LEARNING REWARDS IN STRATEGIC GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the objectives that drive agents' behavior is crucial in strategic environments, where multiple agents interact and adapt to one another. This challenge arises in domains such as cybersecurity, autonomous driving, and online markets, where agents' underlying incentives are typically unknown and must be inferred from their observed behavior. Inverse reinforcement learning (IRL) provides a natural framework for this problem, but extending IRL to strategic multi-agent settings introduces a fundamental challenge: a reward that explains observed behavior does not necessarily induce optimal strategies when used for decision-making. In this work, we study this problem through the lens of *strategic identifiability*, and show that learning from a fixed policy pair can be insufficient to strategically identify rewards, independent on the number of expert samples K. Motivated by this, we introduce an active IRL algorithm for two-player repeated Stackelberg games, in which the leader adaptively selects policies to elicit informative responses about the follower's reward. Combining maximum-likelihood estimation with experimental design, we derive finite-sample guarantees on the Stackelberg sub-optimality induced by the learned reward, matching up to polylogarithmic factors a new minimax learning rate of $K^{-1/2}$. Finally, we evaluate our approach in a realistic anti-poaching setting, demonstrating the benefits of active exploration for reward recovery and downstream strategic performance.[1]

1 INTRODUCTION

Strategic decision-making often requires reasoning about agents whose objectives are unknown. A defender may need to anticipate the incentives of an attacker, an autonomous vehicle the preferences of other drivers, or a wildlife-protection agency the objectives of poachers. In such settings, agents' behavior provides an indirect signal about their underlying incentives. Recovering these objectives is valuable not only for explaining observed behavior, but, crucially, for predicting how agents may respond to new policies or interventions.

Inverse reinforcement learning (IRL) (Ng et al., 2000; Abbeel & Ng, 2004) provides a natural framework for inferring latent reward functions from the behavior of approximately optimal agents. While IRL has been extensively studied in single-agent decision-making, strategic interactions fundamentally change the problem. An agent's observed behavior depends not only on its own reward, but also on the behavior of the agents it interacts with. Consequently, a reward that perfectly explains behavior under one interaction may fail to predict how the same agent responds when its opponent changes strategy. This raises a fundamental question for IRL in games:

when does observed behavior contain enough information about an agent's reward to support reliable strategic decision-making?

Rather than requiring recovery of the ground-truth reward itself, we argue that reward learning should preserve the strategic consequences that matter at deployment. We study this question through the lens of *strategic identifiability*, relating ambiguity in the inferred reward to the sub-optimality incurred when using it for downstream strategic decision-making. We illustrate this objective in Figure 1 for a wildlife protection application, where an error in reward estimation from a ranger patrol can incur highly sub-optimal behavior at deployment.

[1]The code is made available anonymously at https://anonymous.4open.science/r/iclr-learning-rewards/.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

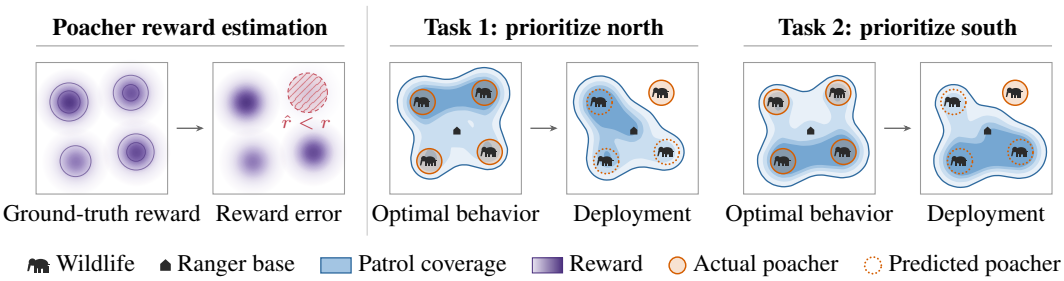

Figure 1: **Illustration of the impact of a poacher reward error on downstream tasks in a wildlife preservation setting.** At deployment, a small error in reward estimation can lead to highly sub-optimal behavior from a rational patrol.

A key difficulty is that passive observations may provide insufficient information for this objective. Observing an agent repeatedly against a fixed policy only reveals its preferences under that particular interaction. Distinct reward functions may therefore generate indistinguishable behavior while inducing substantially different responses to alternative policies. As we show in this work, this can prevent low strategic sub-optimality even with infinitely many observations. This limitation suggests a different approach: rather than passively observing strategic behavior, can we actively choose interactions that reveal the reward information needed for downstream decision-making?

We answer this question in repeated Stackelberg games, where a leader can adaptively select policies and observe the responses of a bounded-rational follower. This setting naturally allows the leader to trade off exploiting its current estimate of the follower's reward with selecting interactions that are informative about the follower's underlying objectives. We develop an active IRL approach that combines maximum-likelihood reward estimation with experimental design to progressively resolve strategically relevant reward ambiguity, and we provide finite-sample guarantees on the strategic performance induced by the learned reward.

Contributions We summarize our main contributions as follows:

- We formalize the objective of IRL in games in terms of the strategic performance induced by the learned reward. We show that observations generated by a fixed policy pair can be insufficient to identify a strategically useful reward, even with infinitely many samples (Lemma 2), while on hard instances any adaptive sampling strategy must suffer a Stackelberg sub-optimality lower bounded by $\Omega(K^{-1/2})$ after K observations (Theorem 1).
- Motivated by this impossibility, we introduce an active IRL framework for repeated Stackelberg normal-form games, in which the leader adaptively selects policies to elicit informative responses from a bounded-rational follower with linear rewards. Our algorithm combines maximum-likelihood reward estimation with experimental design and achieves a high-probabilty Stackelberg sub-optimality of order $\tilde{\mathcal{O}}(d/\sqrt{K})$ after K interactions where d is the dimensionality of the follower features (Theorem 3). This matches the optimal dependence on K up to polylogarithmic factors.
- We then evaluate our approach on synthetic normal-form games and a realistic wildlife-protection environment. Our experiments demonstrate the benefits of actively selecting interactions for reward recovery and downstream strategic performance.

Related work Our work connects inverse reinforcement learning and learning in games. In single-agent settings, IRL has long studied the ambiguity inherent in recovering rewards from demonstrations (Ng et al., 2000; Abbeel & Ng, 2004). In multi-agent settings, recent work has focused on whether policies learned from demonstrations remain robust when interacting with strategic agents, typically through notions of exploitability (Tang et al., 2024; Freihaut et al., 2026a; Bergerault et al., 2026a).

Emerging work on IRL in games has instead characterized sets of rewards compatible with observed strategic behavior (Freihaut & Ramponi, 2026; Barbara et al., 2026). We take a complementary perspective: rather than only characterizing which rewards are consistent with observed behavior, we

ask which interactions provide the information needed to identify a strategically useful reward, and how such interactions can be actively selected with finite-sample guarantees.

A more extensive discussion on previous work can be found in Appendix A.

2 PRELIMINARIES AND FORMAL SETUP

Notation For the rest of the paper, we will use the following notations. For any measurable space $\mathcal{X}$, $\Delta(\mathcal{X})$ denotes the probability simplex over $\mathcal{X}$. Moreover, we define the Shannon entropy $\mathcal{H}_{\mathcal{X}} : \Delta(\mathcal{X}) \to \mathbb{R}$ as $\mathcal{H}_{\mathcal{X}}(P) = -\int \log(P(x))\,\mathrm{d}P(x)$, and we will use the simplified notation $\mathcal{H}$ when $\mathcal{X}$ is clear from context. For any positive definite matrix $P \in \mathbb{R}^{n\times n}$ and any $x \in \mathbb{R}^n$, we denote its norm with respect to P as $\|x\|_P = \sqrt{x^\top P x}$. Finally, for any positive $N \in \mathbb{Z}_+$, we use the notation $[N] \coloneqq \{1, \dots, N\}$, and for any finite set $\mathcal{S}$ we denote by $|\mathcal{S}|$ its cardinality.

We define a two-player normal-form game as a tuple $\mathcal{M} \coloneqq (\mathcal{A}, \mathcal{B}, r_1, r_2)$ where $\mathcal{A}, \mathcal{B}$ represent the finite action spaces of the two players, and for $i \in \{1, 2\}$, we denote by $r_i : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$, $r_i \in \mathcal{R}_i \subseteq \mathbb{R}^{\mathcal{A}\times\mathcal{B}}$ the individual rewards. We denote by $\mu \in \Delta(\mathcal{A})$ and $\nu \in \Delta(\mathcal{B})$ the policies of the two players, such that joint actions are sampled from the product distribution $\mu \otimes \nu \in \Delta(\mathcal{A} \times \mathcal{B})$.

In this environment, player i aims to maximize the objective $J_{r_i}(\mu, \nu) = \mathbb{E}_{(a,b)\sim(\mu,\nu)}\left[r_i(a, b)\right]$ such that given μ, player 2 plays a best-response $\nu^*_\mu(r_2) \in \arg\max_{\tilde{\nu}} J_{r_2}(\mu, \tilde{\nu})$, similarly for the μ-player.

2.1 MULTI-AGENT LEARNING OBJECTIVES

Several metrics can be used to represent the sub-optimality of learning players in games.

Adapted from the single-agent learning literature (Cesa-Bianchi et al., 2006; Orabona, 2019), the *regret* measures the opportunity cost of each agent in the environment induced by the other player.

Definition 1 (Regret)**.** *We define the (external) regret of the players as follows:*

$$\mathrm{Reg}^1_r(\mu, \nu) \coloneqq \max_{\mu^*} J_{r_1}(\mu^*, \nu) - J_{r_1}(\mu, \nu), \quad \mathrm{Reg}^2_r(\mu, \nu) \coloneqq \max_{\nu^*} J_{r_2}(\mu, \nu^*) - J_{r_2}(\mu, \nu).$$

Following Ramponi et al. (2023); Tang et al. (2024), recent multi-agent imitation learning works often aggregate these regrets as[2] $\mathrm{NashGap}_r(\mu, \nu) = \max\left\{\mathrm{Reg}^1_r(\mu, \nu), \mathrm{Reg}^2_r(\mu, \nu)\right\}$ such that each pair (μ, ν) satisfying $\mathrm{NashGap}_r(\mu, \nu) \leq \epsilon$ is an ϵ-Nash equilibrium (Nash Jr, 1950).

This is appealing for two main reasons. First, when offline methods are used as a bootstrap to further policy optimization (Macuglia et al., 2025; Wagenmaker et al., 2025) seeking a Nash equilibrium. Second, as a deployment guarantee with a fixed opponent, i.e. how much a player could improve by deviating when the other players are deployed *without adapting to these deviations*.

However, when learning and deploying in the presence of adaptive strategic players, regret notions are *oblivious*, as they ignore how other agents will react to a change of policy (Mertikopoulos, 2026). To better measure this strategic coupling between the agents, we therefore take the perspective of a Stackelberg setting (Sessa et al., 2020; Bai et al., 2021; Chen et al., 2025).

Under this view, a leader starts by committing to a policy μ, based on which a follower decides on its own policy ν. Given μ, we assume that the follower is optimal by choosing $\nu = \nu^*_\mu(r_2)$[3], i.e. the follower *best-responds* to μ with respect to its reward r_2. The goal of the leader is then to determine a policy $\mu^* \in \arg\max_{\tilde{\mu}} J^{\text{leader}}_r(\tilde{\mu})$ maximizing its expected rewards against an optimal follower:

$$J^{\text{leader}}_r(\mu) \coloneqq J_{r_1}(\mu, \nu^*_\mu(r_2)).$$

From this formulation, we define our measure of sub-optimality as follows.

Definition 2 (Stackelberg sub-optimality)**.** *Given a leader policy μ, we define its Stackelberg sub-optimality with respect to the reward r as:*

$$\mathrm{SubOpt}_r(\mu) \coloneqq \max_{\mu^*} J^{\textit{leader}}_r(\mu^*) - J^{\textit{leader}}_r(\mu).$$

[2] Or similarly $\mathrm{NashGap}_r(\mu, \nu) = \mathrm{Reg}^1_r(\mu, \nu) + \mathrm{Reg}^2_r(\mu, \nu)$.

[3] Here we assume the best-response of the follower is unique to simplify notations. This is true under entropy-regularization of the follower, assumed later (Assumption 2).

Compared to $\mathrm{Reg}_r^1(\mu, \nu)$, this objective no longer measures the learner's sub-optimality against a fixed ν, but rather how the μ-player can improve when facing an adaptive follower.

Given K episodes, we state the goal of an inverse reinforcement learning algorithm as finding an approximate reward $\hat{r}_2$ such that $\mathrm{SubOpt}_r(\hat{\mu})$ is minimal for $\hat{\mu} \in \arg\max_{\tilde{\mu}} J^{\text{leader}}_{(r_1, \hat{r}_2)}(\tilde{\mu})$. In other words, an optimal leader for $\hat{r}_2$ should be close to optimal under the true reward r_2.

2.2 Modeling assumptions

For finite action spaces, $\mathcal{R}_2$ is in general the entire set of tabular rewards. However, in order to better model the reward function, we make the following popular modeling assumption.

Assumption 1 (Linear rewards)**.** *There exists a known feature map $\phi : \mathcal{A} \times \mathcal{B} \to \mathbb{R}^d$, $S > 0$ and a parameter $\omega \in \Omega \coloneqq \left\{\tilde{\omega} \in \mathbb{R}^d : \|\tilde{\omega}\|_2 \leq S\right\}$ for some $d \in \mathbb{Z}_+$ such that:*

$$r_2(a, b) = \langle \phi(a, b), \omega \rangle, \qquad \forall (a, b) \in \mathcal{A} \times \mathcal{B}.$$

Additionally, we assume the features are normalized, i.e. $\max_{a,b} \|\phi(a, b)\|_2 \leq 1$.[4]

The problem of IRL therefore reduces to estimating the parameter ω from observed interactions. This precise structure of r_2 is interpretable (the feature map is known) and allows our algorithm to optimize the information on the reward function via a Fisher information matrix on its coefficients.

As a consequence, we model our estimate $\hat{r}_2$ as a linear map with features ϕ and parameter $\hat{\omega}$, and will use SubOpt_ω and $\mathrm{SubOpt}_{\hat{\omega}}$ to denote SubOpt_r and $\mathrm{SubOpt}_{(r_1, \hat{r}_2)}$, respectively.

Remark 1. *Any reward can be linearly parametrized with features being the canonical basis of $\mathbb{R}^{|\mathcal{A}| \cdot |\mathcal{B}|}$ and coefficients $\omega_{ab} = r_2(a, b)$, therefore satisfying Assumption 1 with $d = |\mathcal{A}| \cdot |\mathcal{B}|$. In general, d can be much smaller with a better choice of features.*

Compared to the tabular setting, Assumption 1 therefore allows to learn at a faster statistical rate, as a single interaction can carry more information on the underlying reward.

Additionally, we assume an entropy regularized follower. This common assumption (Fu et al., 2017; Geist et al., 2019) is motivated by *bounded rationality* in human decision-making and policies.

Assumption 2 (Bounded rationality of the ν-player)**.** *The ν-player is entropy-regularized with inverse temperature parameter $\eta > 0$, i.e. $\nu^*_\mu(\tilde{r}) = \arg\max_{\tilde{\nu}} J_{\tilde{r}}(\mu, \tilde{\nu}) + \eta^{-1}\mathcal{H}(\tilde{\nu})$.*

To make the setting learnable, we ignore the trivial case where the ν-player is completely independent of its reward[5].

Assumption 3 (Non-degeneracy)**.** *There exists $\mu \in \Delta(\mathcal{A})$ and two linear rewards $r_2, \tilde{r}_2 \in \mathcal{R}_2$ with feature map ϕ such that $\nu^*_\mu(r_2) \neq \nu^*_\mu(\tilde{r}_2)$.*

For notational convenience, we also define the bound on the μ-player rewards as follows. Note that $R_{\max}$ is always well-defined since the action spaces are finite.

Definition 3. *We denote $R_{max} \coloneqq \|r_1\|_\infty$, i.e. $R_{max} = \max_{a \in \mathcal{A},\, b \in \mathcal{B}} |r_1(a, b)|$.*

In the next section, we introduce the problem of reward identifiability and its geometric implications for learning from a fixed policy pair. Moving to the statistical perspective, we further derive a lower bound on the sub-optimality of any adaptive algorithm in the finite sample regime. Then, in Section 4, we derive a principled adaptive algorithm (Algorithm 1) for Stackelberg sub-optimality minimization in MA-IRL that matches our lower bound up to poly-logarithmic factors.

3 On reward identifiability in games

The goal of inverse reinforcement learning is to leverage demonstrations to extract a reward function inducing similar behaviors, and identifiability measures the ability of recovering such reward. The strongest notion aims to extract the exact reward $\hat{r}_2 = r_2$, whereas in multi-agent applications it would be enough to preserve strategic behaviors defined from a given learning objective (see Section 2.1).

[4] This is without loss of generality. Indeed, for any $R > 1$ such that $\max_{a,b} \|\phi(a, b)\|_2 = R$, one can always equivalently use the normalized features $\tilde{\phi} = \phi / R$ with the inverse temperature $\tilde{\eta} = \eta R$ (see Assumption 2).

[5] See Appendix B.1 how to make that explicit in terms of the features.

In order to learn reward information, the approach relies on the assumption of some (approximately) optimal players. IRL for the ν-player therefore aims to find a function $\hat{r}_2$ that belongs to the feasible reward set $\mathcal{R}_2^{\text{feas}}(\mu, \nu_\mu^*) \coloneqq \left\{\tilde{r}_2 \in \mathcal{R}_2 : \nu_\mu^* = \nu_\mu^*(\tilde{r}_2)\right\}$ for any sampled policy pair (μ, ν_μ^*), where sub-optimality can be measured using different learning objectives.

In the entropy-regularization setting, where the follower is boundedly-rational by trading-off between rewards and entropy, the relation between the best-response and the reward of the ν-player can be made explicit. In fact, a classical result (see e.g. Geist et al. (2019)) is the following:

Lemma 1. *Under Assumption 2, the unique best-response ν_μ^* to the policy μ is such that:*

$$\nu_\mu^*(b) = \frac{\exp\left(\eta \sum_{a\in\mathcal{A}} \mu(a) r_2(a,b)\right)}{\sum_{b'\in\mathcal{B}} \exp\left(\eta \sum_{a\in\mathcal{A}} \mu(a) r_2(a,b')\right)}, \qquad \forall b \in \mathcal{B}. \tag{1}$$

Hence, the average reward $r_2^\mu(b) \coloneqq \langle \mu, r_2(\cdot, b)\rangle$ such that $\nu_\mu^* = \mathrm{softmax}(\eta \cdot r_2^\mu)$ can be learned by inverting this multinomial logistic model, but only up to constant shifts (Freihaut & Ramponi, 2026).

Under Assumption 1 of a linear follower reward, this relation can be simplified. Specifically, letting $\psi_\mu(b) = \sum_{a\in\mathcal{A}} \phi(a,b)\mu(a)$ be the average feature vector of the ν-player in the environment induced by μ, Liao et al. (2025) notes that ω is solution to the following linear system:

$$\underbrace{\begin{bmatrix} \psi_\mu(b_2)^\top - \psi_\mu(b_1)^\top \\ \vdots \\ \psi_\mu(b_{|\mathcal{B}|})^\top - \psi_\mu(b_1)^\top \end{bmatrix}}_{=:\Phi(\mu)\in\mathbb{R}^{(|\mathcal{B}|-1)\times d}} \omega = \eta^{-1} \begin{bmatrix} \log\left(\nu_\mu^*(b_2)/\nu_\mu^*(b_1)\right) \\ \vdots \\ \log\left(\nu_\mu^*(b_{|\mathcal{B}|})/\nu_\mu^*(b_1)\right) \end{bmatrix}. \tag{2}$$

From this relation, we can study different notions of identifiability for the reward function.

Recovering the exact reward function. As noticed by Liao et al. (2025); Rolland et al. (2022); Freihaut & Ramponi (2026), the linear system in Equation (2) is invertible if and only if $\mathrm{rank}(\Phi(\mu)) = d$. From basic linear algebra, one can already conclude that the reward cannot be recovered if $|\mathcal{B}|-1 < d$. Intuitively, this holds because there always exists a non-zero $\omega_0 \in \mathbb{R}^d$ such that $\psi_\mu(b)^\top\omega_0$ is constant over all b, hence the constant shift ambiguity is not removed by the linear structure of Assumption 1.

When the rank condition holds, previous work such as Liao et al. (2025) learn the ordinary least squares solution of the system, estimating the target vector from samples. The final error bound on $\|\omega - \hat{\omega}\|_2$ relies on concentration results scaling with the inverse minimum probability of actions, of order $|\mathcal{B}|e^{2\eta S}$ in the worst-case. Crucially, this bound is in the standard Euclidean geometry and not weighted by the directions spanned by the features. If the rank condition is not satisfied, such a bound might never vanish and the standard workaround is to study convergence in Hausdorff distance between the true and recovered feasible sets (Metelli et al., 2021; Liao et al., 2025; Freihaut & Ramponi, 2026; Barbara et al., 2026).

In this work, we weaken this goal of exact identifiability and focus on learning a reward $\hat{r}_2$ (approximately) preserving the strategic behaviors of the agents[6].

Strategic identifiability of the reward. Toward this goal, we focus on learning a reward inducing (approximately) the same best-response followers as the true reward. Indeed, as we assume the ν-player to best-respond during training, we also assume this capability at deployment, and therefore take the Stackelberg sub-optimality SubOpt_r instead of Reg_r^1 as the most natural strategic measure for the learned policy.

Learning from a fixed policy pair, exact reward identifiability is impossible when the linear system of Equation (2) is not invertible. We extend below this observation to strategic identifiability, where no choice of μ allows to arbitrarily minimize sub-optimality. We report the proof in Appendix C.

Lemma 2. *Fix $\eta > 0$. There exists a set $(r_1, \mathcal{A}, \mathcal{B}, \phi)$ such that for any reference policy μ_{ref} and any number of samples K, no algorithm only sampling from $(\mu_{ref}, \nu_{\mu_{ref}}^*)$ can guarantee that $\mathrm{SubOpt}_r(\hat{\mu}) \le c_\eta$ with probability greater than $1/2$. $c_\eta > 0$ only depends on η.*

[6]On the contrary, previous algorithms with Hausdorff distance guarantees focus on defining a set that for all plausible rewards *contains* a good approximation of it, and inversely.

The above lemma is a geometric artifact built on the same degeneracy as exact non-identifiability. This degeneracy makes it impossible to learn information that could help distinguish between two, strategically unequal coefficients, even in the infinite-data regime. Observations from a passive leader can therefore be insufficient to recover strategy-preserving rewards.

This motivates the need for active interaction strategies where the learner deliberately selects different leader policies to probe the follower in useful directions. In particular, these adaptively chosen directions should be both informative and strategically important for the reward coefficients. Under this framing, learning rewards remains constrained by the following statistical lower bound.

Theorem 1 (Minimax lower bound)**.** *Let $d \geq 32$, $\epsilon = \frac{1}{2\eta}\sqrt{d/K} \leq S$ and $\omega^{(i)} \coloneqq \varepsilon e_i$ for $i \in [d]$. There exists a game such that for every algorithm producing an output $\hat{\mu}$ after $K \geq 4d$ interactions,*

$$\max_{i\in[d]} \mathbb{E}^{(i)}\left[\mathrm{SubOpt}_{\omega^{(i)}}(\hat{\mu})\right] \;\geq\; \frac{1}{44} R_{\max}\sqrt{\frac{d}{K}},$$

where $\mathbb{E}^{(i)}$ is for the expectation under the trajectory law $\mathbb{P}^{(i)}$ induced by $\omega^{(i)}$.

In the next section, we show how to achieve such a learning rate up to poly-logarithmic factors.

4 Active Stackelberg IRL in normal-form games

In many applications, indeed, one can *actively* interact with the environment over time. Examples include cybersecurity settings where defenders adapt to strategic attackers and online platforms interacting with self-interested users. In those cases, the learner is not restricted to passively observing a fixed behavior, but can deliberately choose policies inducing diverse environments for the ν-player.

This allows the learner to elicit informative responses from the follower, related in spirit to the strategy of Rolland et al. (2022), who modify the dynamics of a single-agent setting to make the reward exactly identifiable from expert demonstrations. Similarly, in our setting, varying the leader policies helps reduce estimation uncertainty in strategically relevant directions of the follower's reward function.

Specifically, we consider a leader who interacts repeatedly with a (boundedly-)rational follower, for a total of N rounds. At each round $\tau \in [N]$, the leader commits to a (potentially mixed) strategy $\mu_\tau \in \Delta(\mathcal{A})$, and observes a sample $b_\tau \sim \nu_\tau$ from the follower's unique best-response $\nu_\tau \coloneqq \nu^*_{\mu_\tau}(r_2)$. After the N iterations, the goal is to return an estimate $\hat{\omega}$ such as to minimize $\mathrm{SubOpt}_\omega(\hat{\mu})$ where $\hat{\mu}$ is optimal under the estimated reward $\hat{r}_2(\cdot,\cdot) = \langle\phi(\cdot,\cdot),\hat{\omega}\rangle$, i.e. $\mathrm{SubOpt}_{\hat{\omega}}(\hat{\mu}) = 0$.

