction number) if the game is almost cooperative or competitive. In detail, we have

$$\left(\max_x IR_k(x)\right) = \left(\max_x \; x^\top A \boldsymbol{y}_k^* \quad \text{s.t. } \boldsymbol{y}_k^* = \arg\max_{y\in\Delta^n} \hat{\boldsymbol{x}}_k^\top B y\right),$$

and

$$\left(\max_x SR(x)\right) = \left(\max_x \; x^\top A y^* \quad \text{s.t. } y^* = \arg\max_{y\in\Delta^n} x^\top B y\right).$$

For an almost cooperative or competitive game, there exists a strategy for the leader such that the expected return of this strategy in the inference setting, is approximately at the level of the expected return $\max_x SR(x)$ of optimal strategies in the full information setting. Consequently, $\max_x IR_k(x)$ is greater than $\max_x SR(x)$ minus a constant depending on the cooperativeness or competitiveness of the game. To formally define the cooperativeness and competitiveness level of the game, we decompose A and B. Let $U^c = A/2 + B/2$ (the cooperative objective) and $U^z = A/2 - B/2$ (the zero-sum objective). The leader's utility matrix is $U^c + U^z$, and the follower's utility matrix is $U^c - U^z$. Let $\bar{U}^c = (U^c - c^c)/\alpha^c$ and $\bar{U}^z = (U^z - c^z)/\alpha^z$ be the shifted and normalized versions of U^c and U^z, respectively, such that $\max_{i,j} \bar{U}^c_{i,j} = 1$, $\min_{i,j} \bar{U}^c_{i,j} = 0$, $\max_{i,j} \bar{U}^z_{i,j} = 1$, and $\min_{i,j} \bar{U}^z_{i,j} = 0$. The game is zero-sum if $\alpha^c = 0$ and fully cooperative if $\alpha^z = 0$.

Proposition 2. *For a repeated static bimatrix Stackelberg game and $k \geq 2$:*

1)
$$\left(\max_x IR_k(x)\right) \geq \left(\max_x SR(x)\right) - 2\alpha^c.$$

2) *If the maximum element of B_{ij} is unique and the follower's estimate of the leader's strategy is unbiased,*
$$\left(\max_x IR_k(x)\right) \geq \left(\max_x SR(x)\right) - 2\alpha^z.$$

Proof of Proposition 2. We prove the statements separately.

Statement 1: Consider that the leader optimizes for the negative of the follower's objective, i.e., $-B = -U^c + U^z$. In this case, the leader's optimal strategy x^* is a Nash equilibrium strategy for the zero-sum objective function $-U^c + U^z$. x^* is also a Nash equilibrium strategy for the zero-sum objective function $-U^c + U^z + 2c^cJ$ since $2c^cJ$ is constant. Define

$$y^* = \arg\max_{y\in\Delta^n} x^\top B y = \arg\max_{y\in\Delta^n} x^\top (U^c - U^z) y,$$

$$y^{**} = \arg\max_{y\in\Delta^n} (x^*)^\top B y = \arg\max_{y\in\Delta^n} (x^*)^\top (U^c - U^z) y,$$

and

$$\boldsymbol{y}_k^* = \arg\max_{y\in\Delta^n} \hat{\boldsymbol{x}}_k^\top B y = \arg\max_{y\in\Delta^n} \hat{\boldsymbol{x}}_k^\top (U^c - U^z) y.$$

Since y^{**} is a Nash equilibrium strategy, for all $k \geq 2$,

$$\begin{aligned}&\mathbb{E}\left[(x^*)^\top(-U^c + c^cJ + U^z + c^cJ)y^{**}\right]\\ &\leq \mathbb{E}\left[(x^*)^\top(-U^c + c^cJ + U^z + c^cJ)\boldsymbol{y}_k^*\right]\end{aligned}$$

Since all entries of $U^c - c^cJ$ are positive, we have

$$\begin{aligned}&\mathbb{E}\left[(x^*)^\top(-U^c + c^cJ + U^z + c^cJ)\boldsymbol{y}_k^*\right]\\ &\leq \mathbb{E}\left[(x^*)^\top(U^c - c^cJ + U^z + c^cJ)\boldsymbol{y}_k^*\right]\\ &= \mathbb{E}\left[(x^*)^\top(U^c + U^z)\boldsymbol{y}_k^*\right]\end{aligned}$$

which is the expected return of the strategy x^* under inference at interaction k. Since x^* is a feasible strategy, we have

$$\begin{aligned}\left(\max_x IR_k(x)\right) &\geq \mathbb{E}\left[(x^*)^\top(-U^c + U^z + 2c^cJ)y^{**}\right]\\ &= \mathbb{E}\left[(x^*)^\top(-U^c + U^z)y^{**}\right] + 2c^c\end{aligned}$$

For all $k \geq 2$, we have

$$\begin{aligned}\max_x \mathbb{E}\left[x^\top A y^*\right] &= \max_x \mathbb{E}\left[x^\top(-U^c + 2U^c + U^z)y^*\right]\\ &\leq \max_x \mathbb{E}\left[x^\top(-U^c + U^z)y^* + 2\alpha^c + 2c^c\right]\\ &= \mathbb{E}\left[(x^*)^\top(-U^c + U^z)y^{**} + 2\alpha^c + 2c^c\right].\end{aligned}$$

The inequality is because for every x there exists a unique y^*, and $\max_{x,y^*} 2x^\top U^c y^* = 2\alpha^c + 2c^c$. Hence, we have

$$\left(\max_x SR(x)\right) \leq \mathbb{E}\left[(x^*)^\top(-U^c + U^z)y^{**} + 2\alpha^c + 2c^c\right].$$

Combining the bounds on $\max_x SR(x)$ and $\max_x IR_k(x)$ yields the desired result.

Statement 2: Note that if B has a unique maximum element then $U^c - U^z + 2c^zJ$ and $\alpha^c\bar{U}^c - \alpha^z\bar{U}^z + (c^c + c^z)J$ also have unique maximum elements at the same locations since they are shifted versions of B by $2c^zJ$. Let (i, j) be the index of the maximum element.

