LEARNING FROM POPULATION MARGINALS: MINIMAX REGRET IN STACKELBERG MEAN-FIELD GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

In Stackelberg mean-field games, a leader interacting with a large population observes population states, while followers' actions are often hidden. We study reinforcement learning (RL) in such games with boundedly rational followers under state observation (SO), where the leader's population feedback consists only of the state distribution. Under SO, we establish the minimax regret $\widetilde{\Theta}(T^{(r+1)/(r+2)})$ over T episodes, with other model parameters fixed, where $r = \min\{d_F, A-1, X(B-1)\}$ is the effective dimension, d_F is the follower feature dimension, A is the number of leader actions, and X, B are the numbers of follower states and actions. We develop an optimistic RL algorithm that attains this minimax rate by evaluating candidate follower responses using observed leader rewards and transitions, and a matching lower bound establishes the optimal dependence on T. The exponent of T increases with r, quantifying how a larger effective dimension makes learning harder in the worst case. We then consider marginal observation (MO), which additionally reveals the action marginal distribution without state–action correspondence. When the SO rate is governed by the follower parameter dimension, we give a necessary and sufficient geometric condition for SO and MO to have the same minimax temporal order. We also establish matching separations for a family allowing arbitrarily large follower action spaces, showing when action marginal observation strictly improves the optimal regret rate.

1 INTRODUCTION

In leader-follower games, leaders may observe population states while followers' actions remain hidden. In electricity demand response, a utility observes consumption profiles as consumers privately choose their adjustment efforts (Élie et al., 2021); the state distribution summarizes consumption, but not the effort chosen within each state. In crowdsourcing, requesters assess completed work while workers' effort remains hidden (Ho et al., 2014). A leader thus influences hidden decisions and evaluates interventions through observable outcomes (Holmström, 1979). With a large population and an initially unknown response, repeated interactions create a learning problem.

We study this problem in Stackelberg mean-field games: in each episode, the leader commits to a policy, followers respond, and the interaction generates rewards and state transitions. Under *state observation* (SO), the leader observes the exact follower-state distribution and its own actions, rewards, and state transitions, but not follower actions. Under complete observation (CO), the joint population distribution is observed, permitting $\widetilde{O}(\sqrt{T})$ regret for fixed horizon length and model dimensions (Anonymous, 2026). Under SO, this joint distribution is a hidden input to the leader's known feature map and changes with the commitment. We ask: *What is the optimal regret for learning leader commitments from population-state observations, and how can a learner attain it?*

We answer this question for finite-dimensional linear Stackelberg mean-field games with myopic quantal responses. Let H be the horizon length, d_F, d_L the follower and leader feature dimensions, S, A the leader state and action cardinalities, and X, B the corresponding follower cardinalities. For $S, H, A, d_L \geq 2$, our expected minimax characterization over T episodes for the class in Section 3 is

$$\mathfrak{R}^*_{\mathrm{SO}}(T) = \widetilde{\Theta}\left(T^{\frac{r+1}{r+2}}\right), \qquad r = \min\{d_F, A-1, X(B-1)\}.$$

The dimensions and norm bounds are fixed, and $\widetilde{\Theta}$ suppresses logarithmic factors in T. The three terms respectively control parameter, commitment, and conditional-response uncertainty; their

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.

minimum determines how model size affects the regret exponent. Our algorithm tests candidate responses and value functions through their Bellman predictions of observed rewards and transitions, obtaining a high-probability upper bound. A same-class lower bound realizes all r difficult directions with common feature maps within each testing family and matches the exponent with constant probability.

We then ask when observing the action marginal distribution accelerates learning. Between SO and CO, *marginal observation* (MO) reveals the two population marginals separately without their state–action correspondence. Conditional response changes can cancel across follower states; the linear blindness capacity $C_{\mathrm{lin}}(X, B)$ measures how many linear logit directions this aggregation can conceal. In the parameter-limited regime $1 \le d_F \le \min\{A-1, X(B-1)\}$, with the same conditions on S, H, A, d_L, SO and MO have the same minimax temporal order exactly when $C_{\mathrm{lin}}(X, B) \ge d_F$. Otherwise, MO achieves $\widetilde{O}(T^{d_F/(d_F+1)})$. For $d_F = X \ge 2$, $X \mid B$, and $A \ge X + 1$, we match both rates, separating the SO exponent $(X + 1)/(X + 2)$ from the MO exponent $X/(X + 1)$ even with arbitrarily large follower action spaces. Numerical experiments in Appendix H illustrate the dimension effect and the contrast between informative and uninformative action marginals.

2 Related Work

The closest population-feedback comparison is Anonymous (2026): CO reveals the joint population distribution and finite-sample observation reveals sampled follower state–action pairs, both retaining state–action correspondence. For quantal Stackelberg learning in episodic Markov games, Chen et al. (2023) infer responses from observed follower actions and combine this inference with optimistic or pessimistic planning. We instead learn from exact population marginals while follower actions or their state–action correspondence remain hidden. Hidden-action incentive problems motivate this distinction: Élie et al. (2021) study private consumer effort with observed consumption, and Ho et al. (2014) learn contracts from observable task outcomes. Our candidate-response Bellman tests connect optimistic linear RL (Jin et al., 2020) with excess-loss and sequential Bellman-error control (Jin et al., 2021), but must also resolve the unknown response inside a known feature map. An extended discussion of related work is provided in Appendix A.

3 Problem Formulation

3.1 Stackelberg mean-field games

We consider an episodic Stackelberg mean-field game M with one leader and a continuum of homogeneous followers. The finite leader state/action spaces are $\mathcal{S}, \mathcal{A}$, and the finite follower state/action spaces are $\mathcal{X}, \mathcal{B}$, with cardinalities S, A, X, B. Each episode has horizon length H and steps $h \in [H] = \{1, \dots, H\}$. For a Borel space $\mathcal{E}$, $\Delta(\mathcal{E})$ denotes its probability measures, and $\mathbb{B}_q = \{v \in \mathbb{R}^q : \|v\|_2 \le 1\}$. At step h, the population state distribution $\mu_h \in \Delta(\mathcal{X})$ and conditional response $\nu_h \in \Delta(\mathcal{B})^{\mathcal{X}}$ induce $\mathcal{P}_h = \mu_h \otimes \nu_h \in \Delta(\mathcal{Z})$, where $\mathcal{Z} = \mathcal{X} \times \mathcal{B}$ and $\mathcal{P}_h(x, b) = \mu_h(x)\nu_h(b \mid x)$. The observed augmented state is $z_h = (s_h, \mu_h)$; every episode starts from the same deterministic pair (s_0, μ_0).

The leader reward $u_h(s, a, \mathcal{P})$ and follower reward $r_h(x, b; s, a)$ lie in $[0, 1]$. The augmented transition kernel $\mathbb{P}_h(\cdot \mid s, a, \mathcal{P})$ is a probability measure on $\mathcal{S} \times \Delta(\mathcal{X})$ and satisfies

$$\mathbb{P}_h(s', d\mu' \mid s, a, \mu \otimes \nu) = \mathcal{T}_{L,h}(s' \mid s, a, \mu, \nu)\, \delta_{\Phi_h(s,a,\mu,\nu)}(d\mu') \tag{1}$$

for every (s, a, μ, ν). Here the Borel kernel $\mathcal{T}_{L,h}$ governs the leader's next state and the Borel map Φ_h gives the next population distribution.

3.2 Linear structure and quantal responses

We use the finite-dimensional linear specialization of the Stackelberg mean-field model in Anonymous (2026). The feature maps are

$$\phi : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{Z}) \to \mathbb{B}_{d_L}, \qquad \varphi : \mathcal{S} \times \mathcal{A} \times \mathcal{X} \times \mathcal{B} \to \mathbb{B}_{d_F}.$$

The leader feature ϕ represents rewards and transitions; the follower feature φ represents immediate rewards and is independent of the population distribution.

Definition 1 (Linear Stackelberg mean-field game)**.** For each step h, there are unknown reward coefficients $\xi_h^* \in \mathbb{B}_{d_L}$ and $\theta_h^* \in \Theta := \{\theta \in \mathbb{R}^{d_F} : \|\theta\|_2 \le B_\Theta\}$, and a vector ω_h^* of d_L finite signed measures on $\mathcal{S} \times \Delta(\mathcal{X})$, such that $u_h(s,a,\mathcal{P}) = \langle \phi(s,a,\mathcal{P}), \xi_h^* \rangle$, $r_h(x,b;s,a) = \langle \varphi(s,a,x,b), \theta_h^* \rangle$, $\mathbb{P}_h(s', d\mu' \mid s,a,\mathcal{P}) = \langle \phi(s,a,\mathcal{P}), \omega_h^*(s', d\mu') \rangle$. These identities hold on the full domains of the reward functions and transition kernel. The true rewards lie in $[0,1]$, and the last identity defines a probability measure for every $(s,a,\mathcal{P})$ and satisfies (1).