Denoting the average features $\psi_{\mu_\tau}(b) = \sum_a \mu_\tau(a)\phi(a,b)$, Lemma 1 gives for all $b \in \mathcal{B}$,

$$\nu_\tau(b) = \frac{1}{Z_{\mu_\tau}(\omega)} \exp\left(\eta\langle\psi_{\mu_\tau}(b),\omega\rangle\right), \qquad Z_{\mu_\tau}(\tilde{\omega}) \coloneqq \sum_{b'\in\mathcal{B}} \exp\left(\eta\langle\psi_{\mu_\tau}(b'),\tilde{\omega}\rangle\right), \tag{3}$$

from which one can derive likelihood values for the ω. Our approach is based on this idea.

4.1 Combining optimal design and maximum likelihood estimation

For any history of n pairs $\mathcal{D} = \{(b_\tau,\mu_\tau)\}_{\tau=1}^n$, let us denote the log-likelihood of $\tilde{\omega} \in \Omega$ as

$$\mathcal{L}(\tilde{\omega}|\mathcal{D}) \coloneqq \sum_{\tau=1}^n l(\tilde{\omega}|b_\tau,\mu_\tau), \qquad l(\tilde{\omega}|b_\tau,\mu_\tau) \coloneqq \log\mathbb{P}(b_\tau|\omega=\tilde{\omega},\mu_\tau),$$

with the instantaneous log-likelihood $l(\tilde{\omega}|b_\tau,\mu_\tau) = \eta\,\langle\psi_{\mu_\tau}(b_\tau),\tilde{\omega}\rangle - \log Z_{\mu_\tau}(\tilde{\omega})$. With the exponential family property of Equation (3), the Fisher information matrix of μ_τ also simplifies[7]:

$$\mathcal{I}_{\mu_\tau}(\tilde{\omega}) \coloneqq -\nabla^2 l(\tilde{\omega}|\mu_\tau) = \eta^2\mathrm{Cov}_{b\sim\mathbb{P}(\cdot|\tilde{\omega},\mu_k)}\left[\psi_{\mu_k}(b)\right],$$

such that $-\nabla^2\mathcal{L}(\tilde{\omega}|\mathcal{D}) = \sum_{\tau=1}^n \mathcal{I}_{\mu_\tau}(\tilde{\omega})$ is the cumulated Fisher information of $\mathcal{D}$.

Following these remarks, we derive an algorithm reducing uncertainty in the Fisher information geometry, by using this relation on the sub-optimality, whose proof is deferred to Appendix F.1.

[7] $\nabla^2 l(\tilde{\omega}|b,\mu_\tau)$ is constant over all $b \in \mathcal{B}$, we note that term $\nabla^2 l(\tilde{\omega}|\mu_\tau)$.

Lemma 3 (Reward transfer lemma). *Let μ^* and $\hat{\mu}$ be optimal leaders under ω and $\hat{\omega}$. Then,*

$$\mathrm{SubOpt}_\omega(\hat{\mu}) \le 2R_{max} \cdot \max_{\tilde{\mu}\in\{\mu^*,\hat{\mu}\}} \chi\left(\eta \cdot d_{\tilde{\mu}}(\omega,\hat{\omega})\right) \cdot \|\omega-\hat{\omega}\|_{\mathcal{I}_{\tilde{\mu}}(\hat{\omega})},$$

with $\chi(x) \coloneqq 2x^{-1}\left(e^{x/2}-1\right)$ for $x>0$ and $d_{\tilde{\mu}}(\omega,\hat{\omega}) = \max_{b,d}\langle \psi_{\tilde{\mu}}(b)-\psi_{\tilde{\mu}}(d), \omega-\hat{\omega}\rangle$.

Essentially, Lemma 3 shows how the sub-optimality aligns with an estimation problem in the Fisher information geometry. This alignment depends on $d_{\tilde{\mu}}(\omega,\hat{\omega})$, which represents the span semi-norm of the difference between logits induced by ω and $\hat{\omega}$.

We optimize this upper bound by combining three ingredients, presented in the proof outline below.

Proof roadmap

Analyzing the second term, for any $\lambda>0$ and $\mathcal{D}\in(\mathcal{B}\times\Delta(\mathcal{A}))^n$, we change the measure as:

$$\|\omega-\hat{\omega}\|^2_{\mathcal{I}_{\tilde{\mu}}(\hat{\omega})} \le \underbrace{\|\omega-\hat{\omega}\|^2_{G^\lambda(\hat{\omega})}}_{\text{estimation}} \cdot \underbrace{\mathrm{Trace}\left(\left[G^\lambda(\hat{\omega})\right]^{-1}\mathcal{I}_{\tilde{\mu}}(\hat{\omega})\right)}_{\text{Fisher uncertainty}},$$

for the regularized $G^\lambda(\tilde{\omega}|\mathcal{D}) \coloneqq \lambda I - \nabla^2\mathcal{L}(\tilde{\omega}|\mathcal{D})$.

We optimize these terms with Maximum Likelihood Estimators (MLE) and optimal design.

(i) *Estimation.* Specifically, for $\{\hat{\omega}_\tau \coloneqq \arg\max_{\tilde{\omega}\in\Omega}\mathcal{L}(\tilde{\omega}|\mathcal{D}_{1:\tau})\}$[8]the constrained MLE sequence, the estimation is controlled using the following confidence result from Lee et al. (2024).

Theorem 2 (Informal version of Theorem 5). *With probability at least $1-\delta$,*

$$\|\omega-\hat{\omega}_\tau\|^2_{\Gamma^\lambda_\tau(\hat{\omega}_\tau)} \le \mathcal{O}\left(\lambda S^2 + \log\delta^{-1} + d\log\left(\max\left\{e, d^{-1}\tau\eta S\right\}\right)\right), \quad \forall\tau\in[n],$$

where we define $\Gamma^\lambda_\tau(\tilde{\omega}) \coloneqq \int_0^1 (1-s)G^\lambda(\tilde{\omega}+s(\omega-\tilde{\omega})|\mathcal{D}_{1:\tau})\,\mathrm{d}s$ and fix $\delta\in(0,1)$.

(ii) *Design.* This holds independently of the policies in $\mathcal{D}$, that we therefore choose to minimize the Fisher uncertainty, or equivalently maximize the Fisher coverage and select:

$$\mu_\tau \in \arg\max_{\mu\in\Delta(\mathcal{A})} \mathrm{Trace}\left(\left[G^\lambda(\hat{\omega}_{k-1}|\mathcal{D}_{1:\tau})\right]^{-1}\mathcal{I}_\mu(\hat{\omega}_{\tau-1})\right),$$

i.e. greedily optimize the *information gain* at each step.

(iii) *Self-concorrdance and warm-up.* We combine everything together by deriving new self-concordance properties for Fisher information matrices and gains expressed in terms of $d_{\tilde{\mu}}$, which can be of independent interest.

To force self-concordance into a harmless regime, we adapt a pure exploration warm-up strategy by generalizing Kiefer-Wolfowitz results to our setting where the leader only has partial control over the feature coverage. Specifically, we collect n_{warm} samples from allocation computed offline, and subsequently perform $K \coloneqq N - n_{\text{warm}}$ adaptive policy selections with the cumulative data:

$$\mathcal{D}_k \coloneqq \mathcal{D}_{1:n_{\text{warm}}+k} = \mathcal{D}_{\text{warm}} \cup \{(b^{\text{on}}_{\tau'}, \mu^{\text{on}}_{\tau'})\}^k_{\tau'=1}, \qquad \forall k\in[K],$$

where $\{(b^{\text{on}}_{\tau'}, \mu^{\text{on}}_{\tau'})\}^k_{\tau'=1}$ are collected online and $\mathcal{D}_{\text{warm}} \coloneqq \mathcal{D}_{1:n_{\text{warm}}}$ is from the warm-up phase.

This warm-up guarantees that all MLE terms $\{\hat{\omega}^{\text{on}}_k \coloneqq \hat{\omega}_{n_{\text{warm}}+k}\}_{k\le K}$ satisfy $\max_\mu d_\mu(\omega,\hat{\omega}^{\text{on}}_k) \le 1/\eta$ simultaneously with high probability. This also bounds the first term of Lemma 3 and concludes the analysis.

4.2 An active algorithm with warm-up

As mentioned in the proof outline and in contrast to standard exploration that would maximize a bonus of the form $\|\phi(\cdot,\cdot)\|_{(\lambda I+\Lambda_k)^{-1}}$ with $\Lambda_k \coloneqq \sum_{\tau=1}^k\left[\psi_{\mu_\tau}(b_\tau)\psi_{\mu_\tau}(b_\tau)^\top\right]$ (e.g. LSVI-UCB (Jin et al., 2020)), we design our objective around minimizing uncertainty in *strategically-relevant directions*, measured by the Fisher information (Lemma 3). We summarize our approach in Algorithm 1.

[6]We define the truncated history as $\mathcal{D}_{1:j} \coloneqq \{(b_\tau,\mu_\tau)\}^j_{\tau=1}, \forall j\in[n]$.

Algorithm 1 Active Stackelberg IRL for two-player normal-form games

Input: horizon $K \geq 0$, confidence δ, burn-in budget $n_{\text{warm}} \geq 0$, regularizer $\lambda \geq 0$, tolerance ϵ.

Warm-up phase (simplified, see the full procedure as Algorithm 2 in Appendix E)

Solve the following optimal design up to a factor $1+\epsilon$, using proxy $\tilde{\omega} = 0$:

$$w \in \arg\min_{\tilde{w} \in \Delta(\mathcal{A})} \max_{a \in \mathcal{A}} \operatorname{Trace}\left(W(\tilde{w})^{\dagger} \mathcal{I}_{\delta_a}(0)\right), \quad \text{s.t.} \quad \operatorname{Im} \mathcal{I}_{\delta_a}(0) \subseteq \operatorname{Im} W(\tilde{w}) \ \ \forall a \in \mathcal{A},$$

where δ_a is the pure policy on a and $W(\tilde{w}) := \mathbb{E}_{a' \sim \tilde{w}}[\mathcal{I}_{\delta_{a'}}(0)]$.

Round w to an allocation of n_{warm} policies and collect $\mathcal{D}_{\text{warm}}$.

Adaptive online phase

$\mathcal{D}_0 := \mathcal{D}_{\text{warm}}$, $\hat{\omega}_0^{\text{on}} \leftarrow \arg\max_{\tilde{\omega} \in \Omega} \mathcal{L}(\tilde{\omega} | \mathcal{D}_{\text{warm}})$ or arbitrary in Ω if $\mathcal{D}_{\text{warm}} = \emptyset$.

for $k = 1, 2, \ldots, K$ **do**

$\mu_k^{\text{on}} \in \arg\max_{\mu \in \Delta(\mathcal{A})} \operatorname{Trace}\left(\left[G^{\lambda}(\hat{\omega}_{k-1}^{\text{on}} | \mathcal{D}_{k-1})\right]^{-1} \mathcal{I}_{\mu}(\hat{\omega}_{k-1}^{\text{on}})\right)$[9]

$b_k \sim \nu^*_{\mu_k^{\text{on}}}(\omega)$; $\mathcal{D}_k \leftarrow \mathcal{D}_{k-1} \cup \{(b_k, \mu_k^{\text{on}})\}$; $\hat{\omega}_k^{\text{on}} \leftarrow \arg\max_{\tilde{\omega} \in \Omega} \mathcal{L}(\tilde{\omega} | \mathcal{D}_k)$.

end for

Output: $\hat{\omega} := \hat{\omega}_K^{\text{on}}$.

After an initial burn-in phase with $n_{\text{warm}} \approx \tilde{\mathcal{O}}\left(\eta S e^{2\eta S} |\mathcal{B}| d^2 \log \delta^{-1}\right)$, we derive the following finite-sample Stackelberg sub-optimality guarantees.

Theorem 3 (Informal)**.** *Let $\hat{\omega}$ the result of Algorithm 1 with $n_{warm} \approx \tilde{\mathcal{O}}\left(\eta S e^{2\eta S} |\mathcal{B}| d^2 \log \delta^{-1}\right)$, $K \geq 1$ and $\lambda = 0$. Then, with probability at least $1-\delta$, for any $\hat{\mu}$ such that $\operatorname{SubOpt}_{\hat{\omega}}(\hat{\mu}) = 0$:*

$$\operatorname{SubOpt}_{\omega}(\hat{\mu}) \leq \tilde{\mathcal{O}}\left(\frac{R_{max}\sqrt{d\left(d\eta S + \log \delta^{-1} + d\right)}}{\sqrt{K}}\right),$$

where $\tilde{\mathcal{O}}$ hides polylog terms in $\log \delta^{-1}$, K and environment parameters.

The proof follows the following thread: after an initial warm-up phase, we ensure that every online MLE is sufficiently localized near the true parameter ω. Given this, we proceed by a greedy online phase using the running estimators in place of the true unknown reward coefficient. Finally, we conclude with a classical elliptical potential bound argument. The formal version of Theorem 3 and its complete proof are in Appendix F.2.

Remark 2. *Our main result assumes an initial warm-up phase for n_{warm} exponential in the parameters η and S. This corresponds to the usual κ-dependency of the various previous works on preference-based learning and bandits with the logistic function (Faury et al., 2020; Abeille et al., 2021; Schlaginhaufen et al., 2026; Bergerault et al., 2026b).*

While Theorem 3 is given for a specific n_{warm} and $\lambda = 0$, we note that in general and in practice, our algorithm can be run with smaller warm-up steps (or no warm-up at all), requiring $\lambda > 0$ for proper matrix inversion in the policy selection.

Lastly, by a standard reduction argument, this PAC result also yields the following:

$$\mathbb{E}\left[\operatorname{SubOpt}_{\omega}(\hat{\mu})\right] \leq \tilde{\mathcal{O}}\left(R_{\max}\sqrt{d/K}\right),$$

where the expectation is taken with respect to the follower best-responses.

This matches the lower bound in Theorem 1 up to polylogarithmic factors for sufficiently large K.

5 Numerical validation

We empirically validate our algorithm on a realistic anti-poaching task from the `GUARD` benchmark (Krever et al., 2026). We simulate both active methods where the controllable leader interacts with

[9] If $\lambda = 0$ and the inverse does not exist, this operation can be carried over a subspace instead. This requires some care, but the intuition remains (see Appendix F).

a best-response follower and fixed-expert methods. These fixed-expert baselines rely on the strong assumption of a known Stackelberg or quantal-response equilibrium oracle that we sample from.

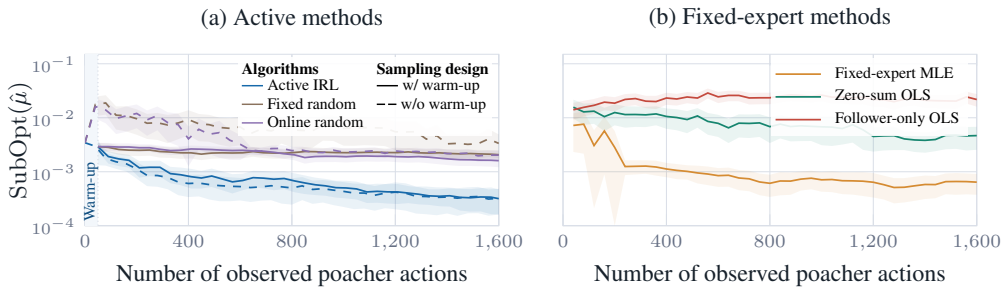

Figure 2: **Mean** SubOpt **with the number of poacher actions.** 95% CIs over 50 seeds.

In Figure 2, we compare our algorithm `Active IRL` with two active MLE baselines where a policy is sampled randomly at each step (`Online random`) or once at the beginning of training then fixed (`Fixed random`). Additionally, we compare to previous (fixed-expert) works with `Zero-sum OLS`[10] (derived from Liao et al. (2025)), `Follower-only OLS` and the version of MaxEntIRL in our setting as `Fixed-expert MLE` (derived from Ziebart et al. (2008)).

As shown, `Active IRL` approximately matches or improves over all the baselines, despite the structural advantage of fixed-expert methods. In terms of *transfer* of the learned rewards to multiple leader objectives, we provide some illustrative results in Figure 3 below for arbitrary downstream tasks[11], comparing our algorithm with the second-to-best algorithm (`Fixed-expert MLE`).

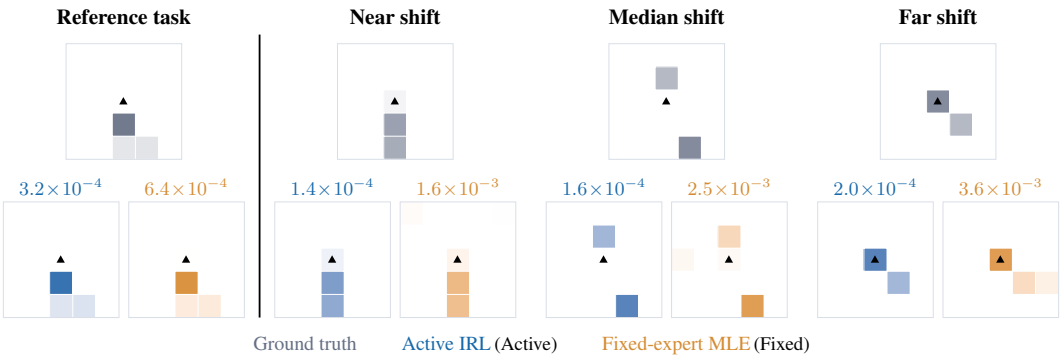

Figure 3: **Patrol coverage and associated mean** SubOpt **across downstream tasks.** Trained on 1,600 poacher actions. Darker map cells indicate greater coverage; the triangle marks the ranger base.

We provide additional details, experiments and ablations in Appendix H.

6 Conclusion

We study IRL for interactive agents in strategic games as the problem of learning reward functions that (approximately) preserve optimality of agents when deployed among strategic players. In two-player normal-form games under linearity and entropy-regularization assumptions, we develop an active algorithm that provably minimizes Stackelberg sub-optimality at a minimax optimal rate, ignoring polylogarithmic factors. We then empirically validate our algorithm on a realistic anti-poaching game and a set of synthetic environments.

Future directions A current open problem is principled learning in dynamic games. Specifically, open challenges include the derivation of exploration strategies over states and the computation of reward estimators from trajectories (beyond strong myopic assumptions of previous works).

[10] Learning from two times more samples by leveraging the known zero-sum structure.

[11] Selected to induce diverse variations of the reference distribution, independently of algorithms performance.

AI USE STATEMENT

In this work, we used generative AI tools for assistance in the writing of proofs and the experimental validation (implementing methods, setting up experiments). We have not used generative AI tools for generating synthetic datasets, developing the conceptual frameworks, designing the methodology of this work and interpreting the results, and thematic data analysis, the formulation of hypotheses and translation/reformatting are not applicable to this work. Additionally, we used generative AI tools to create or edit software code, including the creation of visualization and scientific plots. We have reviewed all AI-assisted work, comprising proofs and software, while we did not use AI to draft or structure this paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The experimental validation of this paper is performed using the public `GUARD` benchmark (Krever et al., 2026) as well as small synthetic experiments. The results and setup details are provided in Section 5 and in the additional experiments section of the appendix (Appendix H). The entire codebase with reproducibility scripts will be made public upon release of this paper.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3691–3699. PMLR, 2021.

Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal discrete optimization for experimental design: A regret minimization approach. *Mathematical Programming*, 186(1): 439–478, 2021.

Zhiyu An and Wan Du. Diml: Differentiable inverse mechanism learning from behaviors of multi-agent learning trajectories. *arXiv preprint arXiv:2601.17678*, 2026.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384 – 414, 2010. doi: 10.1214/09-EJS521.

Yu Bai, Chi Jin, Huan Wang, and Caiming Xiong. Sample-efficient learning of stackelberg equilibria in general-sum games. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21. Curran Associates Inc., 2021. ISBN 9781713845393.

Annalisa Barbara, Riccardo Poiani, Martino Bernasconi, and Andrea Celli. Optimal rates for feasible payoff set estimation in games. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=qV7yFzdK1R`.

Antoine Bergerault, Volkan Cevher, and Negar Mehr. Matching multiple experts: On the exploitability of multi-agent imitation learning. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL `https://openreview.net/forum?id=B9qYvPTbTN`.

Antoine Bergerault, Anna Maria Maddux, Andreas Schlaginhaufen, and Maryam Kamgarpour. Preference-based distributed welfare maximization: A game-theoretic approach. In *ICML 2026 Workshop on New Frontiers in Game-Theoretic Learning (NExT-Game)*, 2026b. URL `https://openreview.net/forum?id=CAS7aInVos`.

Haoyang Cao, Samuel Cohen, and Lukasz Szpruch. Identifiability in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12362–12373, 2021.

Nicolo Cesa-Bianchi, Gábor Lugosi, et al. *Prediction, learning, and games*, volume 1. Cambridge university press Cambridge, 2006.

Jianguo Chen, Jinlong Lei, Biqiang Mu, Yiguang Hong, and Hongsheng Qi. Active inverse methods in stackelberg games with bounded rationality. *arXiv preprint arXiv:2510.15582*, 2025.

Siyu Chen, Mengdi Wang, and Zhuoran Yang. Actions speak what you want: Provably sample-efficient reinforcement learning of the quantal stackelberg equilibrium from strategic feedbacks. *arXiv preprint arXiv:2307.14085*, 2023.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pp. 3052–3060. PMLR, 2020.

Till Freihaut and Giorgia Ramponi. On feasible rewards in multi-agent inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 38:114230–114284, 2026.

Till Freihaut, Luca Viano, Volkan Cevher, Matthieu Geist, and Giorgia Ramponi. Learning equilibria from data: Provably efficient multi-agent imitation learning. *Advances in Neural Information Processing Systems*, 38:97744–97794, 2026a.

Till Freihaut, Luca Viano, Emanuele Nevali, Volkan Cevher, Matthieu Geist, and Giorgia Ramponi. Rate optimal learning of equilibria from data. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026b.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International conference on machine learning*, pp. 2160–2169. PMLR, 2019.

Denizalp Goktas, Amy Greenwald, Sadie Zhao, Alec Koppel, and Sumitra Ganesh. Efficient inverse multiagent learning. *arXiv preprint arXiv:2502.14160*, 2025.

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer, 1988. URL `http://eudml.org/doc/204222`.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pp. 2137–2143. PMLR, 2020.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Noah Krever, Jakub Cerny, Moïse Blanchard, and Christian Kroer. Guard: Constructing realistic two-player matrix and security games for benchmarking game-theoretic algorithms. *Advances in Neural Information Processing Systems*, 38, 2026.

Joongkyu Lee and Min-Hwan Oh. Improved online confidence bounds for multinomial logistic bandits. In *International Conference on Machine Learning*, pp. 33576–33615. PMLR, 2025.

Jungyhun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits. *Advances in Neural Information Processing Systems*, 37:124640–124685, 2024.

Junyi Liao, Zihan Zhu, Ethan X Fang, Zhuoran Yang, and Vahid Tarokh. Decoding rewards in competitive games: Inverse game theory with entropy regularization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37610–37622. PMLR, 13–19 Jul 2025. URL `https://proceedings.mlr.press/v267/liao25i.html`.

David Lindner, Andreas Krause, and Giorgia Ramponi. Active exploration for inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 35:5843–5853, 2022.

Qinghua Liu, Yuanhao Wang, and Chi Jin. Learning markov games with adversarial opponents: Efficient algorithms and fundamental limits. In *International Conference on Machine Learning*, pp. 14036–14053. PMLR, 2022.

Maël Macuglia, Paul Friedrich, and Giorgia Ramponi. Fine-tuning behavioral cloning policies with preference-based reinforcement learning. *arXiv preprint arXiv:2509.26605*, 2025.

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.

Panayotis Mertikopoulos. Regret, equilibrium, and learning in games: A guided tour. *arXiv preprint arXiv:2608.09389*, 2026.

Alberto Maria Metelli, Giorgia Ramponi, Alessandro Concetti, and Marcello Restelli. Provably efficient learning of transferable rewards. In *International Conference on Machine Learning*, pp. 7665–7676. PMLR, 2021.

John F Nash Jr. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287. Citeseer, 1999.

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Giorgia Ramponi, Pavel Kolev, Olivier Pietquin, Niao He, Mathieu Laurière, and Matthieu Geist. On imitation in mean-field games. *Advances in Neural Information Processing Systems*, 36: 40426–40437, 2023.

Paul Rolland, Luca Viano, Norman Schürhoff, Boris Nikolov, and Volkan Cevher. Identifiability and generalizability from multiple experts in inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 35:550–564, 2022.

Andreas Schlaginhaufen and Maryam Kamgarpour. Towards the transferability of rewards recovered via regularized inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 37:21461–21501, 2024.

Andreas Schlaginhaufen, Reda Ouhamma, and Maryam Kamgarpour. Efficient preference-based reinforcement learning: Randomized exploration meets experimental design. *Advances in Neural Information Processing Systems*, 38:12146–12183, 2026.

Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, and Andreas Krause. Learning to play sequential games versus unknown opponents. *Advances in neural information processing systems*, 33:8971–8981, 2020.

Jingwu Tang, Gokul Swamy, Fei Fang, and Zhiwei S Wu. Multi-agent imitation learning: Value is easy, regret is hard. *Advances in Neural Information Processing Systems*, 37:27790–27816, 2024.

Andrew Wagenmaker, Perry Dong, Raymond Tsao, Chelsea Finn, and Sergey Levine. Posterior behavioral cloning: Pretraining bc policies for efficient rl finetuning. *arXiv preprint arXiv:2512.16911*, 2025.