Consider that the leader optimizes for the follower's objective function (with a constant offset of $2c^z$), i.e., $U^c - U^z + 2c^zJ$, and its strategy x^* plays action i deterministically. In this case, the leader's return is $\alpha^c\bar{U}^c_{ij} - \alpha^z\bar{U}^z_{ij} + c^c + c^z$ in the full information setting since the follower will play action j. Similarly in the inference setting, the leader's return is $\alpha^c\bar{U}^c_{ij} - \alpha^z\bar{U}^z_{ij} + c^c + c^z$ in every interaction after the first interaction since the follower will infer the leader's strategy with no error and play action j. Note that $\alpha^c\bar{U}^c_{ij} - \alpha^z\bar{U}^z_{ij} + c^c + c^z \geq \alpha^c - \alpha^z + c^c + c^z$ since

$$\max_{k,l} \alpha^c\bar{U}^c_{kl} - \alpha^z\bar{U}^z_{kl} \geq \max_{k,l} \alpha^c\bar{U}^c_{kl} - \max_{k,l} \alpha^z\bar{U}^z_{kl} \geq \alpha^c - \alpha^z.$$

Consequently, we have $\mathbb{E}\left[(x^*)^\top A\boldsymbol{y}_k^*\right] \geq (\alpha^c - \alpha^z + c^c + c^z)$ which implies that $(\max_x IR_k(x)) \geq (\alpha^c - \alpha^z + c^c + c^z)$.

We have $A = \alpha^c\bar{U}^c + \alpha^z\bar{U}^z + (c^c + c^z)J$ and $\max_{k,l} \alpha^c\bar{U}^c_{kl} + \alpha^z\bar{U}^z_{kl} \leq \max_{k,l} \alpha^c\bar{U}^c_{kl} + \max_{k,l} \alpha^z\bar{U}^z_{kl} = \alpha^c + \alpha^z$ which imply that $(\max_x SR(x)) \leq \alpha^c + \alpha^z + c^c + c^z$.

Arranging the inequalities for $(\max_x IR_k(x))$ and $(\max_x IR_k(x))$ yields the results for statement 2. □

If α^z is small (i.e., the competitive aspect of the game is insignificant,) the leader can optimize for the follower's objective function instead of its own objective function. By considering a joint objective function, we can observe that in the full

information setting, there exists a deterministic strategy for the leader that is near-optimal for the leader's own objective (assuming that the follower has a unique optimal response to the leader's strategy.) Since the strategy is deterministic, the leader does not suffer from an inferability gap in the inference setting compared to the full information setting. Since the ignored part of the objective function is insignificant, the expected return of this strategy in the inference setting is near the optimal return in the full information setting. If α^c is small, (i.e., the cooperative aspect of the game is insignificant,) the leader can optimize for the opposite of the follower's objective function instead of its own objective function. By considering the zero-sum equilibrium, we can observe that in the full information setting, there exists a mixed strategy for the leader that is near-optimal in the absolute sense for the leader's own objective. In the inference setting, the follower may not be able to infer the leader's strategy fully. However, since any error in inferring the leader's strategy leads to a different strategy for the follower, it can only decrease the follower's returns as the leader's strategy is a zero-sum equilibrium strategy. Consequently, it can only increase the leader's returns, and the leader does not suffer from an inferability gap compared to the full information setting. Since the ignored part of the objective function is insignificant, the expected return of this strategy in the inference setting is near the optimal return in the full information setting.

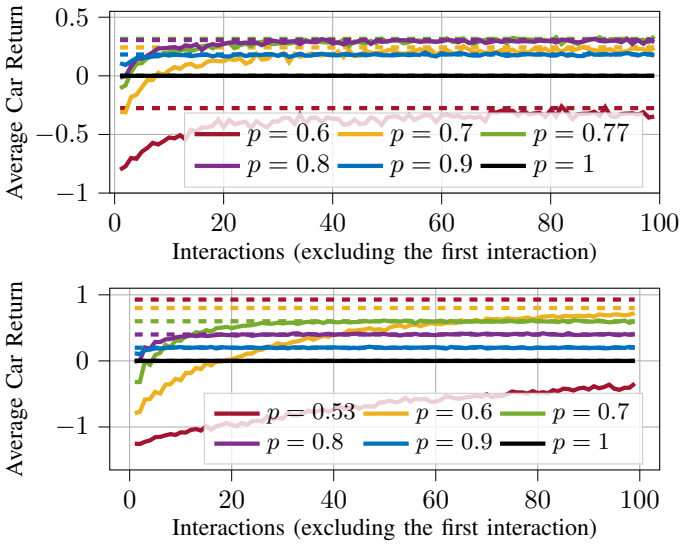

Fig. 1: The car's average return (averaged over 10^4 simulations) in the pedestrian-car example. Solid lines represent the average return for different strategies where p is the probability of the car stopping. Dashed lines represent the average return per interaction under full information, i.e., $x^\top A\sigma_\lambda(B^\top x)$ for $x = [p, 1-p]$. (Top) $\lambda = 5$. (Bottom) $\lambda = 100$.

V. Numerical Examples

In this section, we evaluate the effect of inference on repeated bimatrix Stackelberg games and a repeated Stackelberg game with parametric action spaces. For the bimatrix games, we consider the aforementioned car-pedestrian interaction and randomly generated bimatrix games. For clarity of presentation, we plot the average return $\frac{1}{K}\sum_{k=2}^{K} IR_k(x)$, 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 II with a boundedly rational follower with maximum entropy response. We simulate the gameplay under inference for 100 interactions with rationality constants $\lambda = 5$ and $\lambda = 100$. The car's strategy is determined by p, i.e., the probability that the car stops. For $\lambda = 5$ and $\lambda = 100$, the optimal p are 0.77 and 0.53 in the full information setting, respectively. We show the results in Fig. 1 for different values of p.