The leader feature map obeys the following regularity condition.

Assumption 1 (Population Lipschitzness)**.** For every $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $\mathcal{P}, \mathcal{Q} \in \Delta(\mathcal{Z})$,

$$\|\phi(s,a,\mathcal{P}) - \phi(s,a,\mathcal{Q})\|_2 \le L_\phi \|\mathcal{P} - \mathcal{Q}\|_1.$$

Followers are myopic and adopt a quantal response to the announced mixed action $p \in \Delta(\mathcal{A})$. For known $\eta > 0$, set $\varphi^p(s,x,b) = \sum_a p(a)\varphi(s,a,x,b)$ and, for $\theta \in \Theta$, define

$$\nu^{p,\theta}(b \mid x; s) = \frac{\exp\{\eta \langle \varphi^p(s,x,b), \theta \rangle\}}{\sum_{c \in \mathcal{B}} \exp\{\eta \langle \varphi^p(s,x,c), \theta \rangle\}}.$$

The true step-h response uses θ_h^* and maximizes expected immediate reward under p plus η^{-1} times entropy. It is determined before the leader action is drawn.

3.3 Policy commitments and observation regimes

Let Π be the class of all tuples $\pi = (\pi_1, \dots, \pi_H)$ of Borel maps $\pi_h : \mathcal{S} \times \Delta(\mathcal{X}) \to \Delta(\mathcal{A})$. Before episode t, the learner selects and announces $\pi^t \in \Pi$ using known model information and completed episodes, and keeps the entire tuple fixed. At the observed state $z_h^t = (s_h^t, \mu_h^t)$, it evaluates $p_h^t = \pi_h^t(z_h^t)$. The induced response and action marginal are

$$\mathcal{P}_h^t(x,b) = \mu_h^t(x) \nu^{p_h^t, \theta_h^*}(b \mid x; s_h^t), \qquad \alpha_h^t(b) = \textstyle\sum_x \mathcal{P}_h^t(x,b).$$

The leader draws $a_h^t \sim p_h^t$, receives $u_h^t = u_h(s_h^t, a_h^t, \mathcal{P}_h^t)$, and observes the next augmented state drawn from $\mathbb{P}_h(\cdot \mid s_h^t, a_h^t, \mathcal{P}_h^t)$. SO and MO both reveal the exact μ_h^t before commitment evaluation. MO additionally reports the exact α_h^t from that executed commitment as step feedback, separately from μ_h^t; CO reveals the joint distribution (Table 1). Follower rewards are unobserved in all three regimes.

Regime	Population feedback	Observation structure
State observation (SO)	μ_h	The state marginal distribution is observed.
Marginal observation (MO)	(μ_h, α_h)	The action marginal is observed separately from the state marginal.
Complete observation (CO)	$\mathcal{P}_h$	The joint state–action distribution is observed.

Table 1: Population information in the three observation regimes. All provide the leader's own action, reward, and subsequent augmented state. Action-related feedback is generated by the executed commitment. SO and MO provide exact marginals; CO is the joint-observation benchmark.

The feature maps ϕ, φ, the constants η, B_Θ, L_ϕ, and the Bellman-closure bound R_U in Assumption 2 are known to the leader. The coefficients θ_h^*, ξ_h^* and the transition measures ω_h^* are unknown. For any candidate θ, the learner can compute $\mu \otimes \nu^{p,\theta}$ and its feature vector from the observed μ to test candidate responses.

3.4 Leader values and Bellman closure

For a fixed policy $\pi \in \Pi$, let

$$J_M(\pi) = \mathbb{E}_M^\pi \left[\textstyle\sum_{h=1}^H u_h(s_h, a_h, \mathcal{P}_h) \right], \qquad J_M^* = \sup_{\pi \in \Pi} J_M(\pi),$$

where the trajectory starts from (s_0, μ_0) and uses the true quantal responses. An optimal commitment and its induced response form a quantal Stackelberg equilibrium. For bounded Borel V,

$(\mathbb{P}_h V)(s, a, \mathcal{P})$ denotes its expectation under the augmented transition kernel. The integral notation and policy Bellman recursion are given in Appendix B.1.

Optimistic planning compares candidate responses and commitments; Bellman closure keeps the backups of their value envelopes in the leader's linear class with a uniform coefficient bound. For a bounded Borel function $U : \mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{Z}) \to [0, H]$ and a nonempty compact set $\Theta' \subseteq \Theta$, define

$$W_{\max}^{U,\Theta'}(s, \mu) = \sup_{\substack{p \in \Delta(\mathcal{A}) \\ \theta \in \Theta'}} \sum_{a \in \mathcal{A}} p(a)\, U\big(s, a, \mu \otimes \nu^{p,\theta}(\cdot \mid \cdot;\, s)\big)\,. \tag{2}$$

Assumption 2 (Bellman closure)**.** For every $h \in [H]$, every bounded Borel U as above, and every nonempty compact $\Theta' \subseteq \Theta$ for which the envelope in (2) is Borel, there exists $w_h^{U,\Theta'} \in \mathbb{R}^{d_L}$ with $\|w_h^{U,\Theta'}\|_2 \le R_U H$ such that

$$u_h(s, a, \mathcal{P}) + (\mathbb{P}_h W_{\max}^{U,\Theta'})(s, a, \mathcal{P}) = \langle \phi(s, a, \mathcal{P}), w_h^{U,\Theta'} \rangle$$

for every $(s, a, \mathcal{P})$.

For the minimax class, η, B_Θ, L_ϕ are fixed and strictly positive, and $R_U > 1$ is fixed.

3.5 Regret and the minimax class

The learner interacts with a fixed unknown game for T episodes. Its regret is measured against the optimal commitment in the full policy class Π:

$$\mathrm{Reg}_M(T) = \textstyle\sum_{t=1}^{T} \big[J_M^* - J_M(\pi^t)\big]. \tag{3}$$

Because the selected policies depend on observed data, (3) is a random variable. This is cumulative pseudo-regret, which we call regret throughout. Our upper bounds control this regret with a user-specified confidence level. Minimax regret takes expectation over the game trajectories and the learner's randomization.

Let σ denote the feature maps and the specified dimensions and bounds. For fixed $S, A, X, B, H, d_L, d_F, \eta, B_\Theta, L_\phi, R_U$, let $\mathfrak{C}$ be the class of pairs (σ, M) satisfying Definition 1 and Assumptions 1 and 2, with the interaction protocol above. For an observation regime $\mathsf{O} \in \{\mathrm{SO}, \mathrm{MO}\}$, define

$$\mathfrak{R}_{\mathsf{O}}^*(T) = \inf_{\mathsf{A} \in \mathcal{A}_{\mathsf{O}}} \sup_{(\sigma, M) \in \mathfrak{C}} \mathbb{E}_M^{\mathsf{A}(\sigma)}[\mathrm{Reg}_M(T)]\,,$$

where $\mathcal{A}_{\mathsf{O}}$ is the class of learning algorithms obeying the episode-start commitment and feedback protocol in Section 3.3. Each algorithm receives σ before learning. The model class and its numerical bounds are fixed independently of T. Our matching statements concern the dependence on T over this class, with constants allowed to depend on the fixed dimensions and bounds.

4 Optimistic Learning under State Observation

Under SO, the feature map is known but its true population-response input is hidden. The learner constructs features from candidate responses, tests their Bellman predictions against observed rewards and transitions, and commits to an optimistic policy before each episode.

4.1 Candidate responses and optimistic planning

Representing candidate values. Write $\mathcal{P}_\theta(z, p) = \mu \otimes \nu^{p,\theta}(\cdot \mid \cdot;\, s)$ for $z = (s, \mu)$, and set $W = \max\{1, R_U\}$. We normalize returns by H and define the candidate action-value function

$$Q_h^{\theta,w}(z, p, a) = \big\langle \phi\big(s, a, \mathcal{P}_\theta(z, p)\big), w \big\rangle\,. \tag{4}$$

Let $\mathcal{Q}_h$ be the family of candidate action-value functions indexed by (θ, w), with $\theta \in \Theta$, $\|w\|_2 \le W$, and

$$0 \le Q_h^{\theta,w}(z, p, a) \le \frac{H - h + 1}{H} \qquad \text{for every } (z, p, a).$$

A label records the parameter θ and coefficient w generating the function. The commitment p appears explicitly in (4) because it determines the follower response before the leader draws its action a. For each candidate label, both the joint distribution and its feature vector are computable from the observed augmented state and the features.