Jibang Wu, Weiran Shen, Fei Fang, and Haifeng Xu. Inverse game theory for stackelberg games: the blessing of bounded rationality. *Advances in Neural Information Processing Systems*, 35: 32186–32198, 2022.

Renbo Zhao and Robert M Freund. Analysis of the frank–wolfe method for convex composite optimization involving a logarithmically-homogeneous barrier: R. zhao, rm freund. *Mathematical programming*, 199(1):123–163, 2023.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

APPENDIX

The Appendix contains the proofs of the different results, as well as additional information.

TABLE OF CONTENTS

A Previous work **14**

B Notations and Symbols **15**

B.1 Additional notations for the appendix 15

B.2 Summary of the notations 18

C Hardness results for strategic identifiability of rewards **19**

C.1 Hardness of learning from a fixed policy 19

C.2 A statistical lower bound for adaptive algorithms 21

D Self-concordance properties **23**

D.1 Self-concordance of the log-likelihood in the span semi-norm 23

D.2 Change of coefficients in the Fisher information 25

E Warm-up analysis **27**

E.1 A grouped Kiefer-Wolfowitz theory 28

E.2 Solving for the warm-up allocation 30

E.3 Rounding and certificate 31

F Analysis of the main algorithm **35**

F.1 Reward transfer lemma 35

F.2 Finite-sample Stackelberg sub-optimality guarantee 36

G Useful results **38**

H Additional experiments **41**

H.1 Lobéké environment from the GUARD benchmark 42

H.2 Hyperparameter selection and sensitivity analysis 45

H.3 Diagnosis in synthetic environments 46

A PREVIOUS WORK

We review the literature most relevant to our problem, with a particular focus on reward identifiability and inverse reinforcement learning.

Reward identifiability Strictly convex regularization of the objective, such as entropy, ensures the uniqueness of the optimal policies to any given set of rewards. Nevertheless, inverse RL remains ill-defined as rewards are invariant to potential shaping (Ng et al., 1999). This identifiability issue was studied in the seminal work of Cao et al. (2021), later refined by Rolland et al. (2022) and Schlaginhaufen & Kamgarpour (2024). More recently, Freihaut & Ramponi (2026) extends those results to games, showing that even the nature (e.g. collaborative or competitive) of a game cannot be determined from the feasible reward sets of an expert.

This line of work has motivated inverse RL works like Metelli et al. (2021); Lindner et al. (2022), that minimize the distance between the true and estimated feasible reward sets. Metelli et al. (2021) assumes access to a generative model, an assumption later removed by Lindner et al. (2022). However, these works do not provide a method for extracting a good reward or task representation from the interactions. This is the target of more traditional methods, inspired by the single-agent case.

Inverse reinforcement learning in games When considering multiple interacting agents, inverse RL aims to recover the individual rewards explaining their behaviors. To this end, Mehr et al. (2023) empirically extend the classical MaxEntIRL framework of Ziebart et al. (2008) to games, learning from demonstrations of a Nash equilibrium. Similarly, Liao et al. (2025) study reward recovery from samples of a fixed expert in two-player zero-sum game expert under coverability assumptions. Later, Barbara et al. (2026) derives sample complexity bounds for recovering the feasible reward set by sampling a fixed expert, evaluated in the Hausdorff distance. Under the strong assumption of oracle access to the loss gradient, Goktas et al. (2025) perform principled inverse multi-agent RL with parameter-dependent payoffs. All these settings assume access to reliable equilibrium demonstrations, i.e. that the game has been solved a-priori by a set of players we can learn from (e.g. humans). This is restrictive in many applications, particularly in non-cooperative games where a player must learn to act against an adversary with unknown incentives. This motivates our setting, in which a player learns to behave strategically in arbitrary downstream tasks, playing against a best-responding agent.

Closer to this setup, Chen et al. (2023) extract the reward function from observing the follower in a two-player Stackelberg scenario. By controlling a prescription announced by the leader, they are able to bound the suboptimality of the optimal leader of the estimated environment. In contrast, we assume that the leader directly commits to a policy (as opposed to a prescription), and actively select leader strategies via optimal design to bound the offline suboptimality objective. This has been studied by Chen et al. (2025), where the convergence of their algorithm is studied in the asymptotic regime. Contrary to their work, we derive an algorithm with sample complexity analysis and a matching minimax lower bound. Previously, Wu et al. (2022) also studied bounded rationality in the context of Stackelberg Games. In particular, the authors show that observations are more informative from a QRE than a Nash equilibrium. This means that bounded rationality helps to be more robust against an irrational opponent. Under finite action spaces, this makes it possible to construct algorithms with exponential dependencies on neither the number of leader actions nor the number of follower actions.

Learning against unknown or adversarial opponents. Several works study learning in multi-agent environments with unknown or adaptive opponents, where non-stationarity and strategic interactions make learning particularly challenging. In particular, Sessa et al. (2020) consider learning effective strategies against unknown opponents, while more recent approaches use demonstrations to improve learning in strategic settings, as in DIML (An & Du, 2026). Other works study robustness against adversaries and worst-case behaviors during training (Liu et al., 2022). In contrast, we focus on reward learning under strategic interactions rather than solely on policy optimization or robustness.

Additionally, we expand below on the recent developments around adjacent tasks, such as multi-agent imitation learning, and contrast with the scope and goals of the present paper.

Imitation learning in games Recent imitation learning works in games measure sub-optimality of the learned behavior by the notion of Nash gap, initially introduced by Ramponi et al. (2023). In general Markov games, Tang et al. (2024) showed how harder this objective is compared to the common value gap extension from single-agent problems. Later, these remarks have been extended

to more precise characterizations for imitation learning by Freihaut et al. (2026a;b); Bergerault et al. (2026a). In the non-interactive case, these works show that the hardness of equilibria extraction can be reduced to notions of concentrability and continuity at equilibrium. Further, Freihaut et al. (2026a;b) show how to optimize such Nash gap regrets in the interactive setting by forcing the exploration of the state space to cover uncertain deviations. This is done by controlling all the agents sequentially. We argue that such control over the agents is not realistic in many scenarios, where agents are not collaborative in learning, e.g. in real world environments when learning from interactions with (partially) adversarial players. To this end, we adopt a Stackelberg objective to measure the optimality of extracted policies, evaluating a leader when deployed with other strategic players.

On reward identifiability Furthermore, imitation learning extracts a policy from (approximately) optimal agents corresponding to fixed incentives. Modeling agents' behavior for a fixed task without describing their intents prevents transferring to new environments or learner tasks, limiting adaptation of the learner at deployment. Fortunately, observing (approximately) optimal actions from strategic players allows to extract signals about their individual rewards. However, the complete rewards can in general not be recovered, even with infinite data.

This ambiguity on the reward function has first been explored in the single-agent case, with the analysis of potential-based transformations or potential shaping originating from Ng et al. (1999). In the normal-form game setting, these transformations reduce to constant shifts, whose invariance to optimal policies carry over to the entropy-regularized setting (i.e. softmax policies remain invariant to constant shifts of the rewards). The picture extends to games, where each individual reward is contained in an individual feasible reward set. In this scenario, even the nature of the game (e.g. cooperative, zero-sum) can sometimes not be identifiable from the feasible sets alone (Freihaut & Ramponi, 2026).

To mitigate this issue, one could consider learning in different environments, as pointed out by Rolland et al. (2022); Freihaut & Ramponi (2026). A second possibility is to weaken the exact identifiability objective, which aims to exactly learn the reward function, in particular by minimizing the mismatch with the ground truth in the Euclidean norm (Liao et al., 2025). Our work applies both propositions, where the learner agent induces different environments to the soft-optimal follower, and where only the *strategic identifiability* of its reward is considered.

Empirical inverse reinforcement learning In the single-agent case, the seminal work of Ziebart et al. (2008) develops MaxEntIRL, an algorithm recovering the reward matching the maximum-entropy solution given by the demonstrations. The proposed method operates under linear rewards. Their method essentially boils down to gradient descent on the reward parameters, where gradients, given by expected features, are approximated from rollouts of the policy in the environment.

MaxEntIRL has later been empirically extended to games by Mehr et al. (2023), and the theoretical guarantees of such extensions in terms of strategic exploitability are yet to be understood. Furthermore, one caveat of this method is the requirement that all players are controllable, which can be unreasonable in various applications. In contrast, we develop in this work a principled algorithm with PAC guarantees under a finite number of samples. In our two-player setting, only one agent is assumed controllable, while the non-controllable agent (the follower) is best-responding to the learner (the leader policy).

B Notations and Symbols

B.1 Additional notations for the appendix

For the simplicity of the analysis, we adopt throughout the appendix additional notations not introduced in the main content. We summarize all introduced notations at the end of this section.

Already, let us define the policy spaces as $\Pi_\mu \coloneqq \Delta(\mathcal{A})$ and $\Pi_\nu \coloneqq \Delta(\mathcal{B})$ for the μ- and ν-player, respectively.

Features and feature spaces

We recall from the main content that without loss of generality, the known feature map $\phi : \mathcal{A} \times \mathcal{B} \to \mathbb{R}^d$ is such that $\|\phi(a, b)\|_2 \leq 1$ for all $a, b \in \mathcal{A} \times \mathcal{B}$.

For every $a \in \mathcal{A}$, $b \in \mathcal{B}$, and any policies $\mu \in \Pi_\mu$, $\nu \in \Pi_\nu$, we define the average features as:

$$\psi_\mu(b) \coloneqq \sum_{a' \in \mathcal{A}} \mu(a')\phi(a', b), \qquad \psi_\nu(a) \coloneqq \sum_{b' \in \mathcal{B}} \nu(b')\phi(a, b'),$$

where the notation makes it clear from context if we take the average over $\mathcal{A}$ or $\mathcal{B}$.

These averaged features are used to define the uniform-centered features g_μ spanning the space $\mathcal{G}_1$:

$$g_\mu(b) \coloneqq \psi_\mu(b) - \frac{1}{|\mathcal{B}|} \sum_{b' \in \mathcal{B}} \psi_\mu(b'), \qquad \mathcal{G}_1 \coloneqq \operatorname{span}\{g_\mu(b)\}_{\mu,b}.$$

For any $\mu \in \Pi_\mu$, $\nu \in \Pi_\nu$, let us define the following expected features:

$$\overline{\psi}_\mu(\tilde{\omega}) = \sum_{b \in \mathcal{B}} \mathbb{P}(b|\tilde{\omega}, \mu)\psi_\mu(b) \qquad \text{and} \qquad \overline{\psi}_\nu(\tilde{\omega}) = \sum_{a \in \mathcal{A}} \mathbb{P}(a|\tilde{\omega}, \nu)\psi_\nu(a),$$

where we can therefore recognize $g_\mu(b) = \psi_\mu(b) - \overline{\psi}_\mu(0)$, and similarly for g_ν.

From the definition of the features, d represents the ambient dimension. We also define the following two terms:

$$d_1^* \coloneqq \dim \operatorname{span}\{\psi_\mu(b)\}_{\mu,b}, \qquad d_1^\diamond \coloneqq \dim \mathcal{G}_1,$$

where in general, $d_1^\diamond \leq d_1^* \leq d$.

Using these notations, Assumption 3 is equivalent to assuming $d_1^\diamond \geq 1$.

To simplify notations, we will use $d^\diamond$, d^* and $\mathcal{G}$ to refer to $d_1^\diamond$, d_1^* and $\mathcal{G}_1$, respectively. This is because for our main result, the analysis only depends on these quantities for the leader player, i.e. player 1.

Measuring distances between logits

Our analysis builds on how well the coefficients are estimated throughout the iterations of the algorithm. This is, in part, measured by the differences in induced logits for the best-response player. Intuitively, to allow for arbitrary constant shifts (it doesn't change the softmax), we project away this dimension and express our results with the following span semi-norm.

Definition 4 (Span semi-norm). *For any $n \in \mathbb{Z}_+$ and $v \in \mathbb{R}^n$, $\|v\|_{span} = \max_i v_i - \min_j v_j$.*

In particular, for all $\mu \in \Pi_\mu$, $\omega_1, \omega_2 \in \Omega$,

$$\|\langle \psi_\mu(\cdot), \omega_1 - \omega_2 \rangle\|_{\text{span}} = \max_{b_1, b_2 \in \mathcal{B}} \langle \psi_\mu(b_1) - \psi_\mu(b_2), \omega_1 - \omega_2 \rangle.$$

This latter quantity will be key to our analysis. We therefore define the following two functions:

$$d_\mu(\omega_1, \omega_2) := \|\langle \psi_\mu(\cdot), \omega_1 - \omega_2 \rangle\|_{\text{span}}, \qquad d_*(\omega_1, \omega_2) := \max_{\tilde{\mu}} d_{\tilde{\mu}}(\omega_1, \omega_2),$$

where by Cauchy-Schwarz and the triangle inequality: $d_*(\omega_1, \omega_2) \leq 2\|\omega_1 - \omega_2\|_2 \leq 4S$.

Linear algebra on the identifiable subspace

As mentioned before, the observable space has dimension $d^\diamond \leq d$, and the inequality can be strict. Any interaction can only reveal information about the projection of the reward coefficients in that space, and this is sufficient for any reward representation in terms of strategic identifiability. We therefore define this subspace more precisely here, and in particular appropriate operators over this subspace.

This allows for simpler notations and proofs, but the following is bookkeeping rather than deep machinery. Natural operations resolve cleanly.

Let $P_\mathcal{G}$ denote the orthogonal projector onto $\mathcal{G}$, and define

$$\mathbb{S}(\mathcal{G}) \coloneqq \left\{A \in \mathbb{S}^d : A = P_\mathcal{G} A P_\mathcal{G}\right\}, \qquad \mathbb{S}_+(\mathcal{G}) \coloneqq \mathbb{S}(\mathcal{G}) \cap \mathbb{S}_+^d,$$

so that $\mathbb{S}_+(\mathcal{G})$ is the set of PSD ambient operators supported on $\mathcal{G}$.

We further define the set of positive definite operators on $\mathcal{G}$ as

$$\mathbb{S}_{++}(\mathcal{G}) \coloneqq \left\{A \in \mathbb{S}_{+}(\mathcal{G}) : x^\top A x > 0,\ \forall x \in \mathcal{G} \setminus \{0\}\right\},$$

precisely the set of PSD matrices such that $\operatorname{Im} A = \mathcal{G}$, so that A can cover the information of any pair (μ, ω). We keep ω in $\mathbb{R}^d$ for the simplicity of the analysis and algorithms, so that Images and kernels are always understood as subspaces of the ambient space $\mathbb{R}^d$.

Now fix $U \in \mathbb{R}^{d \times d^\diamond}$ such that the columns of U form an orthonormal basis of $\mathcal{G}$. The definitions below are invariant to the specific choice of U.

For $A \in \mathbb{S}_{+}(\mathcal{G})$, we define its determinant in $\mathcal{G}$ and use the extended-real convention as:

$$\det_{\mathcal{G}}(A) \coloneqq \det(U^\top A U), \qquad \log\det_{\mathcal{G}}(A) = -\infty \quad \text{if} \quad A \notin \mathbb{S}_{++}(\mathcal{G}).$$

This A also implies the following extended dual semi-norm for any $u \in \mathcal{G}$:

$$\|u\|_{A,*} = \begin{cases} \sqrt{u^\top A^\dagger u}, & u \in \operatorname{Im} A, \\ +\infty, & u \notin \operatorname{Im} A, \end{cases}$$

where $A^\dagger$ denotes the ordinary Moore-Penrose pseudo-inverse.

In particular, when $A \in \mathbb{S}_{++}(\mathcal{G})$, we recover the standard semi-norm in $A^\dagger$:

$$\|u\|_{A,*}^2 = \|u\|_{A^\dagger}^2 = u^\top A^\dagger u, \qquad A^\dagger = U(U^\top A U)^{-1} U^\top.$$

For any $M \in \mathbb{S}_{+}(\mathcal{G})$ with eigen value-vector pairs $\{\lambda_i, u_i\}_i$, we extend the trace as:

$$\operatorname{Trace}\left(A^\dagger M\right) = \sum_{i:\,\lambda_i > 0} \lambda_i \|u_i\|_{A,*}^2, \quad \text{such that} \quad \operatorname{Trace}\left(A^\dagger M\right) = +\infty \text{ if } \operatorname{Im} M \not\subseteq \operatorname{Im} A. \tag{4}$$

All the newly introduced operators behave similarly to the usual operators. We show such some properties below.

Lemma 4 (Linear algebra on $\mathcal{G}$)**.** *Let $A, B \in \mathbb{S}_{+}(\mathcal{G})$, $u \in \mathcal{G}$, and $x \in \mathbb{R}^d$. Then the following properties hold.*

1. *Generalized Cauchy-Schwarz inequality: $|\langle u, x\rangle| \le \|x\|_A \|u\|_{A,*}$.*
2. *If $A \succeq cB$ for some $c > 0$, then $\|u\|_{A,*}^2 \le c^{-1}\|u\|_{B,*}^2$.*
   Moreover, if $A, B \in \mathbb{S}_{++}(\mathcal{G})$, then $A^\dagger \preceq c^{-1} B^\dagger$.
3. *If $A \in \mathbb{S}_{++}(\mathcal{G})$ and $H \in \mathbb{S}(\mathcal{G})$, then*

$$\frac{\mathrm{d}}{\mathrm{d}t} \log\det_{\mathcal{G}}(A + tH)\bigg|_{t=0} = \operatorname{Trace}(A^\dagger H), \qquad \operatorname{Trace}(A^\dagger A) = d^\diamond.$$

Proof. Define the following terms:

$$\bar{A} \coloneqq U^\top A U, \qquad \bar{B} \coloneqq U^\top B U, \qquad \bar{u} \coloneqq U^\top u, \qquad \bar{x} \coloneqq U^\top x.$$

The claims then reduce to the corresponding standard facts for $\bar{A}, \bar{B} \in \mathbb{S}_{+}^{d^\diamond}$. □

Coverage and information matrices

For any $\omega_1, \omega_2 \in \Omega$, $\tilde{\mu} \in \Pi_\mu$, we introduce the Fisher information in its integral form as

$$\Gamma_{\tilde{\mu}}(\omega_1, \omega_2) := \int_0^1 \mathcal{I}_{\tilde{\mu}}(\omega_1 + s(\omega_2 - \omega_1))\,\mathrm{d}s,$$

where we recall the Fisher information matrix:

$$\mathcal{I}_{\tilde{\mu}}(\tilde{\omega}) = \eta^2 \mathrm{Cov}_{b \sim \nu^*_{\tilde{\mu}}(\tilde{\omega})}\left[\psi_{\tilde{\mu}}(b)\right] = \eta^2 \mathbb{E}_{b \sim \nu^*_{\tilde{\mu}}}\left[\left(\psi_{\tilde{\mu}}(b) - \overline{\psi}_{\tilde{\mu}}(\tilde{\omega})\right)^{\otimes 2}\right].$$

This motivates the following definitions of feature coverage.

For any policy $\mu \in \Pi_\mu$, we define its uniform coverage as the matrix:

$$M_\mu \coloneqq \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} g_\mu(b) g_\mu(b)^\top \succeq 0,$$

such that the following defines the average uniform coverage for any distribution $w \in \Delta(\Pi_\mu)$:

$$V(w) \coloneqq \int_{\Pi_\mu} M_\mu w(\mathrm{d}\mu) \in \mathbb{S}_+(\mathcal{G}).$$

We additionally call w admissible if $V(w) \in \mathbb{S}_{++}(\mathcal{G})$, i.e. $\operatorname{Im} V(w) = \mathcal{G}$, and denote by $\mathcal{W}_\mathcal{G}$ the set of admissible designs.

As $\operatorname{span} \bigcup_\mu \operatorname{Im} M_\mu = \mathcal{G}$, this implies

$$w \in \mathcal{W}_\mathcal{G} \iff \max_\mu \operatorname{Trace}\left(V(w)^\dagger M_\mu\right) < \infty \iff \log\det{}_\mathcal{G} V(w) > -\infty.$$

Additional notations for the online phase

We introduce additional notations to simplify the indexation of policies and likelihoods during the online phase. Specifically, we define

$$N_k \coloneqq n_{\text{warm}} + k, \quad \mathcal{L}_k(\cdot) \coloneqq \mathcal{L}(\cdot|\mathcal{D}_k), \quad G_k^\lambda(\cdot) \coloneqq G^\lambda(\cdot|\mathcal{D}_k), \qquad \forall k \in [K].$$

B.2 Summary of the notations

General notation

$\Delta(\mathcal{X})$	Probability simplex over a measurable space $\mathcal{X}$
$\mathcal{H}_\mathcal{X}(P)$	Shannon entropy; abbreviated $\mathcal{H}$ when $\mathcal{X}$ is clear
$\|x\|_P$	Mahalanobis (semi-)norm $\sqrt{x^\top P x}$, $P \succeq 0$
$x^{\otimes 2}$	Outer product $xx^\top$
$D_{\mathrm{KL}}(p\|q)$	Kullback–Leibler divergence
$\operatorname{Im} M$, span, rank	Range, linear span, rank
$M^\dagger$	Moore–Penrose pseudo-inverse of a PSD matrix M
$P_\mathcal{G}$	Orthogonal projector onto $\mathcal{G}$
$\mathbb{S}_+(\mathcal{G})$	PSD ambient operators supported on $\mathcal{G}$
$\mathbb{S}_{++}(\mathcal{G})$	Positive definite ambient operators supported on $\mathcal{G}$
$\|u\|_{A,*}$	Dual norm in $\mathcal{G}$
$\det_\mathcal{G}(A)$	Determinant of the restriction of A to $\mathcal{G}$
$\ell(x)$	Logistic function $(1+e^{-x})^{-1}$
$\|v\|_{\text{span}}$	Span seminorm $\max_i v_i - \min_j v_j$

Game and policies

$\mathcal{A}, \mathcal{B}$	Finite action spaces of leader and follower
μ, ν	Policies of the leader and the follower
$\eta > 0$	Inverse temperature (bounded rationality)
$\nu_\mu^*(\tilde{\omega})$	Follower's regularised best response, Equation (1)
$J_r^{\text{leader}}(\mu)$	Leader value $J_{r^1}(\mu, \nu_\mu^*(r^2))$
$\mathrm{SubOpt}_r(\mu)$	Stackelberg sub-optimality
$\mathrm{NashGap}_r(\mu, \nu)$	Maximal regret over the two players
Π_μ, Π_ν	Leader and follower policy spaces

Linear reward model

$\phi(a, b)$	Known feature map, $\max_{a,b}\|\phi(a, b)\|_2 \le 1$
ω, $\hat{\omega}$, $\hat{\omega}_k$	True / final / round-k reward coefficients
S, Ω	Bound $\|\omega\|_2 \le S$ and the ball $\{\tilde{\omega} : \|\tilde{\omega}\|_2 \le S\}$
$R_{\max}$	Upper bound $\|r_1\|_\infty$
$\underline{\psi}_\mu(b)$	Averaged follower feature $\sum_a \mu(a)\phi(a, b)$
$\overline{\psi}_\mu(\tilde{\omega})$	Expected feature $\mathbb{E}_{b\sim\nu^*_\mu(\tilde{\omega})}\psi_\mu(b)$

Identifiable subspace

$g_\mu(b)$	Uniform-centred feature $\psi_\mu(b) - \frac{1}{\|\mathcal{B}\|}\sum_{b'}\psi_\mu(b')$
$\mathcal{G}$	$\mathrm{span}\{g_\mu(b)\}_{\mu,b} = \mathrm{span}\{\psi_\mu(b) - \psi_\mu(d)\}_{\mu,b,d}$
$d^\diamond$	$\dim\mathcal{G}$: the reward directions the follower can distinguish
d^*	$\dim\mathrm{span}\{\psi_\mu(b)\}_{\mu,b}$, so $d^\diamond \le d^* \le d$

Information and estimation

N, n_{warm}, K	Total, warm-up, and online interaction counts, with $K = N - n_{\text{warm}}$
N_k	Total sample count $n_{\text{warm}} + k$ at online checkpoint k
$\mathcal{D}$, $\mathcal{D}_{\text{warm}}$, $\mathcal{D}_k$	Full, warm-up, and online historical data
l, $\mathcal{L}$, $\mathcal{L}_k$	Instantaneous, cumulative and online cumulative log-likelihood
$\mathcal{I}_\mu(\tilde{\omega})$	Fisher information $\eta^2\mathrm{Cov}_{b\sim\nu^*_\mu(\tilde{\omega})}[\psi_\mu(b)]$
G^λ, G^λ_k	Fisher and online cumulative regularized Fisher information
$\Gamma^\lambda_\tau(\tilde{\omega})$	Residual (integrated) information, Theorem 5
$\beta_\tau(\delta)$	MLE confidence radius, Theorem 5
$d_\mu(\omega_1, \omega_2)$	Span seminorm $\|\langle\psi_\mu(\cdot), \omega_1 - \omega_2\rangle\|_{\text{span}}$
$d_*(\omega_1, \omega_2)$	$\max_{\tilde{\mu}} d_{\tilde{\mu}}(\omega_1, \omega_2)$; bounded by $4S$
$\chi(x)$	$2(e^{x/2} - 1)/x$ for $x > 0$, $\chi(0) = 1$, increasing
κ	Coverage constant $2e^{2\eta S}/\eta^2$, Lemma 14

Experimental design

M_μ	Grouped arm $\frac{1}{\|\mathcal{B}\|}\sum_b g_\mu(b)g_\mu(b)^\top \succeq 0$
$V(w)$	Design uniform covariance $\int_{\Pi_\mu} M_\mu w(\mathrm{d}\mu)$
admissible, $\mathcal{W}_\mathcal{G}$	w with $\mathrm{Im}\, V(w) = \mathcal{G}$; $\mathcal{W}_\mathcal{G}$ is the set of such designs
w_D	D-optimal design $\arg\max_w \log\det_\mathcal{G} V(w)$
ϵ	Design / rounding tolerance
n_{warm}, $\mathcal{D}_{\text{warm}}$	Warm-up budget and its sample set
V_{warm}	Uniform allocation coverage matrix

C HARDNESS RESULTS FOR STRATEGIC IDENTIFIABILITY OF REWARDS

We provide in this section the proofs of the impossibility and lower bound results present in the main text, i.e. Lemma 2 and Theorem 1.