For $\lambda = 100$, all strategies receive higher average returns as the number of interactions increases as the pedestrian's estimation improves. In the long run, $p = 0.53$, the optimal strategy for the car in the full information setting, 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 $\hat{\boldsymbol{p}}_k$ 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 a small number of interactions K, and any error in $\hat{\boldsymbol{p}}_k$ results in only small changes to the pedestrian's strategy. For $\lambda = 5$, we also observe that more deterministic strategies such as $p = 0.8$ or $p = 0.9$ achieve higher returns in the transient period than $p = 0.9$, which is optimal for the full information case. In this case, the pedestrian is less rational, caring less about the leader's strategy, and takes actions more uniformly randomly. Consequently, the inferability gaps are smaller, aligned with the bound given in Corollary 1.

B. General-sum Tug of War

We consider a static Stackelberg game with parametric mixed strategies to demonstrate the importance of inferability. The leader's strategies are normal distributions $\mathcal{N}(\mu^{\mathrm{l}}, (s^{\mathrm{l}})^2)$ parametrized by $(\mu^{\mathrm{l}}, s^{\mathrm{l}}) \in \mathbb{R}^2$. Similarly, the follower's are normal distributions $\mathcal{N}(\mu^{\mathrm{f}}, (s^{\mathrm{f}})^2)$ parametrized by $(\mu^{\mathrm{f}}, s^{\mathrm{f}}) \in \mathbb{R}^2$. The leader's action is $a \sim \mathcal{N}(\mu^{\mathrm{l}}, (s^{\mathrm{l}})^2)$, and the follower's action is $b \sim \mathcal{N}(\mu^{\mathrm{f}}, (s^{\mathrm{f}})^2)$. Let $0 \leq c_{low} < c_{high}$ be constants. The return of the leader is 1 if $a + b \in [c_{low}, c_{high}]$, and 0 otherwise. Symmetrically around 0, the return of the follower is 1 if $a + b \in [-c_{high}, -c_{low}]$, and 0 otherwise.

We note that the leader and follower try to pull the sum of their actions in different directions. This aspect of the game resembles a competitive "tug of war". On the other hand, for inconclusive outcomes, i.e., $a+b \in (-c_{low}, c_{low})$, or extreme outcomes, i.e., $a+b \in (-\infty, -c_{high}) \cup (c_{high}, \infty)$, both parties receive the same low return of 0. This aspect of the game is cooperative in that both parties aim to avoid these regions.

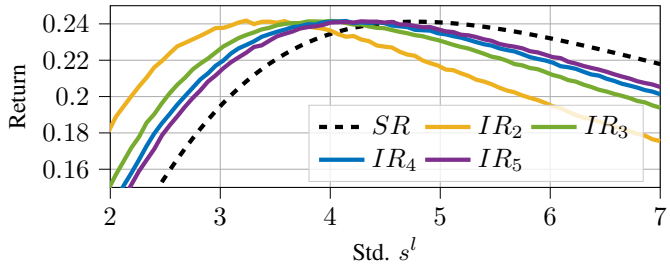

Fig. 2: The leader's expected return in the tug of war game for different mean and standard deviation values ($c_{low} = 0.01$ and $c_{high} = 5$). Lower values of standard deviation lead to higher returns for the lower number of interactions. The leader's mean parameter has no effect on the expected return.

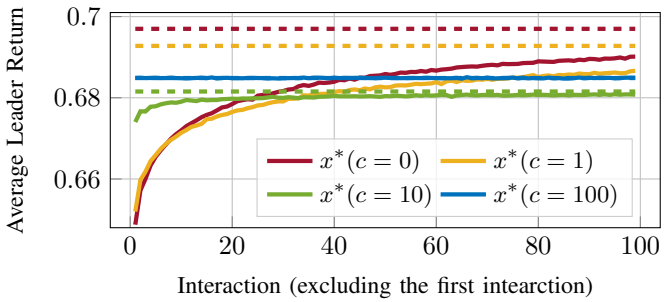

Fig. 3: The leader's average return for the randomly generated bimatrix games. Solid lines represent the average return for the bound's local maxima for different values of the regularization constant c. Dashed lines represent the average return per interaction under full information, i.e., $(x^*(c)))^\top A\sigma_\lambda(B^\top x^*(c))$.

Due to the independent sampling of a and b, we have $a + b \sim \mathcal{N}(\mu^{\mathrm{l}} + \mu^{\mathrm{f}}, (s^{\mathrm{l}})^2 + (s^{\mathrm{f}})^2)$. One can observe that given the leader's strategy $\mathcal{N}(\mu^{\mathrm{l}}, (s^{\mathrm{l}})^2)$, the follower's optimal strategy is $\mathcal{N}((-c_{high}-c_{low})/2-\mu^{\mathrm{l}}, 0)$ since it maximizes the probability of $a+b \in [-c_{high}, -c_{low}]$ which is the follower's expected return. Given the follower's optimal strategy, we have $a + b \sim \mathcal{N}((-c_{high} - c_{low})/2, (s^{\mathrm{l}})^2)$. The leader's expected return is the probability that $a + b \in [c_{low}, c_{high}]$ and the leader's mean parameter μ^{l} has no effect on the leader's expected return. As shown in Figure 2, $SR(\mu^{\mathrm{l}}, s^{\mathrm{l}}) = \Pr(a+b \in [c_{low}, c_{high}])$ is a unimodal function of s^{l} between $[0, \infty)$ and attains its maximum for a finite positive value of s^{l}. Consequently, the leader's optimal strategy is mixed.