Reducing the commitment search. Fix a candidate response parameter θ and an augmented state z. At every occupied follower state, $B-1$ logit contrasts determine the conditional response. Preserving these contrasts imposes linear constraints on the commitment. Once the response is fixed, the leader's objective is linear in its mixture over actions. This observation yields the following support reduction.

Lemma 1 (Response-preserving commitments)**.** *Fix $z = (s, \mu)$, $\theta \in \Theta$, and real-valued functions c_a on the joint-distribution simplex. For every $p \in \Delta(\mathcal{A})$, there exists $\bar{p} \in \Delta(\mathcal{A})$ such that $\mathcal{P}_\theta(z, \bar{p}) = \mathcal{P}_\theta(z, p)$, $\sum_a \bar{p}(a) c_a\big(\mathcal{P}_\theta(z, \bar{p})\big) \geq \sum_a p(a) c_a\big(\mathcal{P}_\theta(z, p)\big)$, and $|\operatorname{supp} \bar{p}| \leq 1 + \min\{A-1, |\operatorname{supp} \mu|(B-1)\}$.*

Accordingly, put $k = \min\{A-1, X(B-1)\}$ and define the commitment set

$$\mathcal{P}_k = \{p \in \Delta(\mathcal{A}) : |\operatorname{supp} p| \leq k+1\}. \tag{5}$$

For a candidate tuple $Q = (Q_1, \ldots, Q_H) \in \prod_h \mathcal{Q}_h$, its greedy values are

$$V_{Q,h}(z) = \max_{p \in \mathcal{P}_k} \sum_a p(a) Q_h(z, p, a), \qquad V_{Q,H+1} = 0. \tag{6}$$

By Lemma 1, this maximum equals the maximum over the full commitment simplex. The algorithm searches the same union of simplex faces for every candidate; its definition requires only the dimensions in (5). The functions in $\mathcal{Q}_h$ remain defined and fitted on the full simplex.

Testing Bellman consistency. For a candidate action-value function q and a continuation function V, define the squared loss on completed episodes by

$$\mathcal{L}_{t,h}(q, V) = \sum_{i<t} \left[q(z_h^i, p_h^i, a_h^i) - \tfrac{u_h^i}{H} - V(z_{h+1}^i) \right]^2. \tag{7}$$

All loss terms are observed or evaluated from candidate functions. Comparing each candidate with its best-fitting alternative against the same continuation V gives the retained set

$$\mathcal{C}_t = \left\{ Q \in \textstyle\prod_{h=1}^H \mathcal{Q}_h : \mathcal{L}_{t,h}(Q_h, V_{Q,h+1}) - \inf_{q \in \mathcal{Q}_h} \mathcal{L}_{t,h}(q, V_{Q,h+1}) \leq \beta_T(\delta) \text{ for every } h \right\}. \tag{8}$$

The auxiliary fit uses the fixed class $\mathcal{Q}_h$. Bellman closure ensures that this class contains the true backup of each candidate continuation. This excess-loss construction follows the optimistic fitted-value approach of Jin et al. (2021), with the functions in (4) accounting for the hidden population response.

For a prescribed failure probability $\delta \in (0, 1)$, choose

$$\beta_T(\delta) = c_\beta \left[H(d_F + d_L) \log\big(c_\beta(T+1)\big) + \log \frac{2H(T+1)}{\delta} \right], \tag{9}$$

where $c_\beta \geq 2$ is a sufficiently large constant depending only on the fixed dimensions and bounds.

Algorithm 1 Optimistic Learning under State Observation

Input: episode count $T \geq 1$, failure probability $\delta \in (0, 1)$, and known model information σ.
Set: the confidence radius $\beta_T(\delta)$ according to (9).

1. Execute one fixed policy, record (s_0, μ_0), and store its observed trajectory in $\mathcal{D}^1$.
2. For episodes $t = 2, \ldots, T$:
   (a) Form $\mathcal{C}_t$ from $\mathcal{D}^{t-1}$ using (7)–(8).
   (b) If $\mathcal{C}_t$ is empty, execute the initialization policy, store the episode, and continue to the next episode.
   (c) Choose $Q^t \in \arg\max_{Q \in \mathcal{C}_t} V_{Q,1}(s_0, \mu_0)$.
   (d) For every h and augmented state z, select a minimum-support maximizer
   $$\pi_h^t(z) \in \arg\max_{p \in \mathcal{P}_k} \sum_a p(a) Q_h^t(z, p, a).$$
   (e) Announce the full policy π^t and execute it for the episode.
   (f) Append $(z_h^t, p_h^t, a_h^t, u_h^t, z_{h+1}^t)_{h=1}^H$ to obtain $\mathcal{D}^t$.

Output: the commitment policies $\pi^1, \ldots, \pi^T$.

Algorithm 1 implements this procedure; Appendix B.2 proves the response-preserving support reduction. We use Borel selections and minimum-support commitment maximizers; compactness and continuity in commitments and candidate labels give attainment. Initialization costs at most H regret and reveals the deterministic initial augmented state for subsequent planning. All fits and the contingent policy are fixed before the episode; its current state is used only to evaluate that policy.

4.2 Theoretical guarantees

The following result holds over the model class $\mathfrak{C}$ in Section 3.5. All constants below are uniform over its admissible feature maps and unknown games, with the dimensions and numerical bounds fixed as specified there.

Theorem 2 (High-probability SO regret upper bound)**.** *Suppose $S, H, A, d_L \geq 2$, and let*

$$r = \min\{d_F, A-1, X(B-1)\}. \tag{10}$$

There are constants $C > 0$ and $q \geq 0$, depending only on the fixed dimensions and bounds defining $\mathfrak{C}$, such that, for every $(\sigma, M) \in \mathfrak{C}$, $T \geq 1$, and $\delta \in (0,1)$, Algorithm 1 with (9) *satisfies*

$$\Pr_M\left\{\mathrm{Reg}_M(T) \leq C\, T^{\frac{r+1}{r+2}} \left[\log \frac{e(T+1)}{\delta}\right]^q\right\} \geq 1-\delta. \tag{11}$$

The confidence level is an algorithm input, and the guarantee concerns the random regret in (3). Choosing $\delta = (T+1)^{-2}$ and using $0 \leq \mathrm{Reg}_M(T) \leq HT$ also yields

$$\sup_{(\sigma,M)\in\mathfrak{C}} \mathbb{E}_M[\mathrm{Reg}_M(T)] \leq C' T^{\frac{r+1}{r+2}} [\log(e(T+1))]^q. \tag{12}$$

This supplies the expected minimax upper bound.

How observed feedback controls regret. For each candidate continuation, Bellman closure places its true backup in the comparison class. Subtracting squared losses against this backup isolates squared Bellman error and a centered fluctuation. A uniform concentration event retains the true optimal tuple and bounds the past errors of all retained candidates. Uniformity over the fixed classes permits candidates and continuations to be fitted from the same completed episodes. Optimism then relates regret to Bellman residuals: sequential complexity controls their accumulated size, and a martingale argument transfers the observed residual bound to policy-value regret.

Why the minimum of three dimensions? Each term in (10) gives a distinct way to control the uncertainty encountered by the same algorithm. The parameter bound covers the unknown follower parameter in d_F dimensions. Within a small parameter neighborhood, the remaining variation in candidate Bellman predictions is linear in the leader coefficients and can be controlled by linear regression. The commitment bound instead covers the simplex on which the leader searches, whose dimension is $A-1$. Finally, Lemma 1 reduces this search to faces of dimension at most $X(B-1)$ by preserving the conditional response. The direct dependence of the leader's objective on its own action is handled by the linear optimization within each response-preserving polytope.

These are alternative bounds on the same uncertainty process, so their minimum governs the exponent. For example, increasing the follower feature dimension stops worsening the exponent of T once the commitment or conditional-response dimension becomes smaller. Conversely, a large follower action space does not by itself force a large exponent when the response parameter dimension is small. The leader feature dimension d_L enters the bounds through linear estimation, while the dimension-dependent power of T comes from the hidden response inside the feature map.