C.1 HARDNESS OF LEARNING FROM A FIXED POLICY

We start by proving the hardness of identifying rewards that maintain the optimal leader by querying a single pair of policies.

Lemma 2. *Fix $\eta > 0$. There exists a set $(r_1, \mathcal{A}, \mathcal{B}, \phi)$ such that for any reference policy μ_{ref} and any number of samples K, no algorithm only sampling from $(\mu_{ref}, \nu^*_{\mu_{ref}})$ can guarantee that $\mathrm{SubOpt}_r(\hat{\mu}) \le c_\eta$ with probability greater than $1/2$. $c_\eta > 0$ only depends on η.*

Proof. Let $\mathcal{M}_+, \mathcal{M}_-$ be two Stackelberg games with $\mathcal{A} = \{a_1, a_2\}$, $\mathcal{B} = \{b_1, b_2\}$, and the tabular feature map $\phi(a_i, b_j) = e_{ij} \in \mathbb{R}^4$.

Fix an arbitrary leader policy μ_{ref} and let us write $p_{\text{ref}} = \mu_{\text{ref}}(a_1)$ and $1 - p_{\text{ref}} = \mu_{\text{ref}}(a_2)$.

Define two follower rewards r_2^+ and r_2^- by setting

$$s_1 \coloneqq 1 - p_{\text{ref}}, \qquad s_2 \coloneqq -p_{\text{ref}},$$

and

$$r_2^+(a_i, b_1) = \frac{1}{2} + \frac{s_i}{2}, \qquad r_2^+(a_i, b_2) = \frac{1}{2} - \frac{s_i}{2},$$

while

$$r_2^-(a_i, b_1) = \frac{1}{2} - \frac{s_i}{2}, \qquad r_2^-(a_i, b_2) = \frac{1}{2} + \frac{s_i}{2}.$$

All rewards lie in $[0, 1]$, and since the features are tabular, both r_2^+ and r_2^- are linear in ϕ.

For any $q \in [0, 1]$, we denote by μ_q the leader policy such that $\mu_q(a_1) = q$. Under the policy μ_q, the averaged reward differences are as follows:

- In $\mathcal{M}_+$:
$$\sum_a \mu_q(a) \left(r_2^+(a, b_1) - r_2^+(a, b_2)\right) = q - p_{\text{ref}}$$
- In $\mathcal{M}_-$:
$$\sum_a \mu_q(a) \left(r_2^-(a, b_1) - r_2^-(a, b_2)\right) = p_{\text{ref}} - q$$

Therefore we have the following entropy-regularized best-responses,

$$\nu^*_{q,+}(b_1) = \ell(\eta(q - p_{\text{ref}})), \qquad \nu^*_{q,-}(b_1) = \ell(\eta(p_{\text{ref}} - q)),$$

where ℓ is the standard logistic function $\ell(x) = (1 + \exp(-x))^{-1}$.

At the queried policy $q = p_{\text{ref}}$, both games induce

$$\nu^*_{p_{\text{ref}},+}(b_1) = \nu^*_{p_{\text{ref}},-}(b_1) = \frac{1}{2}.$$

Hence the full observation law is identical in the two games, for every number of samples K.

Now choose the leader reward $r_1(a, b) = \mathbf{1}\{b = b_1\}$ such that the leader values are:

$$J^{\text{leader}}_{r^\pm}(\mu_q) = \nu^*_{q,\pm}(b_1) = \ell(\pm\eta(q - p_{\text{ref}})).$$

As ℓ is strictly increasing, the optimal leaders of $\mathcal{M}_+$ and $\mathcal{M}_-$ are therefore μ_1 and μ_0, respectively, with corresponding optimal values $\ell(\eta(1 - p_{\text{ref}}))$ and $\ell(\eta p_{\text{ref}})$. Using this fact, for any q,

$$\begin{aligned}
\text{SubOpt}_{r^+}(\mu_q) + \text{SubOpt}_{r^-}(\mu_q) &= \ell(\eta(1 - p_{\text{ref}})) - \ell(\eta(q - p_{\text{ref}})) \\
&\qquad + \ell(\eta p_{\text{ref}}) - \ell(-\eta(q - p_{\text{ref}})) \\
&= \ell(\eta(1 - p_{\text{ref}})) + \ell(\eta p_{\text{ref}}) - 1 \\
&\geq \ell(\eta) - \frac{1}{2},
\end{aligned}$$

where the last inequality follows because $p \mapsto \ell(\eta p) + \ell(\eta(1 - p)) - 1$ is symmetric and minimized at $p = 0$ or $p = 1$.

Hence at least one of $\text{SubOpt}_+(q)$ and $\text{SubOpt}_-(q)$ is larger than $\frac{1}{2}\left(\ell(\eta) - \frac{1}{2}\right)$, therefore strictly larger than $c_\eta \coloneqq \frac{1}{3}\left(\ell(\eta) - \frac{1}{2}\right)$.

Since the two games induce exactly the same observation distribution at the only queried policy, any algorithm has the same output distribution under the two games. Let

$$\mathcal{E}_+ \coloneqq \{\text{SubOpt}_{r_+}(\hat{q}) \leq c_\eta\}, \qquad \mathcal{E}_- \coloneqq \{\text{SubOpt}_{r_-}(\hat{q}) \leq c_\eta\},$$

with $\mathcal{E}_+ \cap \mathcal{E}_- = \varnothing$ from the previous remark.

Since $\hat{q}$ has the same law in the two games,

$$2 \cdot \min\{\mathbb{P}_+(\mathcal{E}_+), \mathbb{P}_-(\mathcal{E}_-)\} \leq \mathbb{P}_{\mathcal{M}_+}(\mathcal{E}_+) + \mathbb{P}_{\mathcal{M}_-}(\mathcal{E}_-) \leq 1.$$

So no algorithm observing only samples from the single policy μ_{ref} can guarantee Stackelberg suboptimality at most c_η with probability larger than $1/2$ uniformly over the two games. □

C.2 A STATISTICAL LOWER BOUND FOR ADAPTIVE ALGORITHMS

The usual strategy for a lower bound is to exhibit a problem instance where no algorithm can quickly distinguish between edge cases. We define such a hard instance as follows.

Definition 5 (Hard instance)**.** *Let $\mathcal{A} = [d]$, $\mathcal{B} = \{b_1, b_2\}$, $R_{\max} \in \mathbb{R}_+$, and*

$$\phi(a_i, b_1) = 0, \qquad \phi(a_i, b_2) = e_i, \qquad r_1(a, b) = R_{\max} \cdot \mathbf{1}\{b = b_2\}.$$

We identify $\mu \in \Delta_d$ and write $\ell(x) \coloneqq (1 + e^{-x})^{-1}$ for the logistic function.

For simplicity of the analysis, we first state some properties of this instance.

Proposition 1 (Properties of the hard instance)**.** *Reusing the notations of the hard instance (Definition 5), for every leader $\mu \in \Delta_d$ and $\tilde{\omega} \in \Omega$:*

- *Structural simplifications*

$$\psi_\mu(b_1) = 0, \qquad \psi_\mu(b_2) = \mu, \qquad \|\psi_\mu(b)\|_2 \le 1,$$
$$\mathcal{G} = \mathbb{R}^d, \qquad d^\diamond = d, \qquad d_\mu(0, \varepsilon e_i) = \varepsilon \mu_i \ \ (\varepsilon \ge 0).$$

- *Implied best-response*

$$\nu^*_\mu(\tilde{\omega})(b_2) = p_\mu(\tilde{\omega}).$$

- *Performance*

$$J^{\text{leader}}_{\tilde{\omega}}(\mu) = R_{\max} \cdot p_\mu(\tilde{\omega}), \qquad \text{SubOpt}_{\tilde{\omega}}(\mu) = R_{\max}\left[\ell\left(\eta \max_j \tilde{\omega}_j\right) - p_\mu(\tilde{\omega})\right].$$

where we denote $p_\mu(\tilde{\omega}) \coloneqq \ell(\eta\langle \mu, \tilde{\omega}\rangle)$.

Proof. We only prove some properties, the rest directly follows from definition.

Implied best-response

Note that since $|\mathcal{B}| = 2$, the best-response is the following logistic law:

$$\nu^*_\mu(\tilde{\omega})(b_2) = \frac{\exp\left(\eta\langle \mu, \tilde{\omega}\rangle\right)}{1 + \exp\left(\eta\langle \mu, \tilde{\omega}\rangle\right)} = \ell\left(\eta\langle \mu, \tilde{\omega}\rangle\right) = p_\mu(\tilde{\omega}). \tag{5}$$

Performance

Given Equation (5), since ℓ is increasing and $\mu \mapsto \langle \mu, \tilde{\omega}\rangle$ is linear on the simplex,

$$\max_\mu J^{\text{leader}}_{\tilde{\omega}}(\mu) = R_{\max} \cdot \ell\left(\eta \max_j \tilde{\omega}_j\right).$$

This implies the $\text{SubOpt}_{\tilde{\omega}}(\mu)$ equality. □

We can now state the lower bound as follows.

Theorem 1 (Minimax lower bound)**.** *Let $d \ge 32$, $\epsilon = \frac{1}{2\eta}\sqrt{d/K} \le S$ and $\omega^{(i)} \coloneqq \varepsilon e_i$ for $i \in [d]$. There exists a game such that for every algorithm producing an output $\hat{\mu}$ after $K \ge 4d$ interactions,*

$$\max_{i \in [d]} \mathbb{E}^{(i)}\left[\text{SubOpt}_{\omega^{(i)}}(\hat{\mu})\right] \ \ge \ \frac{1}{44} R_{\max}\sqrt{\frac{d}{K}},$$

where $\mathbb{E}^{(i)}$ is for the expectation under the trajectory law $\mathbb{P}^{(i)}$ induced by $\omega^{(i)}$.

Proof. Consider the hard instance of Definition 5 and note that in this environment, $d = d^\diamond$.

Let $\mathbb{P}^0$ denote the law of the trajectory under the null hypothesis $\omega^{(0)} := 0$. We use two budget constraints under $\mathbb{P}^0$. First, since $\|\mu_k\|_2 \le \|\mu_k\|_1 = 1$,

$$\mathbb{E}^0\left[\sum_{k\le K}\|\mu_k\|_2^2\right] = \sum_{i=1}^{d}\mathbb{E}^0\left[\sum_{k\le K}\mu_{k,i}^2\right] \le K, \tag{6}$$

which corresponds to a bound on the collected Fisher information[12].

Second, by Markov's inequality and $\sum_i \hat{\mu}_i = 1$,

$$\sum_{i=1}^{d}\mathbb{P}^0\left(\hat{\mu}_i \ge \tfrac{1}{2}\right) \le 2\sum_{i=1}^{d}\mathbb{E}^0\left[\hat{\mu}_i\right] = 2. \tag{7}$$

By Equation (6) at most $d/4$ indices satisfy $\mathbb{E}^0[\sum_k \mu_{k,i}^2] > 4K/d$, and by Equation (7) at most $d/4$ indices satisfy $\mathbb{P}^0(\hat{\mu}_i \ge 1/2) > 8/d$. Since $d/4 + d/4 < d$, there exists $i^\star \in [d]$ with

$$\mathbb{E}^0\left[\sum_{k\le K}\mu_{k,i^\star}^2\right] \le \frac{4K}{d} \qquad \text{and} \qquad \mathbb{P}^0\left(\hat{\mu}_{i^\star} \ge \tfrac{1}{2}\right) \le \frac{8}{d}. \tag{8}$$

Indistinguishability. By the decomposition of the KL (Lemma 22) and the KL of the softmax (Lemma 23), we have:

$$\begin{aligned}
D_{\mathrm{KL}}\left(\mathbb{P}^0 \,\middle\|\, \mathbb{P}^{(i^\star)}\right) &\le \frac{\eta^2}{8}\sum_{k\le K}\mathbb{E}^0\left[d_{\mu_k}(0,\varepsilon e_{i^\star})^2\right] && \text{(Lemmas 22, 23)}\\
&= \frac{\eta^2\varepsilon^2}{8}\,\mathbb{E}^0\left[\sum_{k\le K}\mu_{k,i^\star}^2\right] && \text{(Proposition 1)}\\
&\le \frac{\eta^2\varepsilon^2 K}{2d} && \text{(Equation (8))}\\
&= \frac{1}{8},
\end{aligned}$$

where the last step uses $\eta^2\varepsilon^2 = d/(4K)$. Pinsker's inequality gives $\mathrm{TV}(\mathbb{P}^0, \mathbb{P}^{(i^\star)}) \le \sqrt{D_{\mathrm{KL}}/2} \le 1/4$, so with Equation (8) and $d \ge 32$,

$$\mathbb{P}^{(i^\star)}\left(\hat{\mu}_{i^\star} \ge \tfrac{1}{2}\right) \le \frac{8}{d} + \frac{1}{4} \le \frac{1}{2}, \qquad \text{hence} \qquad \mathbb{P}^{(i^\star)}\left(\hat{\mu}_{i^\star} < \tfrac{1}{2}\right) \ge \frac{1}{2}.$$

Performance implication. From Proposition 1 we have $\mathrm{SubOpt}_{\omega^{(i^\star)}}(\mu) = R_{\max}\left[\ell(\eta\varepsilon) - \ell(\eta\varepsilon\mu_{i^\star})\right]$, where we note that since $K \ge 4d$,

$$\eta\varepsilon = \frac{1}{2}\sqrt{d/K} \le \frac{1}{4}.$$

Therefore, on the event $\{\hat{\mu}_{i^\star} < 1/2\}$ we have,

$$\begin{aligned}
\mathrm{SubOpt}_{\omega^{(i^\star)}}(\hat{\mu}) &\overset{(i)}{\ge} R_{\max}\left[\ell(\eta\varepsilon) - \ell\left(\tfrac{\eta\varepsilon}{2}\right)\right]\\
&\overset{(ii)}{\ge} \frac{R_{\max}\eta\varepsilon}{2}\inf_{|x|\le 1}\ell'(x)\\
&= \frac{R_{\max}\eta\varepsilon}{2}\cdot\frac{e}{(1+e)^2}\\
&\ge \frac{R_{\max}\eta\varepsilon}{11}.
\end{aligned}$$

[12] Indeed, one can show that $\mathbb{E}^0\left[\mathrm{Trace}\left(\sum_{k\le K}\mathcal{I}_{\mu_k}(0)\right)\right] = \frac{\eta^2}{4}\mathbb{E}^0\left[\sum_{k\le K}\|\mu_k\|_2^2\right]$.

where (i) and (ii) come from the monotonicity of ℓ and the mean value theorem, respectively.

Combining with the indistinguishability bound above we get:

$$\mathbb{E}^{(i^\star)}\left[\mathrm{SubOpt}_{\omega^{(i^\star)}}(\hat{\mu})\right] \geq \frac{1}{2}\cdot\frac{R_{\max}\eta\varepsilon}{11} = \frac{R_{\max}\eta\varepsilon}{22} = \frac{1}{44}R_{\max}\sqrt{\frac{d}{K}},$$

where $d = d^\diamond$ (Proposition 1). □

D SELF-CONCORDANCE PROPERTIES

D.1 SELF-CONCORDANCE OF THE LOG-LIKELIHOOD IN THE SPAN SEMI-NORM

In this section, we derive self-concordance properties for the log-likelihood function, inspired by Lee & Oh (2025) but expressed in the span semi-norm.

Extending the standard self-concordance property (Bach, 2010), we define a self-concordance-like function in this semi-norm as follows.

Definition 6 (Self-concordance in the span semi-norm)**.** *A function $f : \mathbb{R}^m \to \mathbb{R}$ is an M-self-concordant-like function in the span semi-norm if:*

$$|\phi'''(s)| \leq M\|b\|_{span}\phi''(s), \qquad \forall s \in \mathbb{R},$$

where $\phi(s) \coloneqq f(a+sb)$ for any $a, b \in \mathbb{R}^m$.

Recall the instantaneous log-likelihood:

$$l(\tilde{\omega}|b_\tau, \mu_\tau) = \eta\langle\psi_{\mu_\tau}(b_\tau), \tilde{\omega}\rangle - \log Z_{\mu_\tau}(\tilde{\omega}).$$

Now, let $-f_\tau$ be a reparameterization with logits $z \in \mathbb{R}^{|\mathcal{B}|}$:

$$f_\tau(z) = \log\left(\sum_{b'} e^{\eta z_{b'}}\right) - \eta z_{b_\tau},$$

and note that f_τ is invariant to constant shifts: $f_\tau(z + a\cdot 1) = f_\tau(z)$ for all $z \in \mathbb{R}^{|\mathcal{B}|},\ a \in \mathbb{R}$.

Then, we have the following Lemma.

Lemma 5. *The function f_τ is η-self-concordant-like in the span semi-norm.*

Proof. For any $a, b \in \mathbb{R}^{|\mathcal{B}|}$, let $\phi(s) \coloneqq f_\tau(a+sb)$ and $p_i(s) \coloneqq e^{\eta(a_i+sb_i)}/\sum_j e^{\eta(a_j+sb_j)}$ for every $i \in [|\mathcal{B}|]$. Further, let X_s be the discrete random variable such that for all i, $\mathbb{P}(X_s = i) = p_i(s)$.

Then,

$$\phi''(s) = \eta^2\mathrm{Var}(b_{X_s}), \qquad \phi'''(s) = \eta^3\mathbb{E}\left[\left(b_{X_s} - \mathbb{E}\left[b_{X_s}\right]\right)^3\right].$$

By Hölder's inequality we then conclude:

$$|\phi'''(s)| \leq \eta\|b - \mathbb{E}[b_{X_s}]\|_\infty\phi''(s) \leq \eta\|b\|_{\text{span}}\phi''(s),$$

and therefore f_τ is η-self-concordant-like. □

Now similarly to Proposition B.4 in Lee & Oh (2025), we can derive the following.

Lemma 6. *For any $\tau \in [N]$ and any $x, u_1, u_2 \in \mathbb{R}^{|\mathcal{B}|}$, we have:*

$$\left|D^3 f_\tau(x)[u_1, u_2, u_2]\right| \leq \eta\|u_1\|_{span}\|u_2\|^2_{\nabla^2 f_\tau(x)},$$

where we define $D^3 f_\tau(x)[u_1, u_2, u_3] \coloneqq \langle D^3 f_\tau(x)[u_1]u_2, u_3\rangle$.

Proof. Fix $x, u_1, u_2 \in \mathbb{R}^{|\mathcal{B}|}$ and define for every $s \in \mathbb{R}$ and $i \in [|\mathcal{B}|]$ the function $p_i(s) := e^{\eta(x_i + s u_{1,i})} / \sum_j e^{\eta(x_j + s u_{1,j})}$.

Additionally, let X_s a discrete random variable such that $\mathbb{P}(X_s = i) = p_i(s)$ and note that for any $g : \mathbb{R} \to \mathbb{R}$,

$$\frac{d}{ds}\mathbb{E}[g(X_s)] = \eta \mathbb{E}[g(X_s) \cdot ((u_1)_{X_s} - \mathbb{E}[(u_1)_{X_s}])].$$

Hence,

$$\begin{aligned} D^3 f_\tau(x)[u_1, u_2, u_2] &= \eta^2 \left.\frac{d}{ds}\mathrm{Var}((u_2)_{X_s})\right|_{s=0} \\ &= \eta^2 \left.\frac{d}{ds}\left(\mathbb{E}[(u_2)_{X_s}^2] - \mathbb{E}[(u_2)_{X_s}]^2\right)\right|_{s=0} \\ &= \eta^3 \mathbb{E}\left[\left((u_2)_{X_0} - \mathbb{E}[(u_2)_{X_0}]\right)^2 \left((u_1)_{X_0} - \mathbb{E}[(u_1)_{X_0}]\right)\right]. \end{aligned}$$

Therefore by Hölder's inequality,

$$\left|D^3 f_\tau(x)[u_1, u_2, u_2]\right| \le \eta^3 \|u_1\|_{\text{span}} \mathrm{Var}\left((u_2)_{X_0}\right) = \eta \|u_1\|_{\text{span}} \|u_2\|^2_{\nabla^2 f_\tau(x)}.$$

□

Finally, we derive the following two lemmas (variants of Propositions B.5 and B.6 of Lee & Oh (2025)), that we will use throughout the analysis of both algorithms.

Lemma 7. *For all $\tau \in [N]$ and any $z_1, z_2 \in \mathbb{R}^{|\mathcal{B}|}$,*

$$e^{-\eta \|z_1 - z_2\|_{span}} \nabla^2 f_\tau(z_1) \preceq \nabla^2 f_\tau(z_2) \preceq e^{\eta \|z_1 - z_2\|_{span}} \nabla^2 f_\tau(z_1).$$

Proof. Fix $z_1, z_2, u \in \mathbb{R}^{|\mathcal{B}|}$, $\tau \in [N]$. Then for any $s \in [0, 1]$, let $g(s) := u^\top \nabla^2 f_\tau\left(z_1 + s(z_2 - z_1)\right) u$ such that $g(0) = \|u\|^2_{\nabla^2 f_\tau(z_1)}$ and $g(1) = \|u\|^2_{\nabla^2 f_\tau(z_2)}$.

Then from Lemma 6,

$$|g'(s)| \le \eta \|z_2 - z_1\|_{\text{span}} g(s).$$

Using $\nabla^2 f_\tau(z) \succeq 0$ for all z, we then integrate as follows:

$$\int_0^1 -\eta \|z_2 - z_1\|_{\text{span}} \,\mathrm{d}s \le \int_0^1 \frac{\mathrm{d} \log g(s)}{\mathrm{d}s} \,\mathrm{d}s \le \int_0^1 \eta \|z_2 - z_1\|_{\text{span}} \,\mathrm{d}s,$$

hence,

$$e^{-\eta \|z_2 - z_1\|_{\text{span}}} \le \frac{g(1)}{g(0)} \le e^{\eta \|z_2 - z_1\|_{\text{span}}},$$

from which we conclude. □

Lemma 8. *For any $\tau \in [N],\ u, z_1, z_2 \in \mathbb{R}^{|\mathcal{B}|}$:*

$$u^\top \left(\int_0^1 (1 - s) \nabla^2 f_\tau(z_1 + s(z_2 - z_1)) ds\right) u \ge \frac{1}{2 + \eta \|z_2 - z_1\|_{span}} u^\top \nabla^2 f_\tau(z_1) u.$$

Proof. If $\|z_2 - z_1\|_{\text{span}} = 0$, the relation holds directly with equality. Hence we assume below $\|z_2 - z_1\|_{\text{span}} > 0$.

Applying Lemma 7 we get

$$\begin{aligned} u^\top \left(\int_0^1 (1 - s) \nabla^2 f_\tau(z_1 + s(z_2 - z_1)) ds\right) u &\ge u^\top \left(\int_0^1 (1 - s) e^{-\eta \|s(z_2 - z_1)\|_{\text{span}}} \nabla^2 f_\tau(z_1) ds\right) u \\ &\ge u^\top \nabla^2 f_\tau(z_1) u \left(\frac{1}{\eta \|z_2 - z_1\|_{\text{span}}} + \frac{e^{-\eta \|z_2 - z_1\|_{\text{span}}} - 1}{(\eta \|z_2 - z_1\|_{\text{span}})^2}\right) \\ &\ge u^\top \nabla^2 f_\tau(z_1) u \left(\frac{1}{2 + \eta \|z_2 - z_1\|_{\text{span}}}\right), \end{aligned}$$

where the third inequality uses the relation $\frac{1}{x}\left(1 + \frac{e^{-x} - 1}{x}\right) \ge \frac{1}{2 + x}$ for all $x > 0$. □

D.2 Change of coefficients in the Fisher information

Noting that $-\nabla^2 l(\tilde{\omega}|b, \mu_\tau) = \mathcal{I}_{\mu_\tau}(\tilde{\omega})$ for any $b \in \mathcal{B}$ and $\tau \in [N]$, we apply the previous results to bound the Fisher information matrix on changes of coefficients.

Specifically, the previous results are expressed in the logit space via the function f_τ. We turn back to the coefficient space by using the following linear lift:

$$z_\mu(\tilde{\omega}) \coloneqq \Psi_\mu \tilde{\omega} \in \mathbb{R}^{|\mathcal{B}|}, \qquad \Psi_\mu \coloneqq \left[\psi_\mu(b)^\top\right]_{b\in\mathcal{B}} \in \mathbb{R}^{|\mathcal{B}|\times d}, \tag{9}$$

where $z_\mu(\omega)$ represents the logits associated to the policy μ and the coefficients $\tilde{\omega}$.