Consider a scenario where the leader and the follower will interact a certain number of times. The follower estimates the leader's mean from the previous interactions using the plug-in estimator. After k interactions, the follower's estimate of the leader's mean parameter is $\hat{\boldsymbol{\mu}}^l_k$ that is distributed as $\mathcal{N}(\mu^{\mathrm{l}}, (s^{\mathrm{l}})^2)/k$. The follower's action b_{k+1} is distributed as $\mathcal{N}((-c_{high} - c_{low})/2 - \hat{\boldsymbol{\mu}}^l_k, 0) = \mathcal{N}((-c_{high} - c_{low})/2 - \mu^{\mathrm{l}}, (s^{\mathrm{l}})^2)/k)$. Since the leader's action $\boldsymbol{a}_{k+1}$ is distributed as $\mathcal{N}(\mu^{\mathrm{l}}, (s^{\mathrm{l}})^2)$, and $\boldsymbol{a}_{k+1}$ and b_{k+1} are independently sampled, we have $a_{k+1} + b_{k+1} \sim \mathcal{N}((-c_{high} - c_{low})/2, (s^{\mathrm{l}})^2) + (s^{\mathrm{l}})^2)/k)$, and the leader's return expected return at $(k+1)$-th interaction is the probability that $a_{k+1} + b_{k+1} \in [c_{low}, c_{high}]$. We note that the leader's mean parameter μ^{l} has no effect on the expected return since the follower's unbiased estimate cancels any changes to μ^{l}. On the other hand, the expected return depends on the leader's variance parameter. If the leader's strategy has high variance, i.e., it is not easily inferable, then the follower's estimation will be inaccurate. In return, the leader will suffer from an inferability gap. Figure 2 shows that strategies with lower variance achieve a higher return for lower number of interactions.

C. Randomly Generated Bimatrix Games

We evaluate the performance under inference for randomly generated bimatrix games when the follower is boundedly rational with maximum entropy response. From the achievability bound given in Corollary 1,

$$(K-1)SR(x) - c\,\nu(x) \leq \sum_{k=2}^{K} IR_k(x)$$

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

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

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. On the other hand, the objective is Lipschitz continuous thanks to the softmax response. To find a local optimum, 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. We note that for the bimatrix game under consideration, by exploiting that the pure strategies are optimal for the follower, such a strategy can be obtained by solving n linear programs [3]. Each of the linear programs fixes the follower's action and constrains the leader's strategies to a polytope to ensure that the considered follower action is optimal.

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 Section IV-G.

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

Fig. 4: The simulated driving environment from the human driver's point of view as the driver approaches to the T-intersection. The autonomous car is the red car on the left.

bimatrix is averaged at each interaction k over all bimatrix games. Results are shown in Fig. 3.

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 full 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 gap 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.

VI. Human Subject Study for the Autonomous Car Example

We assess the importance of inferability in repeated Stackelberg games through an in-person human subject study[3] in a simulated driving environment. The experiment closely follows the autonomous car and pedestrian interaction scenario described in Section III-D. Different from the autonomous car and pedestrian interaction, the human in the study is the driver who is controlling a car.

A. Experiment Scenario

The experiment represents the interaction at T-intersection in the CARLA simulator [34]. There are two agents in the environment, an autonomous car and a human-driven car. The human-driven car approaches the intersection on the terminating road, and the autonomous car approaches the intersection on the through road. A view from the environment is given in Fig. 4.

At the intersection, both the through road and the terminating road have yield signs. If the autonomous car decides to stop, it yields and gives the right of way to the human driver. If it decides not to stop, but the human also does not stop, the autonomous car makes an emergency stop to avoid a crash.

[3]This study was approved by UT IRB study #7222.

The human-driven car is controlled using a commercially available video game driving wheel and pedals.

B. Objectives and Rewards

The objectives of the autonomous car are to pass through the intersection in minimal time and avoid an emergency stop. The instructed objectives of the human driver are completing a left turn at the intersection in minimal time and avoiding getting fined for not stopping. Based on these objectives, we give (virtual) rewards to the participants.

To observe the interaction between the players at the intersection, we do not give a reward to the participants if the turn is not completed within a certain time frame. Completing the turn in time and stopping for the autonomous car is not possible in the experiment if the autonomous car initially decides not to stop for the human driver.

TABLE III: Utilities for the autonomous car and human driven car interaction

		Human's actions	
		Stop	Proceed
Autonomous car's actions	Stop	(0,2)	(0,1)
	Proceed	(2,0)	(-8,1)

Table III shows the (virtual) rewards given to the players when the human completes the turn in time. Similar to the example from Section III-D, the autonomous car collects a reward of 0 if it stops for the human driver. It collects a reward of 2 if it proceeds without stopping and the human driver stops. It collects a reward of -8 if it proceeds and the human decides to proceed as well because of an emergency stop.

The participants collect a reward of 2 if the turn is completed in time and they also stop for the autonomous car. They collect a reward of 1 if the turn is completed in time, but they do not stop for the autonomous car. They collect a reward of 0 if the turn is not completed in time.

As in the example from Section III-D, the optimal strategy for the autonomous car is to choose a stopping probability p such that $p > 0.5$ and $p \approx 0.5$ in the full information setting. Under such a strategy, the human driver's optimal action is to stop, and the autonomous car's expected return is approximately 0.5.

C. Experiment Setting and Independent Variables

We recruited 24 participants with ages ranging from 19 to 31. After the participants completed a series of training turns to get familiar with the controllers, we asked each participant to interact with two types of autonomous cars that have different stopping probabilities. The types consisted of autonomous cars with stopping probabilities of 0.1, 0.25, 0.4, 0.6, 0.75, and 0.9. We randomly assigned the participants to the types.

A participant interacted with a type 15 times, where each interaction lasted less than 45 seconds. At each interaction, the autonomous vehicle stopped with its type probability independent of the other interactions.

Before the interactions began, we instructed the participants to maximize their collected virtual rewards. After each interaction, we showed the participants the outcome of the scenario and the virtual reward.

The video recording of an interaction is available at https://github.com/mustafakarabag/Inferability.

D. Dependent Variables

Before each interaction (including the first interaction), participants filled out a survey regarding their estimations about the autonomous car. The questions and answer scales are (i) "What is the percentage that the other car would do a courtesy stop in the next interaction?" with allowed answers between 0 and 100 with increments of 1, and (ii) "What is your confidence level in this estimation?" with responses "Not confident at all", "Slightly confident", "Somewhat confident", "Fairly confident", and "Completely confident".