Why the exponent $(r+1)/(r+2)$? At value resolution ϵ, the preceding geometric arguments bound sequential exploration complexity by $\widetilde{O}(\epsilon^{-r})$. Predictions accurate to ϵ contribute order $T\epsilon$ regret, while learning unresolved predictions contributes order $\sqrt{T\epsilon^{-r}}$, up to logarithms and fixed structural constants. The leading balance is

$$T\epsilon + \sqrt{T\epsilon^{-r}}, \qquad \epsilon \asymp T^{-1/(r+2)}, \qquad T\epsilon \asymp T^{(r+1)/(r+2)}. \tag{13}$$

The resolution ϵ is a device in the analysis; the algorithm uses the continuous candidate and commitment classes. The statistical term reflects learning expected continuation values from random leader-state transitions. Exact observation of the population state makes the current state available, while the transition law governing future returns still has to be learned. Appendix B formalizes the balance through uniform confidence, sequential residual counting, and a trajectory-to-value martingale transfer; Section 5 proves the same resolution unavoidable.

5 Matching Lower Bounds and Minimax Optimality

The upper bound uses the smallest of the three complexity bounds; we show that the same model class realizes all r difficult directions simultaneously. A localized transition improvement, whose location depends on the unknown follower response, must be found using the complete SO feedback.

5.1 A matching lower bound

Theorem 3 (SO regret lower bound with constant probability)**.** *Suppose $S, H, A, d_L \geq 2$ and let r be as in* (10)*. There is a constant $c > 0$, depending only on the fixed dimensions and bounds defining $\mathfrak{C}$, such that, for every SO algorithm A and $T \geq 1$, some $(\sigma, M) \in \mathfrak{C}$ satisfies*

$$\Pr_M^{\mathsf{A}(\sigma)}\left\{\mathrm{Reg}_M(T) \geq c\, T^{\frac{r+1}{r+2}}\right\} \geq \frac{1}{4}. \tag{14}$$

Consequently,

$$\inf_{\mathsf{A}\in\mathcal{A}_{\mathrm{SO}}} \sup_{(\sigma,M)\in\mathfrak{C}} \mathbb{E}_M^{\mathsf{A}(\sigma)}[\mathrm{Reg}_M(T)] \geq \frac{c}{4}\, T^{\frac{r+1}{r+2}}. \tag{15}$$

At each lower-bound scale, all testing alternatives share the same feature maps.

Why the lower bound has dimension r. For $r \geq 1$, the constraints $r \leq A-1$ and $r \leq X(B-1)$ allow r leader commitment coordinates to interact with r follower-response contrasts. The same feature maps across testing alternatives encode a localized improvement whose location depends on the unknown follower parameter. At resolution ϵ, this gives order ϵ^{-r} disjoint regions with per-episode gap of order ϵ; fixed scalings enforce every budget in $\mathfrak{C}$. A rewarding second-step leader state converts the improvement into a transition signal of size ϵ. The information bound accounts for the complete observable episode transcript, including rewards and population feedback, not just an isolated transition sample. Testing selects one fixed alternative with $\Omega(T\epsilon)$ regret at constant probability for $\epsilon \asymp T^{-1/(r+2)}$. For $r = 0$, two uncertain leader-transition models give $\Omega(\sqrt{T})$ regret with probability at least $1/4$. Appendix C gives the shared-feature construction, fixed-budget verification, and transcript testing argument.

5.2 Minimax characterization and its meaning

Combining the expectation consequence (12) of the high-probability upper bound with (15) completes the SO characterization.

Corollary 4 (Minimax regret under state observation)**.** *For $\mathfrak{C}$ with $S, H, A, d_L \geq 2$,*

$$\boxed{\mathfrak{R}^*_{\mathrm{SO}}(T) = \widetilde{\Theta}\left(T^{\frac{r+1}{r+2}}\right), \qquad r = \min\{d_F, A-1, X(B-1)\}.} \tag{16}$$

The characterization matches the dependence on T with all structural parameters fixed. The learner exploits the smallest available dimension bound, while the lower bound makes all r remaining directions affect the best commitment. The denominator $r+2$ combines r geometric directions with the ϵ^{-2} visits needed to resolve a transition-value difference of size ϵ, giving the continuum-bandit form of the exponent (Bubeck et al., 2011). Relative to the fixed-dimensional $\widetilde{O}(\sqrt{T})$ CO guarantee (Anonymous, 2026), this quantifies the worst-case temporal cost of hidden follower actions and leads naturally to the intermediate MO feedback.

6 When Do Action Marginals Improve the Regret Rate?

MO augments the SO feedback with the executed action marginal α, while the association between follower states and actions remains unobserved. For the same model class and comparator, the information ordering gives $\mathfrak{R}^*_{\mathrm{MO}}(T) \le \mathfrak{R}^*_{\mathrm{SO}}(T)$. The substantive question is when this additional marginal improves the exponent of T. Different follower states can change their conditional actions in offsetting ways, so the answer depends on how much response variation aggregation can conceal.

6.1 A criterion for improvement

To measure what aggregation can conceal, we define the largest linear family of logits with a fixed action marginal. For $1 \le m \le X$ and $u \in \mathbb{R}^{m\times B}$, let $\mathrm{softmax}_b(u_{x\cdot}) = \exp(u_{xb})/\sum_{c=1}^{B}\exp(u_{xc})$. We fix the row-wise additive freedom of logits by requiring $u_{xB} = 0$.

Definition 2 (Linear blindness capacity)**.** The capacity $C_{\mathrm{lin}}(X, B)$ is the largest dimension of a linear subspace $V \subseteq \{u \in \mathbb{R}^{m\times B} : u_{xB} = 0 \text{ for every } x\}$ for which there exist $1 \le m \le X$ and weights $\mu_1, \dots, \mu_m > 0$ with $\sum_x \mu_x = 1$ such that

$$\sum_{x=1}^{m} \mu_x \ \mathrm{softmax}_b(u_{x\cdot}) = \frac{1}{B} \qquad \text{for every } u \in V \text{ and } b \in [B].$$

The weights represent the occupied follower states. The fixed action marginal is uniform because every linear subspace contains the zero logit table. Within such a subspace, the conditional responses can vary while the action marginal distribution remains fixed. The capacity therefore measures how much linear response variation can be hidden by aggregation, and depends only on X and B. The following theorem compares this capacity with the number of unknown follower-parameter directions.

Theorem 5 (When action marginals change the minimax rate)**.** *Suppose $S, H, A, d_L \ge 2$ and*

$$1 \le d_F \le \min\{A - 1, X(B - 1)\}. \tag{17}$$

Over the same model class $\mathfrak{C}$, the following conclusions hold:

1. *If $C_{\mathrm{lin}}(X, B) \ge d_F$, then $\mathfrak{R}^*_{\mathrm{MO}}(T) = \widetilde{\Theta}\left(T^{\frac{d_F+1}{d_F+2}}\right)$.*

2. *If $C_{\mathrm{lin}}(X, B) < d_F$, then $\mathfrak{R}^*_{\mathrm{MO}}(T) \le \widetilde{O}\left(T^{\frac{d_F}{d_F+1}}\right)$.*

Consequently, SO and MO have the same minimax temporal order if and only if $C_{\mathrm{lin}}(X, B) \ge d_F$.

In (17), the SO effective dimension is $r = d_F$. When the capacity is at least d_F, every difficult parameter direction can be placed inside a response family with a constant action marginal, preserving the SO minimax temporal order. When the capacity is smaller, the MO learner can reduce the dimension controlling its exploration by at least one.

6.2 Learning from the extra marginal

The MO learner augments the candidate-response algorithm with exact action-marginal compatibility constraints. At step h, a candidate follower parameter θ must satisfy

$$\sum_x \mu^i_h(x)\nu^{p^i_h,\theta}(b \mid x; s^i_h) = \alpha^i_h(b) \qquad \text{for every } i < t,\ b \in \mathcal{B}. \tag{18}$$

For this learner, candidate follower parameters also satisfy the follower-reward constraints $0 \le \langle\varphi(s, a, x, b), \theta\rangle \le 1$ for every (s, a, x, b). These inequalities define a compact admissible subset $\Theta_{\mathrm{ad}} \subseteq \Theta$. They and (18) are imposed on the retained tuples in (8); the auxiliary regression infimum remains over the fixed class $\mathcal{Q}_h$.

Executed diagnostics and the geometric dichotomy. The action marginals observed on a fixed finite commitment grid determine the entire action-marginal map at each encountered pair (h, z). The learner executes the next grid commitment at unfinished pairs and the first at newly encountered pairs. At completed pairs, it makes non-diagnostic decisions using minimum-support optimistic commitments. Its counters, fitted tuple, and full contingent policy are frozen before each episode, and all constraints use feedback from commitments actually executed. Lemma 10 bounds the number of reachable pairs, giving a total diagnostic count independent of T for fixed dimensions and bounds. Each diagnostic decision is charged at most H in the trajectory-level Bellman-error bound before the martingale transfer to policy-value regret.