Explicitly,

$$-l(\tilde{\omega}|b_\tau, \mu_\tau) = \log\left(\sum_{b'\in\mathcal{B}} e^{\eta\langle\psi_{\mu_\tau}(b'),\tilde{\omega}\rangle}\right) - \eta\langle\psi_{\mu_\tau}(b_\tau), \tilde{\omega}\rangle = f_\tau(z_{\mu_\tau}(\tilde{\omega})), \tag{10}$$

and logit differences can be measured by $d_\mu(\omega_1, \omega_2) = \|z_\mu(\omega_1) - z_\mu(\omega_2)\|_{\text{span}}$.

Further, note that for any τ, τ', the function $f_\tau, f_{\tau'}$ only differ by a linear term. Therefore, the hessian $H(z) \coloneqq \nabla^2 f_\tau(z)$ doesn't depend on τ. Hence, using Equation (10), we have for any μ:

$$\mathcal{I}_\mu(\omega) = \Psi_\mu^\top H\left(z_\mu(\omega)\right)\Psi_\mu. \tag{11}$$

We can now state useful results following Appendix D.1 and these remarks.

Lemma 9. *For any $\omega_1, \omega_2 \in \Omega$ and μ,*

$$e^{-\eta d_\mu(\omega_1,\omega_2)}\mathcal{I}_\mu(\omega_1) \preceq \mathcal{I}_\mu(\omega_2) \preceq e^{\eta d_\mu(\omega_1,\omega_2)}\mathcal{I}_\mu(\omega_1).$$

Proof. By Lemma 7,

$$e^{-\eta d_\mu(\omega_1,\omega_2)}H(z_\mu(\omega_1)) \preceq H(z_\mu(\omega_2)) \preceq e^{\eta d_\mu(\omega_1,\omega_2)}H(z_\mu(\omega_1)).$$

Now note that for any $A \preceq B$, we have $\Psi_\mu^\top A\Psi_\mu \preceq \Psi_\mu^\top B\Psi_\mu$, since for any $v \in \mathbb{R}^d$,

$$v^\top\left(\Psi_\mu^\top A\Psi_\mu - \Psi_\mu^\top B\Psi_\mu\right)v = \left(\Psi_\mu v\right)^\top(A-B)\left(\Psi_\mu v\right) \leq 0.$$

We therefore conclude using Equation (11). □

Lemma 10. *For any $\omega_1, \omega_2 \in \Omega$ and μ, denote $a \coloneqq \eta \cdot d_\mu(\omega_1, \omega_2)$. Then,*

$$\frac{1}{2+a}\mathcal{I}_\mu(\omega_1) \preceq \int_0^1 (1-s)\mathcal{I}_\mu(\omega_1 + s(\omega_2-\omega_1))ds \preceq \frac{e^a - a - 1}{a^2}\mathcal{I}_\mu(\omega_1).$$

Proof. Write $\omega_s \coloneqq \omega_1 + s(\omega_2 - \omega_1)$ and $z_i \coloneqq z_\mu(\omega_i)$, so that $z_\mu(\omega_s) = z_1 + s(z_2 - z_1)$ by linearity of the lift (Equation (9)).

Lower bound. Fix $v \in \mathbb{R}^d$ and apply Lemma 8 to $u \coloneqq \Psi_\mu v$. Using Equation (11),

$$\begin{aligned} v^\top\left(\int_0^1 (1-s)\mathcal{I}_\mu(\omega_s)ds\right)v &= u^\top\left(\int_0^1 (1-s)H(z_1 + s(z_2-z_1))ds\right)u \\ &\geq \frac{1}{2+a}\,u^\top H(z_1)u \\ &= \frac{1}{2+a}\,v^\top\mathcal{I}_\mu(\omega_1)v. \end{aligned}$$

Upper bound. Since $\omega_s - \omega_1 = s(\omega_2 - \omega_1)$ and $\eta d_\mu(\omega_s, \omega_1) = sa$ for every $s \in [0,1]$, Lemma 9 thus gives $\mathcal{I}_\mu(\omega_s) \preceq e^{sa}\mathcal{I}_\mu(\omega_1)$. Integrating,

$$\int_0^1 (1-s)\mathcal{I}_\mu(\omega_s)\,\mathrm{d}s \preceq \left(\int_0^1 (1-s)e^{sa}\,\mathrm{d}s\right)\mathcal{I}_\mu(\omega_1) = \frac{e^a - a - 1}{a^2}\,\mathcal{I}_\mu(\omega_1).$$

Both constants extend continuously to $a = 0$, where they equal $\int_0^1(1-s)ds = 1/2$. □

Lemma 11. *For any $\omega_1, \omega_2 \in \Omega$ and μ, denote $a := \eta \cdot d_\mu(\omega_1, \omega_2)$. Then,*

$$\frac{1 - e^{-a}}{a} \mathcal{I}_\mu(\omega_1) \preceq \Gamma_\mu(\omega_1, \omega_2) \preceq \frac{e^a - 1}{a} \mathcal{I}_\mu(\omega_1),$$

with $\mathcal{I}_\mu(\omega_1) \preceq \Gamma_\mu(\omega_1, \omega_2) \preceq \mathcal{I}_\mu(\omega_1)$ when $a = 0$.

Proof. The proof is identical to the proof of Lemma 9, integrating $\mathcal{I}_\mu(\omega_s)$ directly without the $(1 - s)$ factor. □

Lemma 12 (Best-response stability)**.** *For every ω_1, ω_2 and μ,*

$$\left\| \nu^*_\mu(\omega_1) - \nu^*_\mu(\omega_2) \right\|_1 \leq \int_0^1 \|\omega_1 - \omega_2\|_{\mathcal{I}_\mu(\omega_s)} \, ds \leq \chi\left(\eta \cdot d_\mu(\omega_1, \omega_2)\right) \cdot \|\omega_1 - \omega_2\|_{\mathcal{I}_\mu(\omega_2)},$$

with $\chi(x) := \frac{2(e^{x/2} - 1)}{x}$ (increasing function) and $\omega_s := \omega_2 + s(\omega_1 - \omega_2)$.

Proof. Write $a := \eta \cdot d_\mu(\omega_1, \omega_2)$, $p_s := \nu^*_\mu(\omega_s) \in \Delta(\mathcal{B})$, and let $h(b) := \langle \psi_\mu(b), \omega_1 - \omega_2 \rangle$.

Note: since $\mathcal{B}$ is finite and ν^*_μ is a softmax, $s \mapsto p_s$ is C^1 and $p_s(b) > 0$ for every $b \in \mathcal{B}$.

By Equation (3) and $\mathrm{d}\omega_s = (\omega_1 - \omega_2)\,\mathrm{d}s$, for every $b \in \mathcal{B}$,

$$\frac{\mathrm{d}}{\mathrm{d}s} \log p_s(b) = \eta \left(h(b) - \mathbb{E}_{p_s}[h]\right) \Leftrightarrow \frac{\mathrm{d}}{\mathrm{d}s} p_s(b) = \eta p_s(b) \left(h(b) - \mathbb{E}_{p_s}[h]\right).$$

Taking the ℓ_1 norm and applying Jensen's inequality to the concave map $x \mapsto \sqrt{x}$,

$$\begin{aligned}
\left\| \frac{\mathrm{d}}{\mathrm{d}s} p_s \right\|_1 &= \eta \mathbb{E}_{p_s}\left[|h - \mathbb{E}_{p_s}[h]|\right] \\
&\leq \eta \sqrt{\mathrm{Var}_{p_s}(h)} && \text{(Jensen's inequality)} \\
&= \sqrt{(\omega_1 - \omega_2)^\top \eta^2 \mathrm{Cov}_{b \sim p_s}\left[\psi_\mu(b)\right](\omega_1 - \omega_2)} \\
&= \|\omega_1 - \omega_2\|_{\mathcal{I}_\mu(\omega_s)}. && \text{(exponential family)}
\end{aligned}$$

The fundamental theorem of calculus and the triangle inequality then give

$$\left\| \nu^*_\mu(\omega_1) - \nu^*_\mu(\omega_2) \right\|_1 = \left\| \int_0^1 \frac{\mathrm{d}}{\mathrm{d}s} p_s \, \mathrm{d}s \right\|_1 \leq \int_0^1 \left\| \frac{\mathrm{d}}{\mathrm{d}s} p_s \right\|_1 \mathrm{d}s \leq \int_0^1 \|\omega_1 - \omega_2\|_{\mathcal{I}_\mu(\omega_s)} \, \mathrm{d}s.$$

Now, since d_μ is induced by a semi-norm and $\omega_s - \omega_2 = s(\omega_1 - \omega_2)$, we have $\eta \cdot d_\mu(\omega_s, \omega_2) = sa$. By Lemma 9, this implies $\mathcal{I}_\mu(\omega_s) \preceq e^{sa} \mathcal{I}_\mu(\omega_2)$. Taking square roots and integrating,

$$\int_0^1 \|\omega_1 - \omega_2\|_{\mathcal{I}_\mu(\omega_s)} \, \mathrm{d}s \leq \left(\int_0^1 e^{sa/2} \, \mathrm{d}s \right) \|\omega_1 - \omega_2\|_{\mathcal{I}_\mu(\omega_2)} = \chi(a) \cdot \|\omega_1 - \omega_2\|_{\mathcal{I}_\mu(\omega_2)}.$$

□

Lemma 13 (Fisher uncertainty transfer)**.** *For any $\omega_1, \omega_2, \lambda, \mu$ and allocation $\{\mu_\tau\}_{\tau \leq k}$,*

$$\mathrm{Trace}(G^{-1}(\omega_1)\mathcal{I}_\mu(\omega_1)) \leq e^{2\eta d_*(\omega_1, \omega_2)} \mathrm{Trace}(G^{-1}(\omega_2)\mathcal{I}_\mu(\omega_2)),$$

where $G(\tilde{\omega}) = \lambda I + \sum_{\tau=1}^k \mathcal{I}_{\mu_\tau}(\tilde{\omega})$.

Proof. Set $a_* := \eta \cdot d_*(\omega_1, \omega_2) \geq 0$.

Now noting that $d_{\mu_\tau} \leq d_*$ for all τ, apply Lemma 9 and get:

$$G(\omega_2) = \lambda I + \sum_{\tau=1}^k \mathcal{I}_{\mu_\tau}(\omega_2) \preceq \lambda I + e^{a_*} \sum_{\tau=1}^k \mathcal{I}_{\mu_\tau}(\omega_1) \preceq e^{a_*} G(\omega_1),$$

where we also noted that $\lambda I \preceq e^{a_*}\lambda I$ since $a_* \geq 0$.

Inverting on both sides we get:

$$G^{-1}(\omega_2) \succeq e^{-a_*}G^{-1}(\omega_1) \Leftrightarrow G^{-1}(\omega_1) \preceq e^{a_*}G^{-1}(\omega_2). \tag{12}$$

We then use twice the fact that

$$A_1 \preceq A_2,\ C \succeq 0 \implies \mathrm{Trace}(A_1 C) \leq \mathrm{Trace}(A_2 C), \qquad \forall A_1, A_2, C,$$

which follows from $\mathrm{Trace}\left((A_2 - A_1)C\right) = \mathrm{Trace}\left(C^{1/2}(A_2 - A_1)C^{1/2}\right) \geq 0$.

Specifically,

$$\begin{aligned}
\mathrm{Trace}\left(G^{-1}(\omega_1)\mathcal{I}_\mu(\omega_1)\right) &\leq e^{a_*}\,\mathrm{Trace}\left(G^{-1}(\omega_2)\mathcal{I}_\mu(\omega_1)\right) && \text{(Equation (12))} \\
&\leq e^{2a_*}\,\mathrm{Trace}\left(G^{-1}(\omega_2)\mathcal{I}_\mu(\omega_2)\right) && \text{(Lemma 9)}.
\end{aligned}$$

□

E WARM-UP ANALYSIS

An offline warm-up allows to directly optimize a set of exploratory policies to reduce the maximum uncertainty. However, during this offline phase we cannot collect information about the unknown reward coefficients, which prevents the algorithm to query the best-responding player in the optimal directions.

We therefore perform optimal design assuming a uniform best-responding agents (i.e. $\omega = 0 \in \Omega$), and pay for this choice with a global self-concordance introducing an exponential term κ in the burn-in phase for exploration. This exponential dependency is then removed in the online phase, which relies on a running estimate for ω (namely, the maximum likelihood estimator).

Formally, this price is given by the following lemma.

Lemma 14 (Design curvature)**.** *For $n \geq 1$ and any allocation $\{\mu_\tau\}_{\tau \leq n}$, define the uniform-centered design $V_n \coloneqq \sum_{\tau=1}^{n} M_{\mu_\tau}$ and Γ_n^0 as in Theorem 5 with $\lambda = 0$. Then for any $\tilde{\omega} \in \Omega$,*

$$G^0\left(\tilde{\omega}|\{\mu_\tau\}_{\tau \leq n}\right) \succeq 2\kappa^{-1}V_n \quad \textit{and} \quad \Gamma_n^0(\tilde{\omega}) \succeq \kappa^{-1}V_n,$$

for $\kappa \coloneqq 2\eta^{-2}e^{2\eta S}$.

Hence for any $v \in \mathbb{R}^d$, $\|v\|_{V_n}^2 \leq \kappa \|v\|_{\Gamma_n^0(\tilde{\omega})}^2$.

Proof. Denote by $p_{\min} \geq e^{-2\eta S}/|\mathcal{B}|$ a lower bound on the minimal probability of the softmax distribution with normalized features and coefficient $\tilde{\omega} \in \Omega$. Hence, for any $\tau \in [n]$,

$$\begin{aligned}
\mathrm{Cov}_{b \sim \mathbb{P}(\cdot|\tilde{\omega}, \mu_\tau)}\left[\psi_{\mu_\tau}(b)\right] &\succeq p_{\min} \sum_{b \in \mathcal{B}} \left(\psi_{\mu_\tau}(b) - \overline{\psi}_{\mu_\tau}(\tilde{\omega})\right)^{\otimes 2} \\
&\succeq |\mathcal{B}| p_{\min} \cdot M_{\mu_\tau},
\end{aligned}$$

where the second step re-centers at the uniform mean ($\sum_b (x_b - c)^{\otimes 2} \succeq \sum_b (x_b - \overline{x})^{\otimes 2}$ for any c).

Now fix $s \in [0, 1]$, write $\omega_s \coloneqq \tilde{\omega} + s(\omega - \tilde{\omega}) \in \Omega$, and use the previous inequality to show[13]:

$$\begin{aligned}
G^0\left(\omega_s|\{\mu_\tau\}_{\tau \leq n}\right) &= \eta^2 \sum_{\tau=1}^{n} \mathrm{Cov}_{b \sim \mathbb{P}(\cdot|\omega_s, \mu_\tau)}\left[\psi_{\mu_\tau}(b)\right] \\
&\succeq \eta^2 |\mathcal{B}| p_{\min} \sum_{\tau=1}^{n} M_{\mu_\tau} \\
&\succeq \eta^2 e^{-2\eta S} \cdot V_n.
\end{aligned}$$

[13] This is independent of the sampled follower actions.

Algorithm 2 Warm-up procedure (WARM)

1: **Input:** tolerance ϵ, budget n_{warm}.
2: *Compute warm-up design (offline)* ▷ See Section E.2
3: Approximately solve the following grouped design problem:

$$\arg\min_{\tilde{w}\in\Delta(\Pi_\mu)} \max_{\mu} \max_{b\in\mathcal{B}} \|g_\mu(b)\|^2_{V(\tilde{w}),*},$$

and obtain $w_{\approx}$ an $(1+\epsilon)|\mathcal{B}|$ approximation.
4: This requires an iterative algorithm for optimal design, presented in Algorithm 3.
5: *Rounding procedure* ▷ See Section E.3
6: Apply the rounding procedure of Definition 7 to $w_{\approx}$ and obtain the allocation

$$\mathbf{v} = [v_1, ..., v_s] \in \mathbb{N}^s \qquad \text{such that} \qquad 1^\top \mathbf{v} = n_{\text{warm}},$$

7: where $s \leq d^\diamond(d^\diamond+1)/2 + 1$.
8: *Sample and initialize*
9: Sample $\mathcal{D}_{\text{warm}}$; $\mathcal{L}_0(\tilde{\omega}) \leftarrow \mathcal{L}(\tilde{\omega}|\mathcal{D}_{\text{warm}})$; $\hat{\omega}_0 \leftarrow \arg\max_{\tilde{\omega}\in\Omega} \mathcal{L}_0(\tilde{\omega})$.
10: **Output:** $\mathcal{D}_{\text{warm}}$, $\mathcal{L}_0(\tilde{\omega})$ and $\hat{\omega}_0$.

In particular, for $s = 0$, we have

$$G^0\left(\omega_0|\{\mu_\tau\}_{\tau\leq n}\right) = G^0\left(\tilde{\omega}|\{\mu_\tau\}_{\tau\leq n}\right) \succeq \eta^2 e^{-2\eta S} \cdot V_n = 2\kappa^{-1} V_n,$$

with $\kappa := \frac{2e^{2\eta S}}{\eta^2}$.

Integrating the general term over $[0,1]$ also gives:

$$\Gamma^0_n(\tilde{\omega}) = \int_0^1 (1-s) G^0\left(\omega_s|\{\mu_\tau\}_{\tau\leq n}\right) \mathrm{d}s \succeq \frac{\eta^2 e^{-2\eta S}}{2} V_n = \kappa^{-1} V_n.$$

□

For simplicity, all results in this section will be stated for the μ-player. It is straightforward to extend them to the ν-player by a change of notation.

We first start by introducing the precise version of the warm-up phase in Algorithm 1 from the main text. We will then prove its guarantees and usefulness for our main IRL algorithm (Algorithm 1).

The overall idea is the following: suppose we have a strategy $w \in \Delta(\Pi_\mu)$ such that $\max_{\mu,b}\|g_\mu(b)\|^2_{V(w),*} \leq c$ for some constant c. Then, for any $N \in \mathbb{N}$, $\max_{\mu,b}\|g_\mu(b)\|^2_{N\cdot V(w),*} \leq c/N$. The norm can therefore be arbitrarily small as we increase N.

As w is a probability measure over policies, $N \cdot w$ can be seen as a fractional allocation, and an exact allocation when all masses $N \cdot w(\{\mu\})$ are integers. Instead of stating it as an assumption, we apply a rounding technique inspired by Allen-Zhu et al. (2021) in order to obtain a warm-up allocation v with $1^\top v = n_{\text{warm}}$ for some integer n_{warm}. This allocation determines a sequence $\{\mu^{\text{warm}}_\tau\}_{\tau\leq n_{\text{warm}}}$ that will be used to collect data (Algorithm 2, line 9).

Before explaining how we solve the warm-up optimization problem and obtain the allocation via rounding, let us start by extending Kiefer-Wolfowitz theorems for optimal design to our setting.

E.1 A GROUPED KIEFER-WOLFOWITZ THEORY

The classical Kiefer-Wolfowitz theorem (Kiefer & Wolfowitz, 1960) covers rank-one arms (e.g. see how it is used in Allen-Zhu et al. (2021) in the proof of Lemma 1), and therefore doesn't directly apply to our setting. In our case, we select a leader policy μ and observe (in expectation) a covariance matrix. Hence, a "grouped" update as the expectation of an outer product of averaged features. We thus need a grouped version of KW theory, that we give here in two steps: the trace value (Lemma 15) and the pointwise value needed by the warm-up (Corollary 1).

Lemma 15 (Grouped Kiefer-Wolfowitz)**.** *We have,*

$$\min_{w\in\Delta(\Pi_\mu)} \max_{\mu} \operatorname{Trace}\left(V(w)^\dagger M_\mu\right) = d^\diamond,$$

the minimum being attained at any $w_D \in \arg\max_{\tilde{w}} \log\det_{\mathcal{G}} V(\tilde{w}) \subseteq \mathcal{W}_{\mathcal{G}}$.

Proof. First of all, note that w_D must be admissible, since otherwise the objective is $+\infty$ (Equation (4)) and can be made finite by any admissible distribution, e.g. w_0 the uniform distribution over any basis of $\mathcal{G}$. Therefore,

$$\operatorname{rank} V(w_D) = d^\diamond, \qquad \operatorname{Trace}\left(V(w_D)^\dagger V(w_D)\right) = d^\diamond.$$

Also note that $\log\det_{\mathcal{G}} V(w_D) > -\infty$, and w_D is well defined: $w \mapsto V(w)$ has compact range and $\log\det_{\mathcal{G}}$ is upper semi-continuous on it.

Now, to prove the equality, we will prove the upper and lower bounds separately.

Upper bound

For any $\mu \in \Pi_\mu$ and $t \in [0,1)$, note that $(1-t)w_D + t\delta_\mu \in \mathcal{W}_{\mathcal{G}}$ and:

$$V((1-t)w_D + t\delta_\mu) = (1-t)V(w_D) + tM_\mu.$$

From the first-order optimality condition, and given $V(w_D) \in \mathbb{S}_{++}(\mathcal{G})$ and $M_\mu \in \mathbb{S}_+(\mathcal{G})$,

$$\begin{aligned}
0 &\geq \frac{d}{dt} \log\det_{\mathcal{G}} \left[V((1-t)w_D + t\delta_\mu)\right]\big|_{t=0+} \\
&= \frac{d}{dt} \log\det_{\mathcal{G}} \left[(1-t)V(w_D) + tM_\mu\right]\big|_{t=0+} \\
&= \operatorname{Trace}\left(V(w_D)^\dagger \left(M_\mu - V(w_D)\right)\right) \\
&= \operatorname{Trace}\left(V(w_D)^\dagger M_\mu\right) - d^\diamond.
\end{aligned}$$

Hence, $\max_\mu \operatorname{Trace}\left(V(w_D)^\dagger M_\mu\right) \leq d^\diamond$.

Lower bound

For any $\tilde{w} \in \mathcal{W}_{\mathcal{G}}$, we have:

$$\begin{aligned}
\max_{\mu} \operatorname{Trace}\left(V(\tilde{w})^\dagger M_\mu\right) &\geq \int_{\Pi_\mu} \operatorname{Trace}\left(V(\tilde{w})^\dagger M_\mu\right) \tilde{w}(\mathrm{d}\mu) \\
&= \operatorname{Trace}\left(V(\tilde{w})^\dagger V(\tilde{w})\right) \\
&= d^\diamond.
\end{aligned}$$

This implies the lower bound and concludes the proof. □

One direct consequence of Lemma 15 is that it allows us to directly bound the norm of the uniformly-centered features.

Corollary 1 (Pointwise grouped Kiefer-Wolfowitz)**.** *Using the same notation,*

$$\min_{w\in\Delta(\Pi_\mu)} \max_{\mu,b} \|g_\mu(b)\|^2_{V(w),*} \leq |\mathcal{B}| d^\diamond.$$

Proof. Take w_D the D-optimal design of Lemma 15. Then, for any μ, b:

$$\begin{aligned}\|g_\mu(b)\|^2_{V(w_D),*} &\le \sum_{b'\in\mathcal{B}} \|g_\mu(b')\|^2_{V(w_D),*} \\ &= \operatorname{Trace}\left(V(w_D)^\dagger \sum_{b'\in\mathcal{B}} g_\mu(b')g_\mu(b')^\top\right) \\ &= |\mathcal{B}|\operatorname{Trace}\left(V(w_D)^\dagger M_\mu\right) \\ &\le |\mathcal{B}|d^\diamond.\end{aligned}$$

Hence the upper bound is valid for the minimizer. □

E.2 Solving for the warm-up allocation

The warm-up needs an approximate optimal design w with $\max_{\mu,b}\|g_\mu(b)\|^2_{V(w),*} \le (1+\epsilon)|\mathcal{B}|d^\diamond$. By the proof of Corollary 1 it suffices to compute a design w_D whose grouped trace certificate is at most $(1+\epsilon)d^\diamond$.

We first note that it is sufficient for w_D to assign weights to pure policies, i.e. the optimization can be carried over $\Pi_\mu = \Delta(\mathcal{A})$ directly instead of the more complex $\Delta(\Pi_\mu) = \Delta(\Delta\mathcal{A})$. This is shown below.

Lemma 16 (Reduction to pure leader actions)**.** *Let δ_a denote the pure policy concentrated on action $a \in \mathcal{A}$. Then, for every $V \in \mathbb{S}_{++}(\mathcal{G})$,*

$$\max_{\mu\in\Pi_\mu} \operatorname{Trace}\left(V^\dagger M_\mu\right) = \max_{a\in\mathcal{A}} \operatorname{Trace}\left(V^\dagger M_{\delta_a}\right),$$

and,

$$\max_{\mu\in\Pi_\mu,\, b\in\mathcal{B}} \|g_\mu(b)\|^2_{V,*} = \max_{a\in\mathcal{A},\, b\in\mathcal{B}} \|g_{\delta_a}(b)\|^2_{V,*}.$$

Proof. Note that every policy $\mu \in \Pi_\mu$ is a mixture of pure policies:

$$\mu = \sum_{a\in\mathcal{A}} \mu(a)\delta_a \qquad \text{such that} \qquad g_\mu(b) = \sum_{a\in\mathcal{A}} \mu(a) g_{\delta_a}(b) = \langle \mu, \Phi_b\rangle,$$

with $\Phi_b = [g_{\delta_a}(b)]_{a\in\mathcal{A}}$.