We collected the outcome of each interaction, including whether the autonomous car did a courtesy stop, the autonomous car did an emergency stop, the human driver stopped, and the human driver completed the turn in time.

E. Results and Discussion

We give the results for the study in Figs. 5-6. In Fig. 5, we plot the participant's estimate for the autonomous car's stopping probability after every interaction (solid lines) and the sample mean estimate for the autonomous car's stopping probability after every interaction (dashed lines). The pair of transparent lines for each color represents the data for a participant. The solid black line, A-E (Averaged estimate), represents the estimate of the autonomous car's stopping probability by humans, averaged over all participants. The dashed black line, A-E* (Averaged optimal estimates), represents the sample mean estimate of the autonomous car's stopping probability by humans, averaged over all participants.

The participants reported the following final confidence levels for the estimates after all interactions: 4.12 for $p = 0.1$, 3.71 for $p = 0.25$, 4.14 for $p = 0.4$, 3.77 for $p = 0.6$, 4.22 for $p = 0.75$, and 4.28 for $p = 0.9$ where "Not confident at all" is level 1 and "Completely confident" is level 5.

We observe that the participants' estimates for the autonomous car's stopping probability p converge to the actual value on average for most values of p. The estimates for the more deterministic values of p, such as 0.1 and 0.9, rapidly approached the actual value of p, while the estimates had higher variances for more stochastic values of p, indicating the significance of inferability. The averaged estimate for $p = 0.1$ remains consistently higher than the averaged sample mean value due to the two outlier sets of interactions where the participants' estimates were significantly higher than 0.5, despite the sample mean estimate converging to 0.1. We also note that while the estimates approach the actual value, the initial estimates suggest an anchoring effect around 0.5 rather than immediately following the sample mean estimates. While the participants were told that the car would have an arbitrary type, the majority of the participants' initial estimate was 0.5.

In Fig. 6, we plot the autonomous car's mean return from the first interaction to the last interaction (solid lines) and the autonomous car's hypothetical mean return if the human had acted optimally based on the previous samples after every

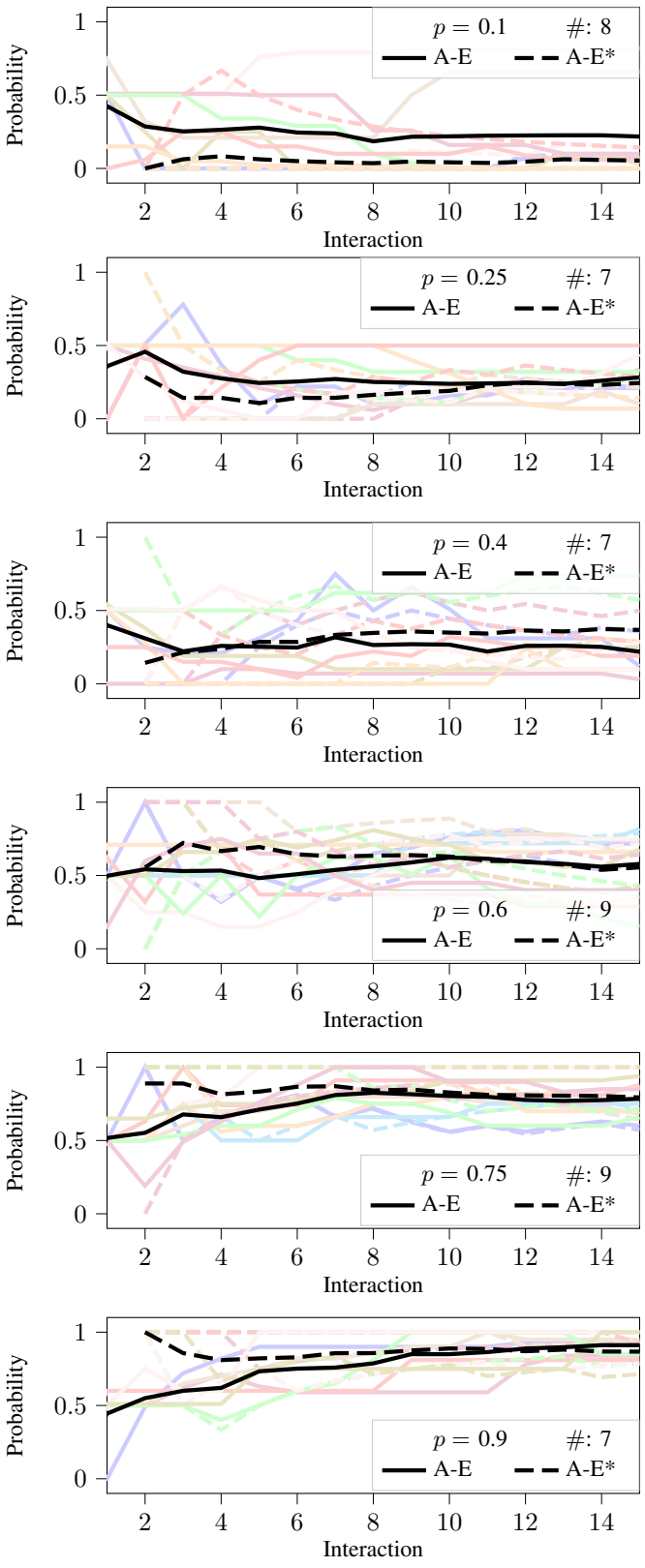

Fig. 5: Human and sample mean estimates after every interaction for different stopping probabilities, p. The number of participants is given with $\#$.