Fix a step h and let F_z be the fiber of admissible parameters whose action-marginal map agrees with the true map at z for every commitment. If its subanalytic dimension is at most $d_F - 1$, a uniform fiber cover reduces parameter-side exploration. If it has full dimension, analytic continuation forces the action marginal to be constant. On the occupied follower states, the logit contrasts are bilinear in p and θ; Lemma 15 places all coefficient tables of the products $p(a)\theta_\ell$ in one blind space of dimension at most $C_{\mathrm{lin}}(X, B)$. Response-preserving optimization then admits support at most $1 + C_{\mathrm{lin}}(X, B)$, which the learner attains by minimum-support tie breaking. These branches bound the same non-diagnostic residual sequence by the dimension $\max\{d_F - 1, C_{\mathrm{lin}}(X, B)\}$, in addition to the SO caps. When $C_{\mathrm{lin}}(X, B) < d_F$, this reduces the effective dimension to $d_F - 1$ in (17). The SO fixed-class confidence event remains valid, and fitting uses the same greedy candidate continuations. Appendix D specifies the full learner, general high-probability bound, and expectation conversion; Appendices E and F prove the geometric ingredients.

Conversely, D independent directions in a blind space give the SO testing construction a constant action marginal whenever $1 \le D \le \min\{d_F, A - 1, C_{\mathrm{lin}}(X, B)\}$. This extra feedback is identical throughout each complete testing transcript, so the lower bound remains $\Omega(T^{(D+1)/(D+2)})$. Taking $D = d_F$ when $C_{\mathrm{lin}}(X, B) \ge d_F$ proves the other direction of Theorem 5; the construction and its constant-probability guarantee are in Lemma 18.

6.3 A MATCHING SEPARATION FOR LARGE ACTION SPACES

The general bound $C_{\mathrm{lin}}(X, B) \le X - 1$ is attained by cyclically balanced action groups when X divides B. Partition the B actions into X equal groups and cyclically permute their group logits across the X equally weighted follower states. Averaging over states keeps every action's marginal probability at $1/B$, while the $X - 1$ relative group logits vary freely.

Proposition 6 (Capacity for divisible action spaces)**.** *If $X \ge 2$ and $X \mid B$, then $C_{\mathrm{lin}}(X, B) = X - 1$.*

Corollary 7 (Matching separation between SO and MO)**.** *Suppose $S, H, d_L \ge 2$, $d_F = X \ge 2$, $X \mid B$, and $A \ge X + 1$. For the same model class $\mathfrak{C}$,*

$$\mathfrak{R}^*_{\mathrm{SO}}(T) = \widetilde{\Theta}\left(T^{\frac{X+1}{X+2}}\right), \qquad \mathfrak{R}^*_{\mathrm{MO}}(T) = \widetilde{\Theta}\left(T^{\frac{X}{X+1}}\right). \tag{19}$$

The capacity identity gives effective dimensions X for SO and $X - 1$ for MO. The lower bound with $D = X - 1$ matches the MO upper bound; Appendix G.3 gives the proof. For every $B = nX$, this separates the optimal rates while state–action correspondence remains hidden, so the improvement persists with arbitrarily large follower action spaces.

7 CONCLUSION

We characterized the expected minimax temporal order for learning from exact population-state observations in linear Stackelberg mean-field games with myopic quantal responses. A high-probability optimistic upper bound and a same-class constant-probability lower bound identify the effective dimension $\min\{d_F, A-1, X(B-1)\}$. Action marginals change this order according to the capacity criterion in the parameter-limited regime, with matched separations for arbitrarily large follower action spaces. These results connect the cost of hidden actions to response geometry and the information retained by population aggregation. The empirical results provided in Appendix H illustrate the dimension effect and the contrast between informative and uninformative action marginals.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the paper, including improving language, clarity, and presentation. We did not use generative AI tools to generate the theoretical results, mathematical claims or proofs, experimental results, or code reported in this work; other required-disclosure tasks are not applicable. All AI-assisted edits were reviewed by the authors, who take responsibility for the final content of the paper.

REFERENCES

Anonymous. Reinforcement Learning for Quantal Stackelberg Equilibrium in Mean-Field Games. *Submitted to The Fifteenth International Conference on Learning Representations*, 2026. Under Review. https://openreview.net/forum?id=pAUFASvZY2.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(46):1655–1695, 2011. https://jmlr.org/papers/v12/bubeck11a.html.

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.

Raf Cluckers, Jonathan Pila, and Alex Wilkie. Uniform parameterization of subanalytic sets and Diophantine applications. *Annales scientifiques de l'École Normale Supérieure*, 53(1):1–42, 2020. https://doi.org/10.24033/asens.2416.

Romuald Élie, Emma Hubert, Thibaut Mastrolia, and Dylan Possamaï. Mean-field moral hazard for optimal energy demand response management. *Mathematical Finance*, 31(1):399–473, 2021. https://doi.org/10.1111/mafi.12291.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. A general framework for learning mean-field games. *Mathematics of Operations Research*, 48(2):656–686, 2023. https://doi.org/10.1287/moor.2022.1274.

Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, 2014. https://arxiv.org/abs/1405.2875.

Bengt Holmström. Moral hazard and observability. *The Bell Journal of Economics*, 10(1):74–91, 1979.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of the Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. In *Advances in Neural Information Processing Systems*, 34, 2021. https://arxiv.org/abs/2102.00815.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2:229–260, 2007.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

Yaolong Yu and Haipeng Chen. Decentralized online learning in general-sum Stackelberg games. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 4056–4077, 2024. https://proceedings.mlr.press/v244/yu24b.html.

A Extended Related Work

Mean-field games and Stackelberg learning. Mean-field games describe strategic interaction in large populations through the distribution of individual states (Lasry and Lions, 2007). Reinforcement-learning approaches have developed algorithms for computing mean-field equilibria from interaction data (Guo et al., 2023). Stackelberg learning adds a leader whose policy shapes the followers' responses. For quantal responses (McKelvey and Palfrey, 1995), Chen et al. (2023) study learning a quantal Stackelberg equilibrium in episodic Markov games with function approximation. Their algorithms infer the response model from observed follower actions and combine this inference with optimistic or pessimistic planning. Our problem concerns a population response whose state–action distribution is hidden from the leader. The resulting minimax rate depends on the geometry of that hidden response.

Complete and finite-sample population observation. Recent work studies linear Stackelberg mean-field games under complete observation and finite-sample observation (Anonymous, 2026). CO supplies the joint population distribution, whereas finite-sample observation supplies sampled follower state–action pairs. Both retain the correspondence between follower states and actions. We study exact population marginals under episode-start commitment and quantal responses: SO supplies the state marginal, and MO additionally supplies the action marginal separately. Our matching SO characterization quantifies the learning cost of hidden follower actions, and the MO criterion determines when action marginal observations improve the minimax temporal order in the parameter-limited regime. This comparison distinguishes information about how often actions are taken from information about which states those actions are taken in.

Hidden actions and incentive learning. Hidden-action models are central to principal–agent theory (Holmström, 1979). Élie et al. (2021) study a mean-field moral-hazard model of energy demand response in which consumers' effort is private and consumption is observed. For online contract design, Ho et al. (2014) develop bandit algorithms that learn from observable task outcomes generated by workers' hidden actions. In repeated finite-action Stackelberg games, Yu and Chen (2024) study decentralized learning with leader feedback consisting of its own reward signal. These works motivate learning through the consequences of strategic behavior. Our analysis quantifies the regret of policy commitments in an episodic mean-field model and compares exact population-state feedback with separately aggregated state and action feedback.

Linear reinforcement learning and geometric complexity. Optimistic methods for linear Markov decision processes use known feature representations to control Bellman estimation errors (Jin et al., 2020). In our setting, the feature map is known, while its joint-population input depends on an unknown follower response. The algorithm therefore constructs candidate features from possible responses and tests their Bellman predictions using observed leader rewards and transitions. Its excess-loss confidence sets build on optimistic fitted-value methods and sequential Bellman-error control (Jin et al., 2021). Dimension-dependent rates also arise in continuum-armed bandits (Bubeck et al., 2011). For the present model, we derive three alternative complexity bounds from follower parameters, leader commitments, and conditional responses, and obtain an SO regret bound governed by their minimum. The matching lower bound realizes all of these remaining difficult directions within the same Stackelberg mean-field model class. The MO analysis then identifies how consistency with the observed action marginals changes these complexity bounds.