Therefore, since $g_\mu(b) \in \mathcal{G}$,

$$\|g_\mu(b)\|^2_{V,*} = \mu^\top U_b \mu, \quad \operatorname{Trace}(V^\dagger M_\mu) = \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} \|g_\mu(b)\|^2_{V,*} = \mu^\top\left(\frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} U_b\right)\mu,$$

with $U_b \coloneqq \Phi_b^\top V^\dagger \Phi_b \succeq 0$

This concludes the proof as convex functions on a polytope attain their maximum at a vertex. □

This was the last key needed for building a Frank-Wolfe iterative algorithm to solve for the policy distribution by only optimizing over pure policies.

Specifically, for any running uniform coverage matrix $V(w)$ with admissible w, we can compute the directional derivative for any $w' \in \Delta(\Pi_\mu)$ as

$$k(w, w') \coloneqq \frac{\mathrm{d}}{\mathrm{d}t} \log\det{}_{\mathcal{G}}\, V(w + t(w' - w))\Big|_{t=0^+} = \int_{\Pi_\mu} h_w(\mu) w'(\mathrm{d}\mu) - d^\diamond,$$

where $h_w(\mu) \coloneqq \operatorname{Trace}\left(V(w)^\dagger M_\mu\right)$. By linearity, we can concentrate the mass of any optimal w' on a single policy $\mu^+ \in \Pi_\mu$, while by Lemma 16, this μ^+ can be taken as a pure policy. These two remarks reduce the optimization of w' to an optimization over $\mathcal{A}$:

$$d^\diamond + \max_{w'\in\Delta(\Pi_\mu)} k(w, w') = \max_{\mu^+\in\Pi_\mu} h_w(\mu^+) = \max_{a\in\mathcal{A}} h_w(\delta_a).$$

Algorithm 3 Frank–Wolfe for optimal design

Input: features $\{\phi(a,b)\}_{a\in\mathcal{A},b\in\mathcal{B}}$, tolerance $\epsilon > 0$.
Admissible initialization
Pick $d^\diamond$ pairs (a_i, b_i) with $\{g_{\delta_{a_i}}(b_i)\}_{i\le d^\diamond}$ a basis of $\mathcal{G}$.
Let w_0 be uniform on the distinct a_i.
for $t = 0, 1, 2, \dots$ **do**
  Linear oracle (enumeration over pure actions, Lemma 16)
  $V_t \leftarrow V(w_t)$; $\quad h_t(a) \leftarrow \mathrm{Trace}\left(V_t^\dagger M_{\delta_a}\right)$ for $a \in \mathcal{A}$; $\quad a_t \in \arg\max_a h_t(a)$.
  Duality gap and stopping test
  $\mathrm{gap}_t \leftarrow h_t(a_t) - d^\diamond$; $\quad$ **if** $\mathrm{gap}_t \le \epsilon\, d^\diamond$ **then return** w_t.
  Update
  $\gamma_t \leftarrow \arg\max_{\gamma\in[0,1]} \log\det_{\mathcal{G}}\left[(1-\gamma)V_t + \gamma M_{\delta_{a_t}}\right]$; $\quad w_{t+1} \leftarrow (1-\gamma_t)w_t + \gamma_t\delta_{a_t}$.
end for

With these simplifications, the algorithm is presented in Algorithm 3. Note that it is purely offline and costs $\mathcal{O}(|\mathcal{A}||\mathcal{B}|d^2)$ computations per iteration. Additionally, it terminates in finite time for every $\epsilon > 0$ by standard Frank-Wolfe analysis (see e.g. Zhao & Freund (2023)).

Running Algorithm 3 provides the following guarantee.

Proposition 2 (Certified output)**.** *If Algorithm 3 returns w_t, then* $\max_\mu \mathrm{Trace}\left(V(w_t)^\dagger M_\mu\right) \le (1+\epsilon)d^\diamond$ *and consequently*

$$\max_{\mu,b} \|g_\mu(b)\|^2_{V(w_t),*} \le (1+\epsilon)\,|\mathcal{B}|\,d^\diamond.$$

Proof. The stopping test gives

$$h_t(a_t) = \max_a \mathrm{Trace}(V(w_t)^\dagger M_{\delta_a}) \le (1+\epsilon)d^\diamond,$$

which extends to all $\mu \in \Pi_\mu$ by Lemma 16.

The result then follows exactly as in the proof of Corollary 1, replacing the bound $d^\diamond$ by $(1+\epsilon)d^\diamond$. □

E.3 Rounding and certificate

We now focus on turning the distribution over policies into an allocation strategy. Precisely, after obtaining a certified output from the Frank-Wolfe algorithm (Algorithm 3) with tolerance ϵ (Proposition 2), we apply a support reduction mechanism (Lemma 17) followed by integer rounding to turn the distribution over policies into an allocation (Definition 7).

Lemma 17 (Carathéodory support reduction)**.** *Let $w^\star$ be an output of Algorithm 3 satisfying $\max_{\mu,b}\|g_\mu(b)\|^2_{V(w^\star)^\dagger} = \rho < \infty$. Then, there exists a design w_0^* of support $s \le d^\diamond(d^\diamond+1)/2 + 1$ such that*

$$\max_{\mu,b} \|g_\mu(b)\|^2_{V(w_0^\star),*} = \rho,$$

and w_0^* can be computed efficiently from w^*.

Proof. Set $\mathcal{V} \coloneqq \mathrm{conv}\,\{M_\mu : \mu \in \Pi_\mu\}$ the convex hull spanned by uniform coverage matrices. Note that these matrices live in a convex cone of ambient dimension

$$D \coloneqq \dim\left(\mathbb{S}(\mathcal{G})\right) = \dim\left(\mathbb{S}^{d^\diamond}\right) = d^\diamond(d^\diamond+1)/2,$$

which follows from the isomorphism $\mathbb{S}(\mathcal{G}) \cong \mathbb{S}^{d^\diamond}$.

$\mathcal{V}$ is compact and convex as $\{M_\mu\}_\mu$ is the continuous image of the compact $\Pi_\mu = \Delta(\mathcal{A})$.

Therefore, by Carathéodory's theorem, for some $s \leq D+1$, there exists a set of policies $\mu_1, \ldots, \mu_s$ with associated matrices $M_{\mu_1}, \ldots, M_{\mu_s}$, such that:

$$\int_{\Pi_\mu} M_\mu w^*(\mathrm{d}\mu) = \sum_{i=1}^{s} \alpha_i M_{\mu_i}, \quad \text{for some } \alpha \in \Delta_{s-1}.$$

In particular, there exists a design w_0^* of support s such that $V(w^*) = V(w_0^*)$.

Now, note that by definition in the Frank-Wolfe procedure:

$$|\mathrm{supp}(w_{t+1})| \leq |\mathrm{supp}(w_t)| + 1,$$

thus $|\mathrm{supp}(w^*)| \leq d^\diamond + T$, where $T < \infty$ is the stopping time of Algorithm 3.

The coefficient w^* has finite support, and therefore we can apply the standard polynomial Carathéodory reduction (Grötschel et al., 1988) to recover some valid w_0^*.

This concludes the proof. □

This support reduction allows for the following simple rounding scheme.

Definition 7 (Warm-up allocation)**.** *Let $w_0^* = \{w_{0,i}^*\}_{i \leq s}$ be the support-reduced version of the output of Algorithm 3, obtained by Lemma 17. Define $o \in \mathbb{N}^s$ the order of w_0^* breaking ties by index such that:*

$$w_{0,o_1}^\star \geq \cdots \geq w_{0,o_s}^\star.$$

Given $n_{warm} \geq s$, the warm-up allocation is the vector $v \in \mathbb{N}^s$ defined by

$$v_{o_i} \coloneqq \underbrace{\left\lceil (n_{warm} - s)\, w_{0,o_i}^\star \right\rceil}_{base_{o_i}} + \mathbf{1}\left\{i \leq \Delta\right\}, \qquad \Delta \coloneqq n_{warm} - \sum_{j \leq s} base_j.$$

That is, the residual samples are given to the heaviest support points and $1^\top v = n_{warm}$.

Remark 3. *We do not argue that this is the most efficient rounding procedure, but this is simple enough for our algorithm.*

Definition 8 (Allocation uniform coverage)**.** *Let v be the warm-up allocation with arbitrary enumeration $\{\mu_\tau^{warm}\}_{\tau \leq n_{warm}}$. We denote the allocation uniform coverage by:*

$$V_{warm} \coloneqq \sum_{\tau=1}^{n_{warm}} M_{\mu_\tau^{warm}} = \frac{1}{|\mathcal{B}|} \sum_{\tau=1}^{n_{warm}} \sum_{b \in \mathcal{B}} g_{\mu_\tau^{warm}}(b) g_{\mu_\tau^{warm}}(b)^\top.$$

Note that V_{warm} is not normalized, so that its dual norm shrinks when n_{warm} grows. We make that precise in the lemma below.

Before introducing the lemma, let us define the following term:

$$C_{\text{rounding}} \coloneqq \epsilon^{-1}(1+\epsilon)s \leq (2\epsilon)^{-1}(1+\epsilon)(d^\diamond(d^\diamond+1)+2).$$

Lemma 18 (Warm-up schedule trace guarantee)**.** *Let V_{warm} denote the allocation uniform coverage. Then, if $n_{warm} \geq C_{rounding}$,*

$$\max_\mu \mathrm{Trace}\left(V_{warm}^\dagger M_\mu\right) \leq \frac{(1+\epsilon)^2 d^\diamond}{n_{warm}} \leq \frac{(1+3\epsilon)d^\diamond}{n_{warm}},$$

the last step since $\epsilon \leq 1$.

Proof. By Proposition 2, $\max_\mu \mathrm{Trace}\left(V(w^*)^\dagger M_\mu\right) \leq \rho$ with $\rho \coloneqq (1+\epsilon)d^\diamond$. Apply the support-reduction of Lemma 17 and the rounding of Definition 7 to obtain the density and allocation w_0^* and v.

Now note that the rounding is such that,

$$v_i \geq (n_{\text{warm}} - s) w_{0,i}^*, \qquad \forall i \in [s],$$

and therefore

$$V_{\text{warm}} = \sum_{\tau=1}^{n_{\text{warm}}} M_{\mu_\tau^{\text{warm}}} = \sum_{i \le s} v_i M_{\mu_i} \succeq (n_{\text{warm}} - s) V(w_0^*),$$

with $\{\mu_i\}_{i \le s}$ the policies in the support of the allocation.

Hence, for any μ (elementary property in $\mathcal{G}$, see Lemma 4),

$$\text{Trace}\left(V_{\text{warm}}^\dagger M_\mu\right) \le \frac{\text{Trace}\left(V(w_0^*)^\dagger M_\mu\right)}{n_{\text{warm}} - s} = \frac{\text{Trace}\left(V(w^*)^\dagger M_\mu\right)}{n_{\text{warm}} - s} \le \frac{\rho}{n_{\text{warm}} - s},$$

where the equality is by Lemma 17.

Replacing $\rho = (1+\epsilon)d^\diamond$ and using the condition $n_{\text{warm}} \ge C_{\text{rounding}}$,

$$n_{\text{warm}} \ge \frac{1+\epsilon}{\epsilon} s \Leftrightarrow \epsilon(n_{\text{warm}} - s) \ge s \Leftrightarrow (1+\epsilon)(n_{\text{warm}} - s) \ge n_{\text{warm}},$$

we conclude:

$$\frac{\rho}{n_{\text{warm}} - s} = \frac{(1+\epsilon)d^\diamond}{n_{\text{warm}} - s} \le \frac{(1+\epsilon)^2 d^\diamond}{n_{\text{warm}}} \le \frac{(1+3\epsilon)d^\diamond}{n_{\text{warm}}},$$

the last step using $(1+\epsilon)^2 = 1 + 2\epsilon + \epsilon^2 \le 1 + 3\epsilon$ for $\epsilon \le 1$. □

Corollary 2 (Warm-up schedule norm guarantee)**.** *Let V_{warm} denote the allocation uniform coverage. Then, if $n_{warm} \ge C_{rounding}$,*

$$\max_{\mu, b} \|g_\mu(b)\|^2_{V_{warm},*} \le \frac{(1+3\epsilon)|\mathcal{B}|d^\diamond}{n_{warm}}.$$

Proof. Using a similar argument to the proof of Corollary 1, for any μ, b,

$$\begin{aligned}
\|g_\mu(b)\|^2_{V_{\text{warm}},*} &= g_\mu(b)^\top V_{\text{warm}}^\dagger g_\mu(b) \\
&\le \sum_{b' \in \mathcal{B}} g_\mu(b')^\top V_{\text{warm}}^\dagger g_\mu(b') \\
&= |\mathcal{B}| \,\text{Trace}(V_{\text{warm}}^\dagger M_\mu) \\
&\le \frac{(1+3\epsilon)|\mathcal{B}|d^\diamond}{n_{\text{warm}}},
\end{aligned}$$

where the last inequality follows from Lemma 18. □

Finally, we can combine these results with the maximum likelihood estimation results to provide a certificate over sequence of MLE estimators.

This lemma depends on a new constant $C_{\text{certificate}}$ that we define as follows:

$$\begin{aligned}
C_{\text{certificate}} &\coloneqq 8\kappa(1+3\epsilon)\eta^2|\mathcal{B}|d^\diamond d \left(\log_+\left(\frac{4e\eta SK}{\delta d}\right) + \log\left(4\kappa(1+3\epsilon)\eta^2|\mathcal{B}|d^\diamond d\right)\right) \\
&\le \mathcal{O}\left(\kappa\eta^2|\mathcal{B}|d^\diamond d \log\left(\frac{\kappa\eta S|\mathcal{B}|d^\diamond K}{\delta}\right)\right) \\
&\le \tilde{\mathcal{O}}\left(e^{2\eta S}|\mathcal{B}|d^\diamond d \log\left(\frac{e^{\eta S}}{\delta}\right)\right),
\end{aligned}$$

where $\kappa = 2e^{2\eta S}/\eta^2$.

Lemma 19 (Warm-up certificate)**.** *Let δ small enough and set $n_{warm} \ge \max\{C_{rounding}, C_{certificate}\}$. Then, with probability at least $1-\delta$, the warm-up ensures $d_*(\omega, \hat{\omega}_k^{on}) \le 1/\eta$ for any $0 \le k \le K$.*

Proof. For any μ, k, we can decompose using the generalized Cauchy-Schwarz (Lemma 4):

$$d_\mu(\omega, \hat{\omega}_k^{\text{on}}) \leq \|\omega - \hat{\omega}_k^{\text{on}}\|_{V_{\text{warm}}} \cdot \max_{b,d} \|\psi_\mu(b) - \psi_\mu(d)\|_{V_{\text{warm}},*},$$

where V_{warm} is the uniform allocation coverage (Definition 7).

For the first term, we combine the design curvature (Lemma 14) with the confidence sets (Theorem 5). Using $\kappa = 2e^{2\eta S}/\eta^2$,

$$\begin{aligned}
\|\omega - \hat{\omega}_k^{\text{on}}\|_{V_{\text{warm}}}^2 &\leq \kappa \|\omega - \hat{\omega}_k^{\text{on}}\|^2_{\Gamma^0_{n_{\text{warm}}}(\hat{\omega}_k^{\text{on}})} \\
&\leq \kappa \|\omega - \hat{\omega}_k^{\text{on}}\|^2_{\Gamma^0_{n_{\text{warm}}+k}(\hat{\omega}_k^{\text{on}})} \\
&\leq \kappa \left(\log \delta^{-1} + d \log_+ \left(4ed^{-1} \eta S (n_{\text{warm}} + K - 1) \right) \right) \\
&:= L(\delta),
\end{aligned}$$

where we define $\log_+(x) := \log(\max\{e, x\})$.

The analysis of the second term comes from the optimal design and the warm-up schedule. Specifically, using $n_{\text{warm}} \geq C_{\text{rounding}}$ we apply Lemma 2 and get:

$$\max_{\mu,b} \|g_\mu(b)\|_{V_{\text{warm}},*} \leq \sqrt{(1+3\epsilon)|\mathcal{B}|d^\diamond / n_{\text{warm}}}.$$

Since $\psi_\mu(b) - \psi_\mu(d) = g_\mu(b) - g_\mu(d)$, we get from the triangle inequality that for all μ:

$$\max_{b,d} \|\psi_\mu(b) - \psi_\mu(d)\|_{V_{\text{warm}},*} = \max_{b,d} \|g_\mu(b) - g_\mu(d)\|_{V_{\text{warm}},*} \leq 2 \cdot \max_b \|g_\mu(b)\|_{V_{\text{warm}},*}.$$

Hence, $\max_{b,d} \|\psi_\mu(b) - \psi_\mu(d)\|_{V_{\text{warm}},*} \leq 2 \cdot \sqrt{\frac{(1+3\epsilon)|\mathcal{B}|d^\diamond}{n_{\text{warm}}}}$.

Combining the two terms, we conclude for every $k \in [0, K]$,

$$d_*(\omega, \hat{\omega}_k^{\text{on}}) \leq \sqrt{L(\delta)} \cdot 2\sqrt{\frac{(1+3\epsilon)|\mathcal{B}|d^\diamond}{n_{\text{warm}}}} = 2\sqrt{\frac{(1+3\epsilon)|\mathcal{B}|d^\diamond L(\delta)}{n_{\text{warm}}}}.$$

Hence with high-probability (the event holds for all k without needing a union bound),

$$d_*(\omega, \hat{\omega}_k^{\text{on}}) \leq 1/\eta, \qquad \text{if } n_{\text{warm}} \geq 4(1+3\epsilon)\eta^2 L(\delta)|\mathcal{B}|d^\diamond.$$

Now let us expand and simplify $L(\delta)$ as follows (using $n_{\text{warm}}, K \geq 1$):

$$\begin{aligned}
L(\delta) &= \kappa \left(\log\left(\frac{1}{\delta}\right) + d \log_+ \left(\frac{4e\eta S(n_{\text{warm}} + K - 1)}{d} \right) \right) \\
&\leq \kappa d \left(\log_+ \left(\frac{4e\eta SK}{\delta d} \right) + \log(n_{\text{warm}}) \right).
\end{aligned}$$

Therefore, a sufficient condition is $n_{\text{warm}} \geq a(b + \log n_{\text{warm}})$ with:

$$a = 4\kappa(1+3\epsilon)\eta^2|\mathcal{B}|d^\diamond d \qquad \text{and} \qquad b = \log_+ \left(\frac{4e\eta SK}{\delta d} \right).$$

Applying Lemma 25 we obtain the following sufficient condition:

$$\begin{aligned}
n_{\text{warm}} &\geq 2a(b + \log a) \\
&= 8\kappa(1+3\epsilon)\eta^2|\mathcal{B}|d^\diamond d \left(\log_+ \left(\frac{4e\eta SK}{\delta d} \right) + \log\left(4\kappa(1+3\epsilon)\eta^2|\mathcal{B}|d^\diamond d\right) \right)
\end{aligned}$$

if $\delta < 16\kappa(1+3\epsilon)\eta^3 S|\mathcal{B}|d^\diamond K$.

We can therefore conclude from the condition $n_{\text{warm}} \geq C_{\text{certificate}}$. □

F Analysis of the main algorithm

Following the proof roadmap in the main text, we analyze the algorithm by first applying the Stackelberg reward transfer lemma (Lemma 20), and then combine the following three ingredients.

i) A confidence sequence for finite-sample estimation We can bound the coefficient distance terms by relating to high probability confidence sets of Generalized Linear Models (GLMs).

Specifically, following (Lee et al., 2024; Schlaginhaufen et al., 2026), we can derive a bound of the following form (Theorem 5):

$$\mathbb{P}\left(\left\{\|\omega - \hat{\omega}_\tau\|_{\Gamma^\lambda_\tau(\hat{\omega}_\tau)} \leq \beta_\tau(\delta),\ \forall \tau \in [N]\right\}\right) \geq 1-\delta,$$

for some $\beta^2_\tau(\delta) \leq \mathcal{O}\left(\lambda S^2 + \log \delta^{-1} + d\log(\max\left\{e, \eta\tau S d^{-1}\right\})\right)$.

This allows, with one good event, to control the concentration of the MLE estimators in both the warm-up and online phases.

ii) Minimizing uncertainty in the Fisher information geometry and the elliptical potential lemma We bound Fisher uncertainty by a greedy selection rule maximizing Fisher information gain. With a self-concordance reparameterization bound (see point iii) below), we apply a variant of the standard elliptical potential lemma (Abbasi-Yadkori et al., 2011) and show how it vanishes at a rate $\approx \sqrt{\log(K)/K}$.

iii) Self-concordance The instantaneous log-likelihood as a function of the logits is self-concordant-like in the span semi-norm. This allows to control the reparameterization of Fisher information matrices. Precisely, an optimal schedule for the leader policies depends on knowing the true coefficient ω. As this is unknown, we instead replace it in the optimization objective by $\tilde{\omega} = 0$ in the warm-up phase, and by the running estimate $\tilde{\omega} = \hat{\omega}^{\text{on}}_k$ during the online phase. More information on the self-concordance properties in Appendix D.

Crucially, this last tool introduces exponential dependency in the induced distance between the proxy $\tilde{\omega}$ and ω. This is the reason for the warm-up phase, which removes this dependency in the final bound by using a burn-in budget. As mentioned, in this offline part we use an arbitrary proxy $\tilde{\omega} = 0$ and rely on global self-concordance. Hence, our offline warm-up would always produce an exponential sample complexity guarantee. We solve this problem by using a subsequent online phase in the improved self-concordance regime induced by the warm-up certificate (Lemma 19).

F.1 Reward transfer lemma

Lemma 20 (Stackelberg reward transfer)**.** *Let μ^* and $\hat{\mu}$ be optimal leaders under ω and $\hat{\omega}$, respectively. Then,*

$$\mathrm{SubOpt}_\omega(\hat{\mu}) \leq 2R_{max} \cdot \max_{\tilde{\mu}\in\{\mu^*,\hat{\mu}\}} \chi\left(\eta \cdot d_{\tilde{\mu}}(\omega,\hat{\omega})\right) \cdot \|\omega - \hat{\omega}\|_{\mathcal{I}_{\tilde{\mu}}(\hat{\omega})},$$

with $\chi(x) := \frac{2(e^{x/2}-1)}{x}$.

Proof.

$$\begin{aligned}
\mathrm{SubOpt}_\omega(\hat{\mu}) &= J^{\text{leader}}_\omega(\mu^*) - J^{\text{leader}}_\omega(\hat{\mu}) \\
&\leq \left[J^{\text{leader}}_\omega(\mu^*) - J^{\text{leader}}_{\hat{\omega}}(\mu^*)\right] + \left[J^{\text{leader}}_{\hat{\omega}}(\mu^*) - J^{\text{leader}}_{\hat{\omega}}(\hat{\mu})\right] \\
&\qquad + \left[J^{\text{leader}}_{\hat{\omega}}(\hat{\mu}) - J^{\text{leader}}_\omega(\hat{\mu})\right] \\
&\leq \mathrm{SubOpt}_{\hat{\omega}}(\hat{\mu}) + 2 \cdot \max_{\tilde{\mu}\in\{\hat{\mu},\mu^*\}} \left|J^{\text{leader}}_\omega(\tilde{\mu}) - J^{\text{leader}}_{\hat{\omega}}(\tilde{\mu})\right| \\
&\leq 2R_{\max} \cdot \max_{\tilde{\mu}\in\{\hat{\mu},\mu^*\}} \left\|\nu^*_{\tilde{\mu}}(\omega) - \nu^*_{\tilde{\mu}}(\hat{\omega})\right\|_1,
\end{aligned}$$

where the last inequality uses the optimality of $\hat{\mu}$ under $\hat{\omega}$.

Then applying Lemma 12 gives the final result. □

Remark 4. *The upper bound of Lemma 20 vanishes exactly when $\hat{r}_2$ is feasible at both evaluated leaders, i.e.*

$$\max_{\tilde{\mu}\in\{\mu^*,\hat{\mu}\}} \|\omega-\hat{\omega}\|_{\mathcal{I}_{\tilde{\mu}}(\hat{\omega})} = 0 \quad \Longleftrightarrow \quad \hat{r}_2 \in \mathcal{R}_2^{\text{feas}}(\mu^*) \cap \mathcal{R}_2^{\text{feas}}(\hat{\mu}),$$

for $\mathcal{R}_2^{\text{feas}}(\mu) = \{\tilde{r}_2 : \nu_\mu^*(\tilde{r}_2) = \nu_\mu^*(r_2)\}$.

Using our specific choice of leader policies in Algorithm 1, we are able to bound the two terms of the upper bound separately. This leads to our main result.

F.2 Finite-sample Stackelberg sub-optimality guarantee

Before deriving the proof, let us quickly note about the approximation given in the main paper for the burn-in phase.

Placing us in the same regime of Lemma 19, we will require $n_{\text{warm}} \geq \max\{C_{\text{rounding}}, C_{\text{certificate}}\}$.

Two things are in order:

1. Typically, $C_{\text{rounding}} \ll C_{\text{certificate}}$.
2. Since for all features we have $d^\diamond \leq d$, we can further bound

$$C_{\text{certificate}} \leq \tilde{\mathcal{O}}\left(e^{2\eta S}|\mathcal{B}|d^\diamond d \log\left(\frac{e^{\eta S}}{\delta}\right)\right) \leq \tilde{\mathcal{O}}\left(\eta S e^{2\eta S}|\mathcal{B}|d^2 \log\frac{1}{\delta}\right).$$

This motivates our approximation $n_{\text{warm}} \approx \tilde{\mathcal{O}}\left(\eta S e^{2\eta S}|\mathcal{B}|d^2 \log \delta^{-1}\right)$ for the informal statement, only using terms introduced in the main text.