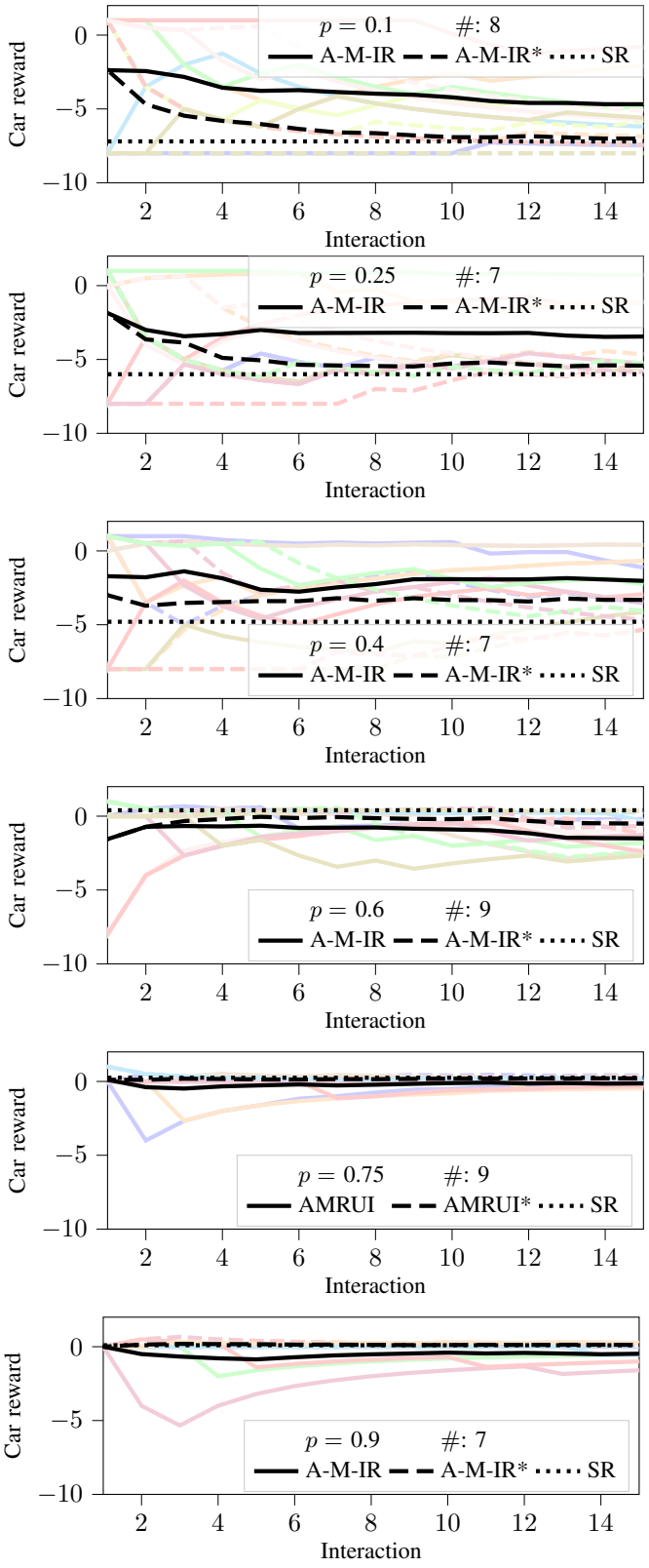

Fig. 6: Mean car return after every interaction for different stopping probabilities, p. The number of participants is given with $\#$.

interaction(dashed lines), and the Stackelberg return (dotted lines). The pair of transparent lines for each color represents the data for a participant. The solid black line, A-M-IR (averaged mean return under inference), represents the autonomous car's mean return, averaged over all participants. The dashed black line, A-M-IR* (averaged mean return under inference under optimal human action), represents the autonomous car's hypothetical mean return if the human had acted optimally based on the previous samples, averaged over all participants.

We observe that, as suggested by our theoretical results, the stochastic strategies that have higher returns in the full information setting suffer from an inferability gap. For example, the strategy with $p = 0.6$ has a higher return in the full information setting. On the other hand, this strategy is outperformed by the strategies $p = 0.75$ and $p = 0.9$ in the inference setting due to its high variance, leading to estimation errors and changes in the human driver's decisions. We note that the strategy $p = 0.6$ even has a negative empirical return under inference, which means that it would also be outperformed by the deterministic strategy $p = 1$ that always stops.

While the strategies with $p < 0.5$ have return $-8(1-p) \in [-8, -4)$ in the full information setting, when interacting with humans, these strategies have higher empirical returns than their respective full information counterparts. This is due to two reasons. Firstly, the lack of inferability, in fact, helps with these strategies. Since the full information returns are already low, the followers' (humans') deviations from the optimal action under full information due to misinference improve the return under inference. Secondly, we observe that these strategies collect rewards (A-M-IR) that are even higher than the levels that would be achieved if the participant were to act optimally based on the previous samples (A-M-IR*). This difference is likely due to the inherent bias of the participants toward stopping, which makes the autonomous car collect a reward of 1 rather than -8 since it is more likely to proceed.

VII. Conclusions

When interacting with other non-competitive agents, an agent should have an inferable behavior to inform others about its intentions effectively. We model the inferability problem using repeated Stackelberg games where the follower infers the leader's strategy via observation from previous interactions. For a variety of repeated Stackelberg game settings, we show that in the inference setting, the leader may suffer from an inferability gap compared to the setting where the follower has full information of the leader's strategy. However, this gap is upper bounded by a function that depends on the stochasticity level of the leader's strategy. The bounds and the results for the experiments and the human subject study 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 full information setting.

Appendix: Proofs of the Technical Results

Proof of Lemma 1. Due to the linearity of expectation, we have $\mathbb{E}[\|\hat{\boldsymbol{x}}_k - x\|^2] = \sum_{i=1}^{m} \mathbb{E}[((\hat{\boldsymbol{x}}_k)_i - (x)_i)^2]$. Note that $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_{k-1}$ is sampled independently from the distribution x,

and $\Pr(\boldsymbol{i}_t = a|x) = (x)_i$ for $t = 1, \ldots, k-1$. Consequently, $(k-1)(\hat{\boldsymbol{x}}_k)_i$ is a binomial distribution with parameters $k-1$ and $(x)_i$, and satisfies

$$\mathbb{E}[((k-1)(\hat{\boldsymbol{x}}_k)_i - (k-1)(x)_i)^2] = (k-1)(x)_i(1-(x)_i).$$

Due to this fact, we have

$$\mathbb{E}[\|\hat{\boldsymbol{x}}_k - x\|^2] = \frac{\nu(x)^2}{k-1}. \qquad \square$$