B Proof of the High-Probability SO Upper Bound

Throughout this appendix, the game and its known model information are fixed, and all constants are uniform over $\mathfrak{C}$ with its dimensions and bounds fixed. We write $\chi_h^t = (z_h^t, p_h^t, a_h^t)$ for a query and $\beta = \beta_T(\delta)$. The following arguments establish Theorem 2 for the observable algorithm in Algorithm 1.

B.1 POLICY-VALUE NOTATION

For a fixed policy π, write $\nu_h^\pi(\cdot \mid x; s, \mu) = \nu^{\pi_h(s,\mu),\theta_h^*}(\cdot \mid x; s)$. The leader's expected return is

$$J_M(\pi) = \mathbb{E}_M^\pi\left[\sum_{h=1}^{H} u_h(s_h, a_h, \mu_h \otimes \nu_h^\pi(\cdot \mid \cdot;\, s_h, \mu_h))\right],$$

where the expectation starts from (s_0, μ_0). For a bounded Borel function W on the augmented state space, define

$$(\mathbb{P}_h W)(s, a, \mathcal{P}) = \sum_{s' \in \mathcal{S}} \int W(s', \mu')\, \mathbb{P}_h(s', d\mu' \mid s, a, \mathcal{P}).$$

The policy value functions satisfy $W_{H+1}^\pi = 0$ and

$$U_h^\pi(s, a, \mathcal{P}) = u_h(s, a, \mathcal{P}) + (\mathbb{P}_h W_{h+1}^\pi)(s, a, \mathcal{P}),$$
$$W_h^\pi(s, \mu) = \sum_{a \in \mathcal{A}} \pi_h(a \mid s, \mu)\, U_h^\pi(s, a, \mu \otimes \nu_h^\pi(\cdot \mid \cdot;\, s, \mu))\,.$$

The function U_h^π evaluates the return conditional on the current leader action and joint population distribution, and W_h^π averages this quantity over the action drawn from the policy. Here $\pi_h(a \mid s, \mu)$ denotes coordinate a of the distribution $\pi_h(s, \mu)$. The optimal leader value is $J_M^* = \sup_{\pi \in \Pi} J_M(\pi)$. An optimal commitment together with its induced quantal follower response forms a quantal Stackelberg equilibrium.

B.2 PROOF OF RESPONSE-PRESERVING SUPPORT REDUCTION

The support reduction used by the main-text learner preserves each candidate response while optimizing the leader's action mixture.

Proof of Lemma 1. Choose a baseline follower action and let $L_\theta(z)$ map a commitment to the logit contrasts in the occupied rows. The polytope

$$\{v \geq 0 : \mathbf{1}^\top v = 1,\ L_\theta(z)v = L_\theta(z)p\}$$

preserves the entire candidate response. Maximizing the resulting linear objective over this polytope has an extreme-point solution with at most $\operatorname{rank}[\mathbf{1}^\top; L_\theta(z)]$ positive coordinates. This rank is at most $\min\{A, 1 + |\operatorname{supp}\mu|(B-1)\}$. □

B.3 THREE DIMENSION CONTROLS

Table 2 summarizes the complementary controls used by the same learner.

Dimension	Statistical or decision freedom	Mechanism controlling regret
d_F	Unknown follower-response parameters	Cover candidate parameters and use linear prediction within each neighborhood.
$A-1$	Mixed leader commitments	Cover the commitment simplex.
$X(B-1)$	Conditional follower-response contrasts	Preserve responses while optimizing over smaller commitment supports.

Table 2: Three bounds on the same sequential learning difficulty. Their minimum is the effective dimension r in Theorem 2.

B.4 HOW BELLMAN CONSISTENCY CONTROLS EXPLORATION

The analysis links the observable loss in (7) to prediction errors and then uses the three dimensions to control how often those errors can be large. For analysis, define the normalized true backup

$$(\mathcal{B}_h V)(z, p, a) = \frac{u_h(s, a, \mathcal{P}_{\theta_h^*}(z, p))}{H} + (\mathbb{P}_h V)(s, a, \mathcal{P}_{\theta_h^*}(z, p)),$$

and let $g_{t,h} = Q_h^t - \mathcal{B}_h V_{Q^t, h+1}$.

Uniform confidence and optimism. A simultaneous concentration argument over the fixed candidate classes yields an event of probability at least $1 - \delta/2$ on which the true optimal value tuple belongs to every $\mathcal{C}_t$ and

$$\sum_{i<t} g_{t,h}(z_h^i, p_h^i, a_h^i)^2 \le c\,\beta_T(\delta) \qquad \text{for every } t, h. \tag{20}$$

To see the role of excess loss, fix a continuation V and compare a candidate prediction with its true backup. Conditioning on the observed history and current leader action, the difference of their squared losses equals the squared Bellman error plus a centered term linear in that error. A uniform bound on these centered terms gives (20). Taking the event over the fixed classes before selecting Q^t permits the continuation and predictor to be fitted from the same completed episodes. Retention of the true optimal tuple then supplies optimism at the initial state.

Controlling new prediction errors. Bellman closure places the residuals inside a class of functions of the form

$$g(z, p, a) = \langle \phi(s, a, \mathcal{P}_\theta(z, p)), w\rangle - \left\langle \phi(s, a, \mathcal{P}_{\theta_h^*}(z, p)), v\right\rangle, \qquad \|w\|_2, \|v\|_2 \le W. \tag{21}$$

For a fixed game, let $e(\epsilon)$ be the maximum over steps of the sequential dimension of this class on reachable augmented states and commitments in $\mathcal{P}_k$. Specifically, it is the longest sequence of queries $\chi_j = (z_j, p_j, a_j)$ for which, at a common scale $\epsilon' \ge \epsilon$, each query has a witness g_j satisfying

$$\sum_{i<j} g_j(\chi_i)^2 \le (\epsilon')^2, \qquad |g_j(\chi_j)| > \epsilon'.$$

It counts how long a prediction can remain accurate on previous queries and substantially wrong on a new one. We establish the uniform bound

$$e(\epsilon) \le C\epsilon^{-r}\big[\log(C/\epsilon)\big]^q, \qquad 0 < \epsilon \le 1. \tag{22}$$

The parameter argument partitions Θ into small neighborhoods. Centering the nonlinear feature map within each neighborhood turns (21) into a linear prediction problem with $2d_L$ coefficients, up to a controlled approximation error. An elliptical-potential argument bounds the number of independent witnesses in one neighborhood logarithmically in the resolution, leading to the cap $\widetilde{O}(\epsilon^{-d_F})$. For the commitment argument, the linear augmented transition and the causal factorization give at most Sd_L reachable augmented states per step. Covering the finitely many faces in $\mathcal{P}_k$ at each such state gives the alternative cap $O(\epsilon^{-k})$. Lipschitz continuity prevents too many independent witnesses within one commitment cell. Taking the smaller cap proves (22).

From prediction error to regret. The sequential counting argument for (20), of the type used in Jin et al. (2021), bounds the residuals at the observed queries. Optimism and a Bellman telescoping argument bound the regret in an episode by the conditional expected sum of those residuals. Because the fitted tuple and the policy are fixed before the episode, the difference between this conditional expectation and the observed sum is a bounded martingale difference. Concentration of this martingale, together with the uniform confidence event, gives, with probability at least $1 - \delta$,

$$\mathrm{Reg}_M(T) \le CH^2\left[1 + T\epsilon + e(\epsilon) + \sqrt{\beta_T(\delta)Te(\epsilon)} + \sqrt{T\log\frac{2}{\delta}}\right]. \tag{23}$$

The final term is the trajectory-to-value concentration cost. The initial 1 includes the first episode. Substituting (22) into (23) and choosing the resolution in (13) proves Theorem 2. At this resolution, $e(\epsilon)$ has order $\widetilde{O}(T^{r/(r+2)})$ and is dominated by the leading regret term. The additional square-root term also preserves the exponent, including the endpoint $r = 0$. The remaining subsections provide the confidence, sequential counting, and martingale arguments in detail.

B.5 Backups and uniform confidence

Coefficient control. Every Borel continuation $v : \mathcal{S} \times \Delta(\mathcal{X}) \to [0, 1]$ has a normalized backup $u_h/H + \mathbb{P}_h v$ represented by a coefficient of norm at most W. Indeed, choose $U(s, a, \mathcal{P}) =$

$Hv(s, \operatorname{marg}_{\mathcal{X}} \mathcal{P})$ in Assumption 2. The envelope in (2) is exactly $Hv(s, \mu)$, and dividing the resulting backup coefficient by H gives the claim. For a continuation of the range appropriate to step $h+1$, this backup takes values in $[0, (H-h+1)/H]$. Evaluating it at the true response therefore gives an element of the full auxiliary class $\mathcal{Q}_h$. In particular, backward induction represents the normalized optimal value tuple in $\prod_h \mathcal{Q}_h$. Lemma 1 preserves its maximization over the full policy class.