Theorem 4. *Set $n_{warm} = \max\{C_{rounding}, C_{certificate}\}$, $\lambda = 0$ and let $\hat{\omega}$ be the result of Algorithm 1 after $K \geq 1$ online iterations. Then, with probability at least $1-\delta$, for any leader $\hat{\mu}$ such that $\mathrm{SubOpt}_{\hat{\omega}}(\hat{\mu}) = 0$:*

$$\mathrm{SubOpt}_{\omega}(\hat{\mu}) \leq \tilde{\mathcal{O}}\left(\frac{R_{max}\sqrt{d^\diamond\left(d\eta S + \log\delta^{-1} + d\right)}}{\sqrt{K}}\right),$$

where $\tilde{\mathcal{O}}$ hides polylog terms in $\log\delta^{-1}$, K and environment parameters.

Proof. In the following, we assume the certificate over the entire sequence of maximum likelihood estimators, which occurs with probability at least $1-\delta$, assuming a sufficient burn-in phase budget (Lemma 19).

Therefore, combining the reward transfer lemma (Lemma 20) with the warm-up result (Lemma 19), the bound reduces to the coefficient approximation in the Fisher information norm:

$$\mathrm{SubOpt}_{\omega}(\hat{\mu}) \leq 2\chi(1)\, R_{\max} \cdot \max_{\tilde{\mu}\in\{\mu^*,\hat{\mu}\}} \|\omega-\hat{\omega}\|_{\mathcal{I}_{\tilde{\mu}}(\hat{\omega})},$$

where $\chi(1) = 2(\sqrt{e}-1) \approx 1.297$ and $\hat{\omega} = \hat{\omega}_K^{\text{on}}$.

Now, recall that for any k and λ, the curvature of the log-likelihood corresponds to the cumulative information collected up to online iteration k:

$$-\nabla^2\mathcal{L}_k(\tilde{\omega}) = \eta^2 \sum_{\tau=1}^{N_k} \mathbb{E}_{b\sim\mathbb{P}(\cdot|\tilde{\omega},\mu_\tau)}\left[\left(\psi_{\mu_\tau}(b) - \overline{\psi}_{\mu_\tau}(\tilde{\omega})\right)^{\otimes 2}\right] = \sum_{\tau=1}^{N_k} \mathcal{I}_{\mu_\tau}(\tilde{\omega}),$$

and in its regularized form $G_k^\lambda(\tilde{\omega}) \coloneqq \lambda I - \nabla^2\mathcal{L}_k(\tilde{\omega})$.

By Lemma 18, the warm-up uniform allocation must be admissible. Hence, for all $k \in [K]$ we have $G_k^0(\tilde{\omega}) \in \mathbb{S}_{++}(\mathcal{G})$, i.e. the warm-up phase guarantees that the unregularized total information matrix covers the space $\mathcal{G}$ for all k. Therefore, setting $\lambda = 0$, we simplify notations by denoting $G_k \coloneqq G_k^0$.

For any $\tilde{\mu} \in \{\mu^*, \hat{\mu}\}$, under the certificate,

$$\begin{aligned}
\|\omega - \hat{\omega}\|^2_{\mathcal{I}_{\tilde{\mu}}(\hat{\omega})} &= (\omega - \hat{\omega})^\top \mathcal{I}_{\tilde{\mu}}(\hat{\omega})(\omega - \hat{\omega}) \\
&\leq \|\omega - \hat{\omega}\|^2_{G_K(\hat{\omega})} \cdot \mathrm{Trace}\left(G_K^\dagger(\hat{\omega})\mathcal{I}_{\tilde{\mu}}(\hat{\omega})\right) && \text{(Lemma 21 in } \mathcal{G}) \\
&\leq 3\,\|\omega - \hat{\omega}\|^2_{\Gamma_K(\hat{\omega})} \cdot \mathrm{Trace}\left(G_K^\dagger(\hat{\omega})\mathcal{I}_{\tilde{\mu}}(\hat{\omega})\right) && \text{(Lemma 10, warm-up)} \\
&\leq 3\beta^2_{N_K}(\delta) \cdot \mathrm{Trace}\left(G_K^\dagger(\hat{\omega})\mathcal{I}_{\tilde{\mu}}(\hat{\omega})\right). && \text{(Theorem 5)}
\end{aligned}$$

This last term is controlled by our greedy design.

Indeed, first notice that:

$$\begin{aligned}
&\mathrm{Trace}\left(G_K^\dagger(\hat{\omega})\mathcal{I}_{\tilde{\mu}}(\hat{\omega})\right) \\
&\leq \frac{1}{K}\sum_{k=1}^{K} \mathrm{Trace}\left(G_{k-1}^\dagger(\hat{\omega})\mathcal{I}_{\tilde{\mu}}(\hat{\omega})\right) && \text{(non-increasing sequence)} \\
&\leq \frac{1}{K}\sum_{k=1}^{K} e^{2\eta d_*(\hat{\omega},\hat{\omega}_{k-1})}\,\mathrm{Trace}\left(G_{k-1}^\dagger(\hat{\omega}^{\text{on}}_{k-1})\mathcal{I}_{\tilde{\mu}}(\hat{\omega}^{\text{on}}_{k-1})\right) && \text{(self-concordance)} \\
&\leq \frac{e^{2\eta d_*(\hat{\omega},\omega)}}{K}\sum_{k=1}^{K} e^{2\eta d_*(\omega,\hat{\omega}^{\text{on}}_{k-1})}\,\mathrm{Trace}\left(G_{k-1}^\dagger(\hat{\omega}^{\text{on}}_{k-1})\mathcal{I}_{\tilde{\mu}}(\hat{\omega}^{\text{on}}_{k-1})\right) && \text{(triangle inequality)} \\
&\leq \frac{e^4}{K}\sum_{k=1}^{K} \mathrm{Trace}\left(G_{k-1}^\dagger(\hat{\omega}^{\text{on}}_{k-1})\mathcal{I}_{\tilde{\mu}}(\hat{\omega}^{\text{on}}_{k-1})\right). && \text{(certificate)}
\end{aligned}$$

Then, from our selection rule we get:

$$\begin{aligned}
\mathrm{Trace}(G_K^\dagger(\hat{\omega})\mathcal{I}_{\tilde{\mu}}(\hat{\omega})) &\leq \frac{e^4}{K}\sum_{k=1}^{K} \mathrm{Trace}(G_{k-1}^\dagger(\hat{\omega}^{\text{on}}_{k-1})\mathcal{I}_{\mu^{\text{on}}_k}(\hat{\omega}^{\text{on}}_{k-1})) && \text{(selection rule)} \\
&\leq \frac{e^6}{K}\sum_{k=1}^{K} \mathrm{Trace}(G_{k-1}^\dagger(\omega)\mathcal{I}_{\mu^{\text{on}}_k}(\omega)) && \text{(certificate)}
\end{aligned}$$

We will then apply the elliptical potential lemma (see classical isotropic version as Lemma 24) in the information geometry (Corollary 3), leveraging the warm-up phase.

Specifically, let $U_0(\cdot) \coloneqq G^0\left(\cdot|\mathcal{D}_{\text{warm}}\right)$ be the warm-up cumulative Fisher information. From the warm-up guarantees, the information uncertainty is bounded as follows:

$$\begin{aligned}
\sup_{\mu\in\Pi_\mu} \mathrm{Trace}\left(U_0(\omega)^\dagger \mathcal{I}_\mu(\omega)\right) &\leq \frac{\kappa}{2} \cdot \sup_{\mu\in\Pi_\mu} \mathrm{Trace}\left(V^\dagger_{\text{warm}}\mathcal{I}_\mu(\omega)\right) && \text{(Lemma 14)} \\
&\leq \frac{\kappa\eta^2|\mathcal{B}|}{2} \cdot \sup_{\mu\in\Pi_\mu} \mathrm{Trace}\left(V^\dagger_{\text{warm}}M_\mu\right) && \text{(second moment)} \\
&\leq \frac{\kappa\eta^2|\mathcal{B}|}{2} \cdot \frac{(1+3\epsilon)d^\diamond}{n_{\text{warm}}}. && \text{(Lemma 18)}
\end{aligned}$$

We define this last term as $q_{\text{warm}} \coloneqq \frac{(1+3\epsilon)\kappa d^\diamond \eta^2|\mathcal{B}|}{2n_{\text{warm}}} \leq \frac{1}{16d} \leq 1$, where the simplification is from the condition $n_{\text{warm}} \geq C_{\text{certificate}}$.

We can then apply the elliptical potential lemma in $\mathcal{G}$ (Corollary 3) with $U_0 \coloneqq U_0(\omega)$ and $A_k = \mathcal{I}_{\mu^{\text{on}}_k}(\omega)$ and obtain

$$\mathrm{Trace}\left(G_K^\dagger(\hat{\omega})\mathcal{I}_{\tilde{\mu}}(\hat{\omega})\right) \leq \frac{2d^\diamond e^6 \log\left(1 + \frac{q_{\text{warm}}K}{d^\diamond}\right)}{K} \leq \frac{2d^\diamond e^6 \log\left(1 + \frac{K}{16d^\diamond d}\right)}{K}.$$

To simplify notations, set $C \coloneqq \sqrt{d^\diamond \log\left(1 + \frac{K}{16 d^\diamond d}\right)}$ and use $\log_+(x) \coloneqq \log\left(\max\{e, x\}\right)$. Then, putting everything together, we get with probability at least $1-\delta$,

$$\begin{aligned}
\mathrm{SubOpt}_\omega(\hat{\mu}) &\leq 2\chi(1) R_{\max} \sqrt{3\beta_{N_K}^2(\delta)\, \mathrm{Trace}\left(G_K^\dagger(\hat{\omega}) \mathcal{I}_{\tilde{\mu}}(\hat{\omega})\right)} \\
&\leq \frac{9e^3 C R_{\max}}{\sqrt{2K}} \sqrt{\log \delta^{-1} + d \log_+\left(\frac{4e\eta N_K S}{d}\right)} \\
&\leq \frac{9e^3 C R_{\max}}{\sqrt{2K}} \sqrt{\log \delta^{-1} + d \log\left(\frac{4e\eta n_{\text{warm}} S}{d}\right) + d \log_+\left(\frac{4e\eta K S}{d}\right)} \\
&\leq \tilde{\mathcal{O}}\left(\frac{C R_{\max}}{\sqrt{K}} \sqrt{\log \delta^{-1} + d \log\left(e^{2\eta S} \log\left(\frac{e^{2\eta S}}{\delta}\right)\right) + d}\right) \\
&\leq \tilde{\mathcal{O}}\left(\frac{R_{\max} \sqrt{d^\diamond \left(d\eta S + \log \delta^{-1} + d \log\log \delta^{-1}\right)}}{\sqrt{K}}\right) \\
&\leq \tilde{\mathcal{O}}\left(\frac{R_{\max} \sqrt{d^\diamond \left(d\eta S + \log \delta^{-1} + d\right)}}{\sqrt{K}}\right).
\end{aligned}$$

□

G USEFUL RESULTS

Theorem 5 (Corollary of Theorem 3.1 in Lee et al. (2024)). *Fix $\lambda \geq 0$ and $\delta \in (0,1)$. Then, with probability at least $1-\delta$,*

$$\|\omega - \hat{\omega}_\tau\|^2_{\Gamma^\lambda_\tau(\hat{\omega}_\tau)} \leq \beta^2_\tau(\delta) \coloneqq 2\lambda S^2 + \log\frac{1}{\delta} + d \log\left(\max\left\{e, \frac{4(\tau-1)e\eta S}{d}\right\}\right),$$

for all $\tau \in [N]$, where we define $\Gamma^\lambda_\tau(\tilde{\omega}) \coloneqq \int_0^1 (1-s) G^\lambda\left(\tilde{\omega} + s(\omega - \tilde{\omega}) | \mathcal{D}_{1:\tau}\right) \mathrm{d}s$.

Proof. Combining a first-order Taylor expansion and the optimality of $\hat{\omega}_\tau$ with respect to $\mathcal{L}\left(\cdot | \mathcal{D}_{1:\tau}\right)$,

$$\begin{aligned}
\mathcal{L}\left(\omega | \mathcal{D}_{1:\tau}\right) - \mathcal{L}\left(\hat{\omega}_\tau | \mathcal{D}_{1:\tau}\right) &= \left\langle \nabla \mathcal{L}\left(\hat{\omega}_\tau | \mathcal{D}_{1:\tau}\right), \omega - \hat{\omega}_\tau \right\rangle \\
&\quad + (\omega - \hat{\omega}_\tau)^\top \left(\int_0^1 (1-s) \nabla^2 \mathcal{L}\left(\hat{\omega}_\tau + s(\omega - \hat{\omega}_\tau) | \mathcal{D}_{1:\tau}\right) \mathrm{d}s\right) (\omega - \hat{\omega}_\tau) \\
&\leq \frac{\lambda}{2} \|\omega - \hat{\omega}_\tau\|_2^2 - \|\omega - \hat{\omega}_\tau\|^2_{\Gamma^\lambda_\tau(\hat{\omega}_\tau)} \\
&\leq 2\lambda S^2 - \|\omega - \hat{\omega}_\tau\|^2_{\Gamma^\lambda_\tau(\hat{\omega}_\tau)}.
\end{aligned}$$

Now, note that $\mathcal{L}\left(\cdot | \mathcal{D}_{1:\tau}\right)$ is $2\eta(\tau-1)$-Lipschitz by Rademacher's theorem.

We can then apply Theorem 3.1 in Lee et al. (2024) and conclude. □

Lemma 21 (Trace splitting). *Let $x \in \mathbb{R}^d$, $A \succeq 0$, and $B \succ 0$. Then*

$$\|x\|_A^2 \leq \mathrm{Trace}\left(B^{-1} A\right) \cdot \|x\|_B^2.$$

Proof. Set $y \coloneqq B^{1/2} x$, so that $\|x\|_B^2 = \|y\|_2^2$ and

$$\|x\|_A^2 = x^\top A x = y^\top \left(B^{-1/2} A B^{-1/2}\right) y \leq \lambda_{\max}\left(B^{-1/2} A B^{-1/2}\right) \|y\|_2^2.$$

Now since $B \succ 0$, we note that:

$$\lambda_{\max}\left(B^{-1/2} A B^{-1/2}\right) \leq \mathrm{Trace}\left(B^{-1/2} A B^{-1/2}\right) = \mathrm{Trace}\left(B^{-1} A\right),$$

which gives the claim. □

Lemma 22 (Divergence decomposition)**.** *Let $K \in \mathbb{Z}_{>0}$ and let $Q^{(1)}, Q^{(2)}$ be two kernels from $\mathcal{X}$ to a common measurable space such that $Q^{(1)}(\cdot|x) \ll Q^{(2)}(\cdot|x)$ for every $x \in \mathcal{X}$. Define a common kernel family $\{\pi_k\}_{k \le K}$ inducing*

$$\mathbb{P}^{(j)} = \mathrm{Law}\left(\tau_K \coloneqq (x_1, y_1, \ldots, x_K, y_K)\right), \qquad y_k \sim Q^{(j)}(\cdot|x_k), \qquad \forall j \in \{1, 2\},$$

with $x_k \sim \pi_k(\cdot \mid \tau_{<k})$ for the sub-trajectory $\tau_{<k} \coloneqq (x_1, y_1, \ldots, x_{k-1}, y_{k-1})$.

Then,

$$D_{\mathrm{KL}}\left(\mathbb{P}^{(1)} \,\middle\|\, \mathbb{P}^{(2)}\right) = \sum_{k=1}^{K} \mathbb{E}^{(1)}\left[D_{\mathrm{KL}}\left(Q^{(1)}(\cdot|x_k) \,\middle\|\, Q^{(2)}(\cdot|x_k)\right)\right].$$

Proof. By construction, the trajectory laws factorize as:

$$\mathbb{P}^{(j)}(\mathrm{d}\tau_K) = \prod_{k=1}^{K} \pi_k(\mathrm{d}x_k \mid \tau_{<k})\, Q^{(j)}(\mathrm{d}y_k \mid x_k),$$

so that $\mathbb{P}^{(1)} \ll \mathbb{P}^{(2)}$.

Taking the likelihood ratio, the common kernels cancel:

$$\log \frac{\mathrm{d}\mathbb{P}^{(1)}}{\mathrm{d}\mathbb{P}^{(2)}}(\tau_K) = \sum_{k=1}^{K} \log \frac{\mathrm{d}Q^{(1)}(\cdot \mid x_k)}{\mathrm{d}Q^{(2)}(\cdot \mid x_k)}(y_k), \qquad \mathbb{P}^{(1)}\text{-a.s.}$$

Taking expectations on both sides and using the tower property and substitution law for conditional expectations give the final result. □

Lemma 23 (KL divergence of softmax policies)**.** *Let $v, v' \in \mathbb{R}^d$ and $\pi = \mathrm{softmax}(v), \pi' = \mathrm{softmax}(v')$ the two corresponding softmax policies with inverse temperature η. Then,*

$$D_{\mathrm{KL}}(\pi||\pi') \le \frac{\eta^2}{8}\|v - v'\|^2_{\mathit{span}}$$

where $\|x\|_{\mathit{span}} = \max_a x_a - \min_a x_a$.

Proof. Define the partition terms Z, Z' such that for any $1 \le i \le d$,

$$\pi_i = e^{\eta v_i}/Z \qquad \text{and} \qquad \pi'_i = e^{\eta v'_i}/Z'.$$

Using the notation $\Delta_i = v'_i - v_i$, notice that:

$$Z' = \sum_i e^{\eta v'_i} = \sum_i e^{\eta v_i} \cdot e^{\eta \Delta_i} = Z\mathbb{E}_\pi\left[e^{\eta\Delta_i}\right].$$

Therefore,

$$\begin{aligned}
D_{\mathrm{KL}}(\pi||\pi') &= \sum_i \pi_i \log(\pi_i/\pi'_i) \\
&= \sum_i \pi_i \cdot [\eta v_i - \log Z - \eta v'_i + \log Z'] \\
&= \log Z' - \log Z - \eta \sum_i \pi_i \cdot \Delta_i \\
&= \log\left(\mathbb{E}_\pi\left[e^{\eta\Delta_i}\right]\right) - \eta\mathbb{E}_\pi\left[\Delta_i\right] \\
&= \log\left(\mathbb{E}_\pi\left[e^{\eta(\Delta_i - \mathbb{E}_\pi[\Delta_i])}\right]\right) \\
&\le \frac{\eta^2\|\Delta\|^2_{\text{span}}}{8}
\end{aligned}$$

where the last line holds by Hoeffding's lemma. □

Lemma 24 (Elliptical Potential Lemma)**.** *Fix $K \geq 1$, $m \geq 1$, $\rho > 0$ and some sequence $A_1, \ldots, A_K \in \mathbb{S}_+^m$ such that $\operatorname{Trace}(A_k) \leq L$ for all $k \in [K]$. Define for all $k \in [K]$,*

$$U_k := U_0 + \sum_{\tau=1}^{k} A_\tau, \qquad \text{for a given} \quad U_0 \succeq \rho I_m.$$

If $\rho \geq L$, then

$$\sum_{k=1}^{K} \operatorname{Trace}\left(U_{k-1}^{-1} A_k\right) \leq 2 \log \frac{\det U_K}{\det U_0} \leq 2m \log\left(1 + \frac{KL}{m\rho}\right).$$

Proof. Define $M_k := U_{k-1}^{-1/2} A_k U_{k-1}^{-1/2} \succeq 0$ such that

$$\lambda_{\max}(M_k) \leq \operatorname{Trace}(M_k) = \operatorname{Trace}\left(U_{k-1}^{-1} A_k\right) \leq \frac{\operatorname{Trace}(A_k)}{\lambda_{\min}(U_{k-1})} \leq \frac{L}{\rho} \leq 1.$$

Therefore, using the relation $\log(1 + x) \geq x/2$ for every $x \in [0, 1]$,

$$\log\left(\frac{\det U_k}{\det U_{k-1}}\right) = \log\det(I + M_k) = \sum_{i=1}^{m} \log(1 + \lambda_i(M_k)) \geq \frac{1}{2} \operatorname{Trace}(M_k).$$

Hence, by a telescopic argument,

$$\sum_{k=1}^{K} \operatorname{Trace}\left(U_{k-1}^{-1} A_k\right) = \sum_{k=1}^{K} \operatorname{Trace}(M_k) \leq 2 \log\left(\frac{\det U_K}{\det U_0}\right),$$

which gives the first inequality.

The second inequality follows from the AM-GM inequality on the eigenvalues of U_K. □

Corollary 3 (Whitened Elliptical Potential Lemma on $\mathcal{G}$)**.** *Fix $K \geq 1$, $q > 0$, $U_0 \in \mathbb{S}_{++}(\mathcal{G})$ and some sequence $A_1, \ldots, A_K \in \mathbb{S}_+(\mathcal{G})$ such that for all $k \in [K]$,*

$$\operatorname{Trace}\left(U_0^{\dagger} A_k\right) \leq q \quad \textit{and} \quad U_k := U_0 + \sum_{\tau=1}^{k} A_\tau.$$

If $q \leq 1$, then

$$\sum_{k=1}^{K} \operatorname{Trace}\left(U_{k-1}^{\dagger} A_k\right) \leq 2 \log \frac{\det_{\mathcal{G}} U_K}{\det_{\mathcal{G}} U_0} \leq 2d^{\diamond} \log\left(1 + \frac{qK}{d^{\diamond}}\right).$$

Proof. Let $Q \in \mathbb{R}^{d \times d^{\diamond}}$ have orthonormal columns spanning $\mathcal{G}$, and write the projections of U_k and A_k via Q as:

$$\bar{U}_k := Q^{\top} U_k Q \in \mathbb{S}_{++}^{d^{\diamond}}, \qquad \bar{A}_k := Q^{\top} A_k Q \in \mathbb{S}_+^{d^{\diamond}}.$$

Define the following matrices, whitened by the initial $\bar{U}_0$:

$$\tilde{U}_k := \bar{U}_0^{-1/2} \bar{U}_k \bar{U}_0^{-1/2}, \qquad \tilde{A}_k := \bar{U}_0^{-1/2} \bar{A}_k \bar{U}_0^{-1/2},$$

such that $\tilde{U}_0 = I_{d^{\diamond}}$ and $\operatorname{Trace}(\tilde{A}_k) = \operatorname{Trace}(U_0^{\dagger} A_k) \leq q$.

We can then apply Lemma 24 with $m = d^{\diamond}$, $\rho = 1$, and $L = q$ and obtain, given $q \leq \rho$:

$$\sum_{k=1}^{K} \operatorname{Trace}\left(\tilde{U}_{k-1}^{-1} \tilde{A}_k\right) \leq 2 \log \frac{\det \tilde{U}_K}{\det \tilde{U}_0} \leq 2d^{\diamond} \log\left(1 + \frac{qK}{d^{\diamond}}\right).$$

We then conclude by the following equalities:

$$\mathrm{Trace}(\tilde{U}_{k-1}^{-1}\tilde{A}_k) = \mathrm{Trace}(U_{k-1}^{\dagger}A_k), \qquad \frac{\det \tilde{U}_K}{\det \tilde{U}_0} = \frac{\det_{\mathcal{G}} U_K}{\det_{\mathcal{G}} U_0}.$$

□

Lemma 25. *Let $a, b \in \mathbb{R}_+$ such that $ae^b > e$. Then we have $n \geq a(b + \log n)$ if $n \geq 2a(b + \log a)$.*

Proof. In general for every positive integer n, some algebraic manipulations give:

$$\begin{aligned} n \geq a(b + \log n) &\Leftrightarrow n \geq ab + a\log n \\ &\Leftrightarrow \frac{n}{a} - \log n \geq b \\ &\Leftrightarrow \log n - \frac{n}{a} \leq -b \\ &\Leftrightarrow n e^{-\frac{n}{a}} \leq e^{-b} \\ &\Leftrightarrow \left(-\frac{n}{a}\right) e^{-\frac{n}{a}} \geq -\frac{1}{a} e^{-b} \end{aligned}$$

which gives the sufficient condition

$$n \geq -a \cdot W_{-1}\left(-\frac{1}{a}e^{-b}\right) = -2a' W_{-1}\left(-\frac{1}{2a'\sqrt{b'}}\right),$$

where we define $a' \coloneqq \frac{a}{2}$ and $b' \coloneqq e^{2b}$. W_{-1} represents the lower branch of the Lambert W function.

Applying the simplification of Metelli et al. (2021, Lemma B.8) we get:

$$-2a' W_{-1}\left(-\frac{1}{2a'\sqrt{b'}}\right) \leq 4a' \log(2a'\sqrt{b'}) = 2a \log(ae^b) = 2a(\log a + b),$$

which works under the assumption $2a'\sqrt{b'} = ae^b > e$.

Hence, given $2a'\sqrt{b'} = ae^b > e$, we can conclude:

$$n \geq 2a(b + \log a) \implies n \geq a(b + \log n).$$

□

H Additional experiments

We provide below additional experiments and ablations, complementing Section 5 in the main text. Unless stated otherwise, the plots report the mean and two-sided 95% confidence interval across independent seeds.

We study the following algorithms.

Active algorithms. These methods are implemented with and without an initial warm-up phase (using Algorithm 2), and are operating under the assumption of a follower best-response oracle, given a leader policy that is actively selected.