Proof of Lemma 3. We first define

$$q^{\mathrm{l}} = \max_{i,j} A_{ij} + \min_{i,j} A_{ij}, \quad q^{\mathrm{f}} = \max_{i,j} B_{ij} + \min_{i,j} B_{ij}.$$

Let J be a matrix of ones. We note that

$$\sigma_\lambda(B^\top \hat{x}_k) = \sigma_\lambda\left(\left(B - {}^{Jq^{\mathrm{f}}}\!/_2\right)^\top \hat{x}_k\right)$$
$$\sigma_\lambda(B^\top x) = \sigma_\lambda\left(\left(B - {}^{Jq^{\mathrm{f}}}\!/_2\right)^\top x\right)$$

since subtracting the same constant from all elements does not change the result of the softmax function. Let z_i be the i^{th} column of $B - {}^{Jq^{\mathrm{f}}}\!/_2$. We have

$$\begin{aligned}
&\|\sigma_\lambda(B^\top x) - \sigma_\lambda(B^\top \hat{x}_k)\| \\
&= \left\|\sigma_\lambda\left(\left(B - \frac{Jq^{\mathrm{f}}}{2}\right)^\top x\right) - \sigma_\lambda\left(\left(B - \frac{Jq^{\mathrm{f}}}{2}\right)^\top \hat{x}_k\right)\right\| \\
&\le \lambda \left\|\left(B - \frac{Jq^{\mathrm{f}}}{2}\right)^\top x - \left(B - \frac{Jq^{\mathrm{f}}}{2}\right)^\top \hat{x}_k\right\| && \text{(9a)} \\
&= \lambda \sqrt{\sum_{i=1}^n \langle x - \hat{x}_k, z_i\rangle^2} && \text{(9b)} \\
&\le \lambda \sqrt{\sum_{i=1}^n \|x - \hat{x}_k\|^2 \|z_i\|^2} && \text{(9c)} \\
&\le \|x - \hat{x}_k\| \, \frac{\sqrt{mn}}{2} && \text{(9d)}
\end{aligned}$$

where (9a) is due to the λ-Lipschitzness of σ_x, (9c) is due to Cauchy-Schwartz ineq., (9d) is due to $\max_{i,j} |z_{i,j}| = {}^1\!/_2$. $\square$

Proof of Lemma 2. Let J be a matrix of ones, and z_i be the i^{th} column of $A - {}^{Jq^{\mathrm{f}}}\!/_2$. We have

$$\begin{aligned}
&|x^\top A y_k - x^\top A y| \\
&= \left|x^\top A y_k - x^\top A y - x^\top \frac{q^{\mathrm{l}}}{2}(Jy_k - Jy)\right| && \text{(10a)} \\
&\le \left\|x^\top \left(A - \frac{q^{\mathrm{l}}}{2} J\right)\right\| \|y_k - y\| && \text{(10b)} \\
&\le \max_{x' \in \Delta^m} \left(\sum_i^n (x'^\top z_i)^2\right)^{1/2} \|y_k - y\| \\
&\le \left(\sum_i^n \max_{x' \in \Delta^m} (x'^\top z_i)^2\right)^{1/2} \|y_k - y\| \\
&\le \left(\frac{n^2}{4}\right)^{1/2} \|y_k - y\| = \frac{\sqrt{n}}{2}\|y_k - y\| && \text{(10c)}
\end{aligned}$$

where (10a) is due to $Jy = Jy_k$, (10b) is due Cauchy-Schwartz ineq., and (10c) is due to $\max_j |z_{ij}| \le {}^1\!/_2$. $\square$

Proof sketch for Lemma 4. The proof follows from considering the dynamic game as a discounted MDP (since the leader's strategy is fixed and we have Assumption 1), where the follower's strategies in different settings in the dynamic game are different policies for the MDP. $\square$

Proof sketch for Lemma 5. Lemma 3 [30] uses Theorem 2.1 of [24], the upper bound 2 on $\varphi(p_s)$, and the union bound to derive a similar looser bound that does not depend on the concentration of $x(s)$. The proof of Lemma 5 follows from not invoking the upper bound on $\varphi(p_s)$ and following the same steps as in Lemma 3 [30]. $\square$

Proof sketch for Lemma 6. The proof directly follows from combining Lemmas 4 and 5 from [30]. $\square$

Acknowledgement

This work was supported in part by the National Science Foundation (NSF), under grant numbers 1836900, 2145134, 2211432, 2211548, 2336840, and 2423131. The authors would like to thank Tichakorn Wongpiromsarn for insightful discussions that motivated the considered problem. The authors would like to thank Neel Bhatt for his help in creating the simulation environment for the human subject study.

References

[1] Vincent Conitzer. On Stackelberg mixed strategies. *Synthese*, 193(3):689–703, 2016.

[2] Mustafa O Karabag, Sophia Smith, David Fridovich-Keil, and Ufuk Topcu. Encouraging inferable behavior for autonomy: Repeated bimatrix Stackelberg games with observations. *2024 American Control Conference (ACC)*, 2024.

[3] Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 82–90, 2006.

[4] Bernhard Von Stengel and Shmuel Zamir. Leadership games with convex strategy sets. *Games and Economic Behavior*, 69(2):446–457, 2010.

[5] Dmytro Korzhyk, Vincent Conitzer, and Ronald Parr. Solving Stackelberg games with uncertain observability. In *AAMAS*, pages 1013–1020, 2011.

[6] James Pita, Manish Jain, Milind Tambe, Fernando Ordónez, and Sarit Kraus. Robust solutions to Stackelberg games: Addressing bounded rationality and limited observations in human cognition. *Artificial Intelligence*, 174(15):1142–1171, 2010.

[7] Jan Karwowski, Jacek Mańdziuk, and Adam Żychowski. Sequential Stackelberg games with bounded rationality. *Applied Soft Computing*, 132:109846, 2023.