Lemma 8 (Simultaneous confidence)**.** *With the radius in* (9)*, there is an event G of probability at least $1-\delta/2$ such that the true optimal tuple belongs to $\mathcal{C}_t$ for every t and*

$$\sum_{i<t} \left[Q_h(\chi_h^i) - (\mathcal{B}_h V_{Q,h+1})(\chi_h^i)\right]^2 \le C\beta \qquad \text{for every } Q \in \mathcal{C}_t,\ t \le T,\ h \le H. \tag{24}$$

The same event can be used when retained candidate labels also satisfy the exact MO constraints.

Proof of Lemma 8. First fix a candidate prediction $q \in \mathcal{Q}_h$ and a candidate continuation V. Condition on the observed history through the action at step h of episode i. For this fixed pair, set

$$Y_i = u_h^i/H + V(z_{h+1}^i), \qquad q_i^0 = (\mathcal{B}_h V)(\chi_h^i), \qquad \zeta_i = Y_i - q_i^0, \qquad g_i = q(\chi_h^i) - q_i^0.$$

The variable Y_i lies in $[0,1]$, and ζ_i has conditional mean zero and conditional range length at most one. The excess loss against the true backup is

$$\mathcal{L}_{t,h}(q, V) - \mathcal{L}_{t,h}(\mathcal{B}_h V, V) = \sum_{i<t} g_i^2 - 2\sum_{i<t} g_i \zeta_i. \tag{25}$$

Conditional Hoeffding bounds imply

$$\mathbb{E}[\exp(2\lambda g_i \zeta_i) \mid \text{history through } a_h^i] \le \exp(\lambda^2 g_i^2/2).$$

Multiplying these inequalities over episodes and applying Markov's inequality with $\lambda = \pm 1/2$ gives, for a fixed prefix, a failure probability at most $2e^{-\ell}$ for

$$\left| 2\sum_{i<t} g_i \zeta_i \right| \le \frac{1}{4}\sum_{i<t} g_i^2 + 2\ell. \tag{26}$$

Let $\mathcal{V}_{h+1}$ be the fixed class of continuations induced by candidate tuples, and set $\rho_T = (T+1)^{-2}$. Choose a proper finite cover $\mathcal{N}_h$ of $\mathcal{Q}_h \times \mathcal{V}_{h+1}$ such that every pair (q, V) has a center $(\bar{q}, \bar{V})$ satisfying

$$\|q - \bar{q}\|_\infty \le \rho_T, \qquad \|V - \bar{V}\|_\infty \le \rho_T.$$

These covers follow from Euclidean covers of the bounded candidate labels: the softmax response and, by Assumption 1, the candidate predictions are uniformly Lipschitz in their labels, while maximization over commitments is a supremum-norm contraction. Proper centers preserve the candidate range constraints, and $N_T = \max_h |\mathcal{N}_h|$ satisfies

$$\log N_T \le CH(d_F + d_L)\log(C(T+1)).$$

For a pair and its center, define $\bar{g}_i$ and $\bar{\zeta}_i$ by replacing (q, V) with $(\bar{q}, \bar{V})$ in the definitions above. Conditional expectation is also a supremum-norm contraction, so $|g_i - \bar{g}_i| \le 2\rho_T$ and $|\zeta_i - \bar{\zeta}_i| \le 2\rho_T$. Since all four variables have absolute value at most one, for $n = t-1$ we obtain

$$\left| \sum_{i<t} g_i^2 - \sum_{i<t} \bar{g}_i^2 \right| \le 4n\rho_T,$$

$$\left| 2\sum_{i<t} g_i \zeta_i - 2\sum_{i<t} \bar{g}_i \bar{\zeta}_i \right| \le 8n\rho_T.$$

Apply (26) to the centers with $\ell = \log(4H(T+1)N_T/\delta)$. A union bound over centers, steps, and prefixes has failure probability at most $2H(T+1)N_T e^{-\ell} = \delta/2$. Transferring from a center to an arbitrary pair adds at most $9n\rho_T$ to the remainder in that inequality. Choosing the fixed constant c_β

in the confidence radius sufficiently large gives $2\ell + 9n\rho_T \le \beta/4$ simultaneously for all prefixes. Together with (25), this yields

$$\mathcal{L}_{t,h}(q, V) - \mathcal{L}_{t,h}(\mathcal{B}_h V, V) \ge \frac{3}{4} \sum_{i<t} \big[q(\chi_h^i) - (\mathcal{B}_h V)(\chi_h^i)\big]^2 - \frac{\beta}{4} \tag{27}$$

simultaneously for every candidate pair, step, and prefix.

For the true optimal tuple, its prediction is its own true backup, which belongs to $\mathcal{Q}_h$ by coefficient control. Applying (27) to every $q \in \mathcal{Q}_h$, with V equal to this tuple's continuation, gives

$$0 \le \mathcal{L}_{t,h}(\mathcal{B}_h V, V) - \inf_{q \in \mathcal{Q}_h} \mathcal{L}_{t,h}(q, V) \le \frac{\beta}{4} \le \beta,$$

which proves retention.

For any retained tuple Q, now take $V = V_{Q,h+1}$. Coefficient control likewise puts its true backup in the full auxiliary class, so

$$\mathcal{L}_{t,h}(Q_h, V) - \mathcal{L}_{t,h}(\mathcal{B}_h V, V) \le \mathcal{L}_{t,h}(Q_h, V) - \inf_{q \in \mathcal{Q}_h} \mathcal{L}_{t,h}(q, V) \le \beta.$$

Combining this inequality with (27) bounds the squared-error sum by $5\beta/3$, proving (24). The event was constructed on fixed global classes before any data-dependent fitting. Intersecting retained labels with exact MO equations preserves the true tuple and the same squared-error bound, while leaving the auxiliary regression class unchanged. □

B.6 Sequential dimension and the realized residual sum

We give the two bounds underlying (22). For this purpose, the enlarged class in (21) has uniformly bounded coefficients and uniform Lipschitz constants in both its candidate parameter and its commitment.

The parameter bound. Write an enlarged residual as a linear prediction in the $m = 2d_L$-dimensional feature vector

$$x_\theta(z, p, a) = \big(\phi(s, a, \mathcal{P}_\theta(z, p)), \phi(s, a, \mathcal{P}_{\theta_h^*}(z, p))\big)$$

with coefficient $(w, -v)$ of norm at most $R_b = \sqrt{2}W$. Fix a common independence scale $\epsilon' \ge \epsilon$ and put $L_\epsilon = \lceil Cm \log(C/\epsilon) \rceil$. Partition the candidate parameter ball into proper cells so that replacing a parameter by its cell representative changes each residual prediction by at most $\epsilon/(16\sqrt{L_\epsilon})$. The number of cells is at most $C\epsilon^{-d_F}[\log(C/\epsilon)]^{d_F/2}$.

Suppose a cell contained $L_\epsilon + 1$ witnesses to an independent sequence, and retain the first such witnesses in their original order. At that cell's fixed representative, the previous predictions for each witness have Euclidean norm at most $(17/16)\epsilon'$, while the absolute current prediction exceeds $(15/16)\epsilon'$. Let x_i be the feature vectors at this cell representative and set

$$V_{i-1} = \lambda I + \sum_{j<i} x_j x_j^\top, \qquad \lambda = (17/16)^2(\epsilon')^2/R_b^2.$$

The prediction inequalities and Cauchy–Schwarz in the V_{i-1} norm imply $x_i^\top V_{i-1}^{-1} x_i \ge c > 0$. The determinant identity and the bound $\|x_i\|_2 \le \sqrt{2}$ therefore give, for the number M of such witnesses,

$$M \log(1 + c) \le \log \frac{\det V_M}{\det(\lambda I)} \le m \log\left(1 + \frac{CM}{m\epsilon^2}\right).$$

For the constant in L_ϵ sufficiently large, this rules out $M = L_\epsilon + 1$. Multiplying the number of cells by L_ϵ gives the bound $\widetilde{O}(\epsilon^{-d_F})$.

The commitment bound. At a fixed step, the population marginals of all augmented transition kernels lie in the span of the d_L population-marginal coordinate measures of ω_h^*. By the causal factorization in (1), each such marginal is a Dirac measure. Distinct Dirac measures are linearly

independent, so at most d_L population state distributions can occur at the next step. Including the deterministic initial state gives at most $Z = \max\{1, Sd_L\}$ reachable augmented states per step. This counting bound is used only in the analysis.