- `Active IRL` (ours): An implementation of Algorithm 1.
- `Fixed random`: An MLE-based reward extraction similar to Algorithm 1, where the leader policy is randomly sampled then kept fixed throughout all online iterations.
- `Online random`: A similar MLE-based algorithm using a new randomly sampled leader policy at each online iteration.

Fixed-expert algorithms. These methods assume access to a fixed expert, from which they can collect samples. The expert is defined with respect to a fixed (but unknown) leader reward, which is also in contrast to active algorithms.

- `Fixed-expert MLE`: Uses the same constrained MLE as the active methods, but fixing the policy as a known Stackelberg optimal leader.
- `Follower-only OLS`: Recovering the ordinary least-square solution of the empirical version of the linear system of Equation 2. This assumes access to an entropy-regularized joint-QRE expert oracle.
- `Zero-sum OLS`: Similar to `Follower-only OLS` and following the analysis of Liao et al. (2025), leverages the fact that the game is known to be zero-sum and taking the ordinary least squares solution of the stacked system formed by the linear equality constraint of Equation 2 and its leader counter-part.

For each algorithm, we term target SubOpt (respectively target Fisher) the Stackelberg sub-optimality of the policy obtained by optimizing the leader on the learned reward (respectively the norm in the Fisher information geometry described by the reward transfer lemma, Lemma 3), for the true leader reward used during training.

Similarly, the averaged transfer SubOpt and averaged transfer Fisher describe the same quantities, but averaged over ten selected leader rewards. These ten downstream tasks have been chosen from a pool of 500 randomly chosen reward candidates, by greedily maximizing diversity of payoff matrices. The pool is disjoint from the tasks used for hyperparameter selection. Averaged transfer SubOpt and averaged transfer Fisher are used as proxies for measuring the strategic-transferability of rewards.

H.1 Lobéké environment from the GUARD benchmark

We provide additional experiments using the Lobéké National Park environment from the GUARD benchmark (Krever et al., 2026), in the same setup as in the main text. We describe it below.

Task The environment models an anti-poaching scenario in the Lobéké National Park (Cameroon), designed from a dataset of recorded elephant movements. This classical green security game involves patrols (leader player) to protect wildlife against poachers (follower player). Both players distribute in the park and the followers are rewarded based on the area they occupy (as they can poach more animals) and the presence of patrols (as patrols are limiting their capacity to poach). On the other hand, the patrols can be incentivized in minimizing the rewards for the poachers, but also in trading-off this objective with the covering of prioritized areas.

Model (illustrated in Figure 4) The game is model as a 5×5 grid comprised of 23 accessible locations. A ranger action action $a \in \mathcal{A}$ is one of the 23 feasible ten-step patrol schedules generated by GUARD, while the follower response $b \in \mathcal{B}$ is the choice of one of the 23 targets. At a given time step t, the leader first chooses a distribution over patrol schedules, to which the follower reacts by choosing a location distribution, modeled as the entropy-regularized best-response to the patrol policy. Taking the role of the leader and observing sampled actions $b_t \sim \nu^*_{\mu_t}$, the learner can therefore learn information about the follower's reward and, in the active case, adapt its subsequent patrol policies.

Additionally, Zero-sum OLS also observes a sampled expert leader action from its known expert follower. Therefore, this algorithm observes twice as many samples as the other method (one additional leader action for every follower action).

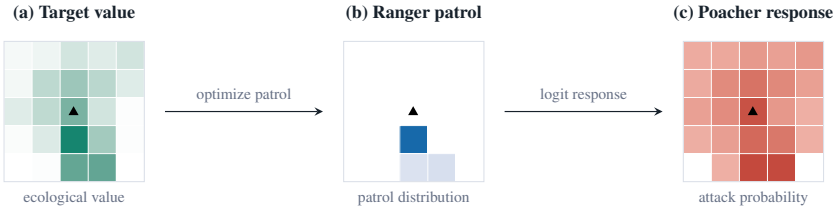

Figure 4: **Rewards, patrols, and poacher behavior in the Lobéké environment.** The 23 accessible poaching targets are embedded in a 5×5 park grid, darker cells indicate larger concentration, and the triangle marks the ranger base.

For every pair $a, b \in \mathcal{A} \times \mathcal{B}$, let us denote by v_b the poacher's value for target b and $c_{a,b} \in [0, 1]$ the fraction of target b covered by the patrol at a. From this, we define the following security-game types

of features:

$$\phi_{\text{raw}}(a,b) \coloneqq (c_{a,b},\ v_b,\ c_{a,b}v_b)\,, \qquad \phi(\cdot,\cdot) \coloneqq \phi_{\text{raw}}(\cdot,\cdot)/\max_{a,b}\|\phi_{\text{raw}}(a,b)\|_2.$$

From this, we define the true reward as:

$$r_2(a,b) \coloneqq \langle \phi(a,b), \omega^* \rangle, \qquad \omega^* \coloneqq (0,\ 1,\ -2/3),$$

where ω^* is unknown to the algorithms and extracted from the generated GUARD matrix.

We define during training the target leader objective as $r_1 \coloneqq -r_2$ to make the game zero-sum and be compatible with all the baselines. In particular, this allows for the `Zero-sum OLS` method extracted from Liao et al. (2025). In general, this is not imposed, and we will measure transfer behaviors to multiple leader rewards. Throughout, we fix $\eta \approx 2.165$.

We parametrize the transfer objectives of the patrols with a target-priority vector $z \in [-1,1]^{23}$. For each z, r_1^z trades-off between covering the same cells as the poachers and covering prioritized cells:

$$r_1^z(a,b) = (2c_{ab} - 1)z_b.$$

To ease interpretation, we will report metrics computed from the *normalized* objective $\overline{r}_1(\cdot,\cdot) = (r_1(\cdot,\cdot) - \min_{a,b} r_1(a,b))/(\max_{a,b} r_1(a,b) - \min_{a,b} r_1(a,b))$, such that $R_{\max} = 1$ and to simplify the interpretation[14].

Sampling priority vectors uniformly as random from $[-1,1]^{23}$, we generated 500 candidate transfer objectives, before flattening their induced reward matrices. From those we select 10 rewards maximizing diversity in payoff vectors. These are used to define our transfer metrics as averaged over these candidate objectives.

In the following, we provide a completion of the results present in Section 5. For the different experiments, we summarize the tuning procedure in Appendix 8.

H.1.1 Completed comparison of the algorithms: target and averaged transfer

We evaluate how the different algorithms perform for the target task and show how they transfer to multiple downstream tasks, without additional interactions in the environment.

As can be seen on Figure 5, our algorithm `Active IRL` presents a real advantage in terms of transfer to multiple downstream tasks compared to fixed-expert methods, and outperforms the other active method baselines. For the target task, the gap with the `Fixed-expert MLE` algorithm is significantly reduced, but our algorithm remains competitive without assuming access to a fixed-expert oracle.

H.1.2 Completed comparison of the algorithms: behaviors at different transfer difficulties

To deepen the understanding of transfer abilities of the different algorithms, we evaluate them on a broader spectrum of tasks

First, we draw a set of 180 priority vectors independently at random, before normalizing their induced reward matrices to $[0,1]$ using a positive affine transformation.

Second, we measure a notion of task difficulty through the sub-optimality of the source optimal leader policy:

$$D_{\text{gap}}(z, \text{src}) = \text{SubOpt}_z(\mu^\star_{\text{src}}), \qquad \mu^\star_{\text{src}} \in \arg\max_{\tilde{\mu}} J^{\text{leader}}_{\text{src}}(\tilde{\mu}).$$

Finally, we construct difficulty buckets and measure the average performance of the different algorithms for each of those buckets. Each bucket comprises exactly 60 tasks.

The results can be found in the heatmaps of Figure 6 below.

[14]Note that for constant rewards we would simply rescale, but this always induce a zero SubOpt.

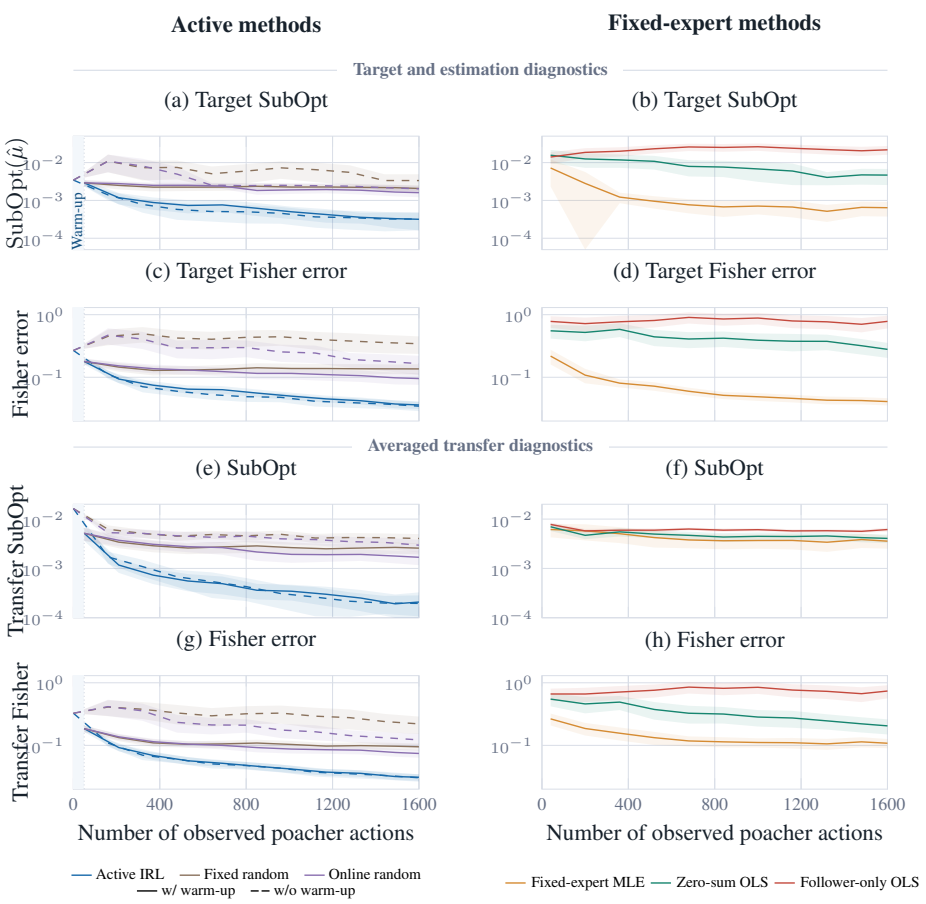

Figure 5: **Target and transfer diagnostics as a function of the number of observed poacher actions.** We evaluate the algorithms across 50 distinct seeds.

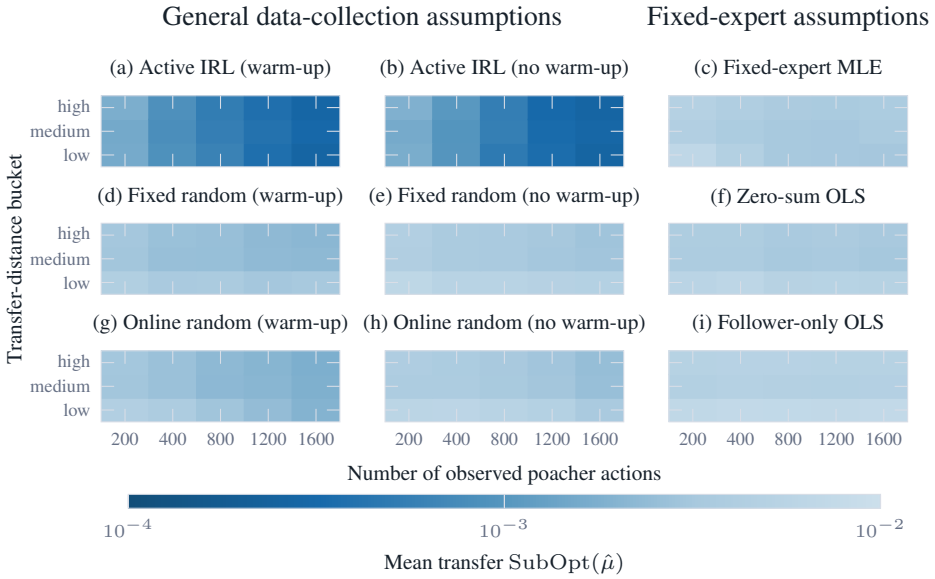

Figure 6: **Transfer SubOpt by the downstream value gap of the true source-optimal leader policy.** Cells average results for tasks in their respective bucket, each forming exactly $1/3$ of the tasks. Averaged over 50 seeds.

H.1.3 COMPLETED COMPARISON OF THE ALGORITHMS: ILLUSTRATIVE TRANSFER TASKS

We now extend the Figure 3 from the main document by reporting in Figure 7 the results for the same downstream tasks and for all the algorithms, with and without initial warm-up.

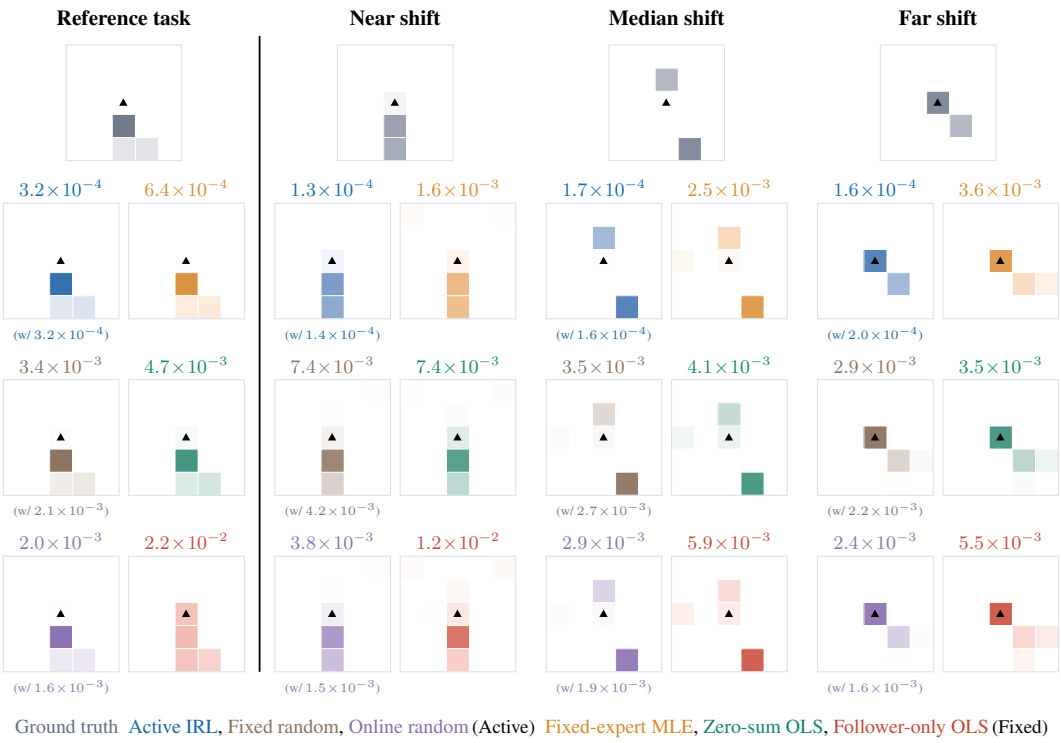

Figure 7: **Illustrative patrol policies for diverse tasks.** Algorithms with logit-response and fixed-expert assumptions. Illustrations of the learned policies for different leader rewards. Averaged over 50 seeds, reporting mean target and final SubOpt. w/ is for the performance when using warm-up.

These illustrations are mostly for intuition, we do not argue that they represent neither typical realistic leader tasks nor that they highlight the worst-case leader rewards for the algorithms.

H.2 HYPERPARAMETER SELECTION AND SENSITIVITY ANALYSIS

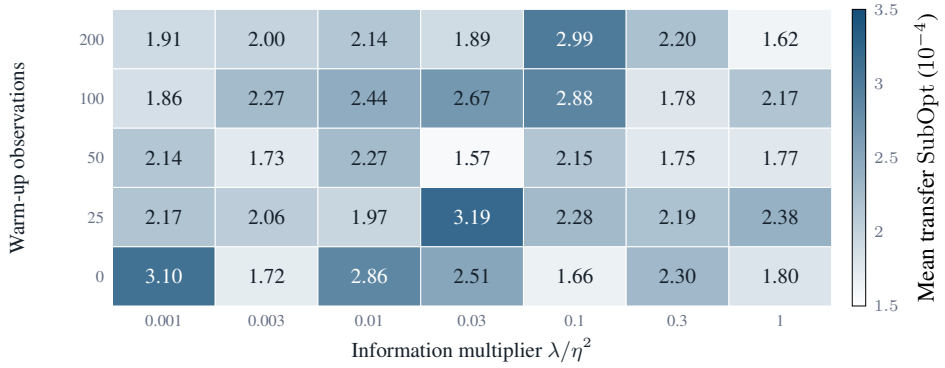

Figure 8: **Joint selection of the warm-up budget and information multiplier for `Active IRL` after 1,600 observed poacher actions.** Means computed over 100 fixed tuning tasks and 32 paired seeds. We use an MLE ridge of 10^{-6} for speed and stability (same as in the evaluation pipeline).

We present in Figure 8 a sweep over pairs of information multiplier λ/η^2 and warm-up budget, to describe the sensitivity of `Active IRL` to both parameters. We compute each transfer SubOpt mean over 100 fixed and randomly selected tasks, each task performance averaged over 32 seeds.

We use this to select hyperparameters for our experiments. Specifically, we select the minimum mean transfer SubOpt. The tasks and corresponding random seeds are only used for tuning and not shared with the evaluation pipeline.

As can be seen in the heatmap, the performances are similar across the different configurations, and sensitive to the exact seeds and tasks used. Among those configurations, we therefore choose the best-performing configuration with positive warm-up and without warm-up ($n_{\text{warm}} = 50$, $\lambda/\eta^2 = 0.03$ and $n_{\text{warm}} = 0$, $\lambda/\eta^2 = 0.1$). This allows to understand the impact of the warm-up, which is minimal in the Lobéké task.

H.3 Diagnosis in synthetic environments

H.3.1 Active selection in sparse-information games

We now highlight the benefits of active policy selection by designing a game where the Fisher information is very sparse, i.e. most of the possible leader actions are not informative for learning ω. To this end, we set up a game with m leader actions, such that only the first leader action affects the follower's reward.

Specifically, the game is defined as $\mathcal{A} = [m]$, $\mathcal{B} = [2]$, $\omega = 1$, $\eta = 2$ and $r_2^\omega(a, b) = \omega \cdot \mathbf{1}\{a = 1, b = 2\}$, where only the leader action $a = 1$ is informative. All other leader actions provide no information about ω. The leader rewards are $r_1(1, b) = \mathbf{1}\{b = 2\}$ and $r_1(a, b) = 3/4$ for $a > 1$.

For this experiment, we report results for `Active IRL` without warm-up to isolate the online policy-selection behavior of our algorithm.

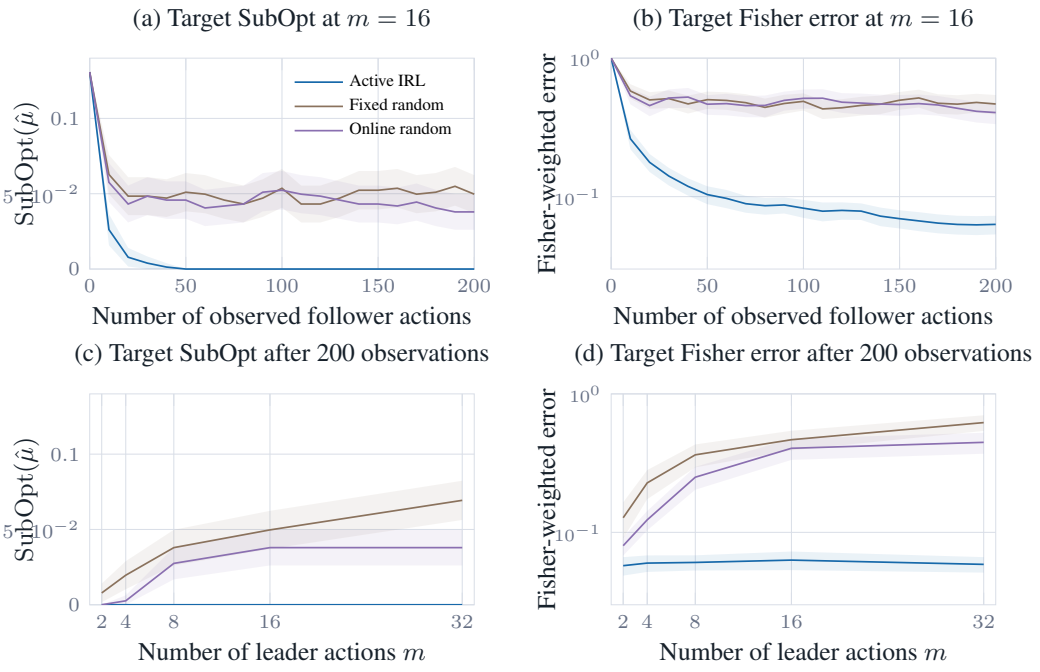

Figure 9: **Target metrics in the sparse-information game.** Averaged over 100 seeds.

As can be seen in Figure 9, our algorithm `Active IRL` is able to quickly learn ω and thus minimizing the sub-optimality of the computed policy $\hat{\mu}$. In comparison, the two baselines are not able to find useful information efficiently. As expected, the performance discrepancy is increasing with the number of leader actions m.

H.3.2 Active IRL on the lower-bound instance

We recall the hard instance as follows.

Definition 5 (Hard instance)**.** *Let* $\mathcal{A} = [d]$, $\mathcal{B} = \{b_1, b_2\}$, $R_{\max} \in \mathbb{R}_+$, *and*

$$\phi(a_i, b_1) = 0, \qquad \phi(a_i, b_2) = e_i, \qquad r_1(a, b) = R_{\max} \cdot \mathbf{1}\{b = b_2\}.$$

We identify $\mu \in \Delta_d$ and write $\ell(x) := (1 + e^{-x})^{-1}$ for the logistic function.

From this hard instance, we were able to derive the following lower bound.

Theorem 1 (Minimax lower bound)**.** *Let $d \geq 32$, $\epsilon = \frac{1}{2\eta}\sqrt{d/K} \leq S$ and $\omega^{(i)} := \varepsilon e_i$ for $i \in [d]$. There exists a game such that for every algorithm producing an output $\hat{\mu}$ after $K \geq 4d$ interactions,*

$$\max_{i \in [d]} \mathbb{E}^{(i)} \left[\mathrm{SubOpt}_{\omega^{(i)}}(\hat{\mu}) \right] \;\geq\; \frac{1}{44} R_{\max} \sqrt{\frac{d}{K}},$$

where $\mathbb{E}^{(i)}$ is for the expectation under the trajectory law $\mathbb{P}^{(i)}$ induced by $\omega^{(i)}$.

This minimax result lower bounds the worst-case SubOpt over all learning algorithms. As we cannot enumerate and run all such possible algorithms, we instead report below the performance of our algorithm `Active IRL` for the hard instance. This is more illustrative than a direct validation of the lower bound (see the full proof in Appendix C), but it does not invalidate the bound with similar rates, hence working as a sanity check.

In our experiment below, we simplify the instance by selecting $\eta = 1$, $R_{\max} = 1$ and $S = 1$. To ease the interpretation, we disable warm-up and set $\lambda = \eta^2$, therefore only keeping the online part of Algorithm 1. Finally, we directly use the available closed-form leader policy for evaluation.

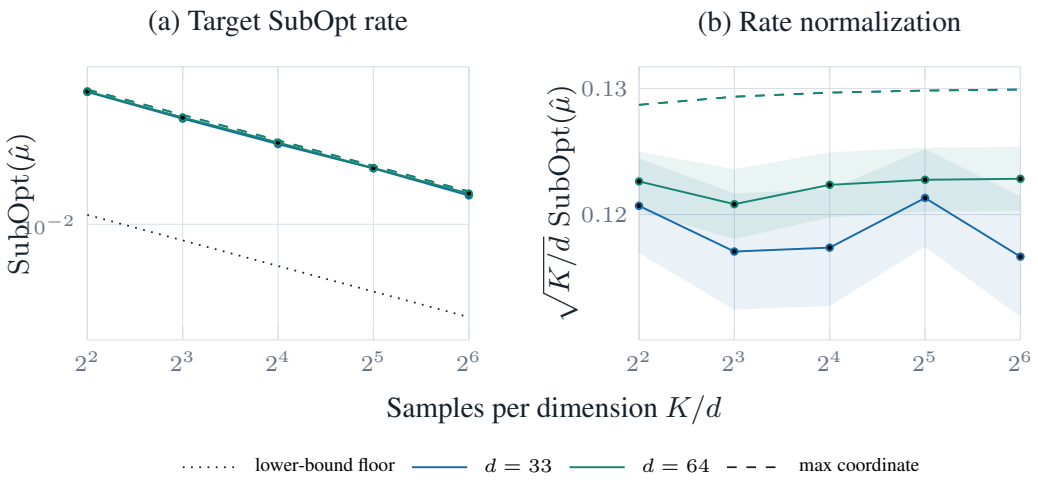

Figure 10: **Empirical scaling on the minimax hard instance.** We use 16 independent seeds per coordinate and report the maximum sub-optimality mean over all coordinates.