[8] Zhengyu Yin, Manish Jain, Milind Tambe, and Fernando Ordonez. Risk-averse strategies for security games with execution and observational uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25-1, pages 758–763, 2011.

[9] Farhad Farokhi and Henrik Sandberg. Fisher information as a measure of privacy: Preserving privacy of households with smart meters using batteries. *IEEE Transactions on Smart Grid*, 9(5):4726–4734, 2017.

[10] Farhad Farokhi and Henrik Sandberg. Ensuring privacy with constrained additive noise by minimizing Fisher information. *Automatica*, 99:275–288, 2019.

[11] Mustafa O Karabag, Melkior Ornik, and Ufuk Topcu. Least inferable policies for Markov decision processes. In *2019 American Control Conference (ACC)*, pages 1224–1231. IEEE, 2019.

[12] Christopher Bodden, Daniel Rakita, Bilge Mutlu, and Michael Gleicher. A flexible optimization-based method for synthesizing intent-expressive robot arm motion. *The International Journal of Robotics Research*, 37(11):1376–1394, 2018.

[13] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308. IEEE, 2013.

[14] Yasmine Beck, Ivana Ljubić, and Martin Schmidt. A survey on bilevel optimization under uncertainty. *European Journal of Operational Research*, 2023.

[15] Michael Patriksson and Laura Wynter. Stochastic nonlinear bilevel programming. In *Technical report. PRISM*. Citeseer, 1997.

[16] Michael Patriksson and Laura Wynter. Stochastic mathematical programs with equilibrium constraints. *Operations research letters*, 25(4):159–167, 1999.

[17] Gui-Hua Lin and Masao Fukushima. Stochastic equilibrium problems and stochastic mathematical programs with equilibrium constraints: A survey. *Pacific Journal of Optimization*, 6(3):455–482, 2010.

[18] Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Zebang Shen. Stochastic conditional gradient++. *arXiv preprint arXiv:1902.06992*, 2019.

[19] Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

[20] Negar Mehr, Mingyu Wang, Maulik Bhatt, and Mac Schwager. Maximum-entropy multi-agent dynamic games: Forward and inverse solutions. *IEEE Transactions on Robotics*, 39(3):1801–1815, 2023.

[21] M Hosein Zare, Oleg A Prokopyev, and Denis Sauré. On bilevel optimization with inexact follower. *Decision Analysis*, 17(1):74–95, 2020.

[22] Craig R Fox and Robert T Clemen. Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science*, 51(9):1417–1432, 2005.

[23] Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *CoRR*, abs/1704.00805, 2017.

[24] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, page 125, 2003.

[25] Ariel Rubinstein. *Modeling bounded rationality*. MIT press, 1998.

[26] Daniel A Braun and Pedro A Ortega. Information-theoretic bounded rationality and ε-optimality. *Entropy*, 16(8):4662–4676, 2014.

[27] Tamer Başar and Geert Jan Olsder. *Dynamic noncooperative game theory*. SIAM, 1998.

[28] Xuan Di, Xu Chen, and Eric Talley. Liability design for autonomous vehicles and human-driven vehicles: A hierarchical game-theoretic approach. *Transportation research part C: emerging technologies*, 118:102710, 2020.

[29] Adam Millard-Ball. Pedestrians, autonomous vehicles, and cities. *Journal of planning education and research*, 38(1):6–12, 2018.

[30] Mustafa O Karabag and Ufuk Topcu. On the sample complexity of vanilla model-based offline reinforcement learning with dependent samples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37-7, pages 8195–8202, 2023.

[31] Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

[32] Sever S Dragomir, Marcel L Scholz, and Jadranka Sunde. Some upper bounds for relative entropy and applications. *Computers & Mathematics with Applications*, 39(9-10):91–100, 2000.

[33] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[34] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

Mustafa O. Karabag is a postdoctoral fellow in the Oden Institute for Computational Engineering & Sciences at the University of Texas at Austin. He received his Ph.D. degree from the University of Texas at Austin in 2023. His research focuses on developing theory and algorithms to control the information flow of autonomous systems to succeed in information-scarce or adversarial environments.

Sophia Smith received her B.S. in mathematics and her B.A. in Physics at the University of Chicago, Chicago IL, USA in 2021. She is currently working towards the Ph.D. degree in Computational Science Engineering in Math within the Oden Institute at the University of Texas at Austin, Austin, TX, USA. Her research interests include developing theory and algorithms for decomposing teams of autonomous agents.

Negar Mehr is an assistant professor in the Department of Mechanical Engineering at the University of California, Berkeley. Before that, she was an assistant professor of Aerospace Engineering at the University of Illinois Urbana-Champaign. She was a postdoctoral scholar at Stanford Aeronautics and Astronautics department from 2019 to 2020. She received her Ph.D. in Mechanical Engineering from UC Berkeley in 2019 and her B.Sc. in Mechanical Engineering from Sharif University of Technology, Tehran, Iran, in 2013. The focus of her research is to develop control algorithms that allow autonomous systems to safely and intelligently interact with each other and with humans. She draws from the fields of control theory, robotics, game theory, and machine learning. She is a recipient of the NSF CAREER Award. She was awarded the IEEE Intelligent Transportation Systems best Ph.D. dissertation award in 2020.

David Fridovich-Keil received the B.S.E. degree in electrical engineering from Princeton University, and the Ph.D. Degree from the University of California, Berkeley. He is an Assistant Professor in the Department of Aerospace Engineering and Engineering Mechanics at the University of Texas at Austin. Fridovich-Keil is the recipient of an NSF Graduate Research Fellowship and an NSF CAREER Award.

Ufuk Topcu received the Ph.D. degree from the University of California at Berkeley, Berkeley, CA, USA, in 2008. He joined the Department of Aerospace Engineering, University of Texas at Austin, Austin, TX, USA, in Fall 2015. He held research positions with the University of Pennsylvania, Philadelphia, PA, USA, and California Institute of Technology, Pasadena, CA, USA. His research focuses on the theoretical, algorithmic and computational aspects of design and verification of autonomous systems through novel connections between formal methods, learning theory, and controls.