For each reachable state, actual leader action, and face in $\mathcal{P}_k$, partition the face into cells of ℓ_1 diameter at most $\epsilon/(2L_g)$, where $L_g \geq 1$ is a uniform commitment Lipschitz bound for the enlarged residuals. There are at most $C\epsilon^{-k}$ cells in total, since the number of faces, states, and leader actions is fixed. If five queries in one cell belonged to an independent sequence at scale $\epsilon' \geq \epsilon$, a witness for the fifth would have magnitude greater than $\epsilon'/2$ at all four preceding queries in that cell. Their squared predictions would sum to more than $(\epsilon')^2$, a contradiction. The resulting dimension bound is $O(\epsilon^{-k})$. Taking the smaller bound proves (22) with $r = \min\{d_F, k\}$.

From squared past error to cumulative current error. Consider a sequence of residuals g_t with $|g_t(\chi_t)| \leq 1$ and $\sum_{i<t} g_t(\chi_i)^2 \leq b$. For a threshold u, greedily assign the queries with $|g_t(\chi_t)| > u$ to $\lfloor b/u^2 \rfloor + 1$ subsequences. At least one subsequence has preceding squared predictions at most u^2, so the current query extends an independent sequence there at scale u. Each subsequence has at most $e(u)$ queries. The number of errors exceeding u is therefore at most $(1 + b/u^2)e(u)$. Since $e(u) \leq E := e(\epsilon)$ for $u \geq \epsilon$, define

$$N(u) = \#\{t \leq T : |g_t(\chi_t)| > u\}.$$

The preceding counting bound and $N(u) \leq T$ give

$$N(u) \leq \min\{T, E(1 + b/u^2)\} \leq \min\{T, E\} + \min\{T, bE/u^2\}, \qquad u \geq \epsilon.$$

Because $|g_t(\chi_t)| \leq 1$, integrating the tail counts yields

$$\begin{aligned}\sum_{t=1}^{T} |g_t(\chi_t)| &= \int_0^1 N(u)\, du \\ &\leq T\epsilon + \min\{T, E\} + \int_0^\infty \min\{T, bE/u^2\}\, du.\end{aligned}$$

If $bE > 0$, splitting the last integral at $u_0 = \sqrt{bE/T}$ gives $Tu_0 + bE/u_0 = 2\sqrt{bTE}$; if $bE = 0$, the integral is zero. Therefore

$$\sum_{t=1}^{T} |g_t(\chi_t)| \leq T\epsilon + \min\{T, e(\epsilon)\} + 2\sqrt{bTe(\epsilon)}. \tag{28}$$

On G, apply (28) at each step with $b = C\beta$. The resulting bound is simultaneous for the fitted tuples actually selected by the algorithm.

B.7 A MARTINGALE TRANSFER AND COMPLETION OF THE PROOF

Let $\mathcal{F}_{t-1}$ contain the completed episodes and the learner's episode-start choices for episode t. The learner's choices in Algorithm 1 use fixed Borel selections, so they are determined by the completed data. For $t \geq 2$ put

$$W_t = \sum_{h=1}^{H} |g_{t,h}(\chi_h^t)|, \qquad m_t = \mathbb{E}[W_t \mid \mathcal{F}_{t-1}].$$

When the retained set is empty, use the zero candidate tuple to define the residuals for this analysis and execute the fallback policy specified by the algorithm. The residuals remain bounded by one, so $0 \leq W_t, m_t \leq H$ on every outcome. Consequently $m_t - W_t$ is a bounded martingale difference under the original trajectory law. Azuma–Hoeffding concentration yields an event G' of probability at least $1 - \delta/2$ on which

$$\sum_{t=2}^{T} (m_t - W_t) \leq H\sqrt{2(T-1)\log(2/\delta)}. \tag{29}$$

On G, optimism and the Bellman telescoping identity for the greedy policy give

$$\frac{J_M^* - J_M(\pi^t)}{H} \leq V_{Q^t,1}(z_1) - \frac{J_M(\pi^t)}{H} = \mathbb{E}\left[\sum_{h=1}^{H} g_{t,h}(\chi_h^t) \,\middle|\, \mathcal{F}_{t-1}\right] \leq m_t.$$

Here $z_1 = (s_0, \mu_0)$ denotes the initial augmented state. The equality uses the tuple and policy frozen at the beginning of the episode, and averages the realized leader action according to the announced commitment. The concentration statement (29) was established under the original law before intersecting with G. On $G \cap G'$, whose probability is at least $1 - \delta$, initialization and the preceding bounds imply

$$\begin{aligned}\mathrm{Reg}_M(T) &\le H + H\sum_{t=2}^{T} W_t + H^2\sqrt{2(T-1)\log(2/\delta)} \\ &\le CH^2\left[1 + T\epsilon + e(\epsilon) + \sqrt{\beta T e(\epsilon)} + \sqrt{T\log(2/\delta)}\right].\end{aligned}$$

This is (23). For $T \ge 2$, substituting (22) with $\epsilon = T^{-1/(r+2)}$ proves (11) after enlarging the fixed constants and logarithmic exponent. For $T = 1$, the bound follows from $\mathrm{Reg}_M(1) \le H$. Finally, setting $\delta = (T+1)^{-2}$ and bounding the regret on failure by HT proves (12).

C Proof of the SO Lower Bound with Constant Probability

We first implement the local return differences in the original game class and then prove the probability statement in (14). All numerical scalings below are fixed independently of T and of the localization scale.

C.1 Construction overview and the minimax scale

Encoding difficult commitments. For $r \ge 1$, choose r nonbaseline leader actions and r distinct follower state–action contrasts. These choices are possible because $r \le A - 1$ and $r \le X(B-1)$. Assign follower features so that, on the resulting family of population responses, a fixed Lipschitz transformation of the joint state–action distribution recovers the coordinates

$$\ell_i\big(\mathcal{P}_\theta(z,p)\big) = \theta_i p_i, \qquad i = 1, \dots, r.$$

Here the hidden coordinates θ_i range over a fixed positive interval inside Θ. The transformation uses log ratios of conditional response probabilities, extended to the full joint-distribution simplex by clipping. For a fixed $c_* > 0$ and a small localization scale ϵ, the leader feature contains the function

$$b_\epsilon(\mathcal{P}) = \left[\epsilon - \max_{1 \le i \le r} |\ell_i(\mathcal{P}) - c_*|\right]_+, \tag{30}$$

where $[v]_+ = \max\{v, 0\}$. Its Lipschitz bound is independent of ϵ. The maximizing commitments have $p_i = c_*/\theta_i$, so varying the hidden parameter moves the optimum through an r-dimensional region. A packing gives $K_\epsilon \asymp \epsilon^{-r}$ alternatives with disjoint regions of improved value. The feature map specifies the shape and scale in (30); the unknown parameter determines where that shape is encountered as the leader varies its commitment.

Realizing the signal in the game. Two leader states and two leader feature coordinates suffice. At the first step, the probability of reaching a rewarding state is a fixed interior probability plus a fixed positive multiple of $b_\epsilon(\mathcal{P})$. The second-step reward is a deterministic function of the reached state, and the population is fixed across the testing family. All additional steps and coordinates can be padded without changing the learning problem. The construction uses fixed scalings that satisfy the reward bounds, feature norms, population Lipschitzness, transition factorization, and Bellman-closure budget of $\mathfrak{C}$ for every ϵ.

Let $\mathbb{Q}_0$ be the law of the complete observable transcript under a reference model with no local improvement, and let $\mathbb{Q}_j$ be its law under alternative j. If N_j counts commitments in alternative j's improved region, the conditional transition probabilities give

$$\mathrm{KL}(\mathbb{Q}_0 \,\|\, \mathbb{Q}_j) \le C\epsilon^2 \mathbb{E}_0[N_j]. \tag{31}$$

Pinsker's inequality and the disjointness of the improved regions imply

$$\frac{1}{K_\epsilon}\sum_{j=1}^{K_\epsilon} \frac{\mathbb{E}_j N_j}{T} \le \frac{1}{K_\epsilon} + C\epsilon\sqrt{\frac{T}{K_\epsilon}}.$$

Choose $\epsilon = c_{\mathrm{loc}} T^{-1/(r+2)}$ with a sufficiently small fixed $c_{\mathrm{loc}} > 0$ so that the right-hand side is at most $1/4$. For some alternative j, $\mathbb{E}_j N_j \le T/4$, and Markov's inequality gives $\Pr_j\{N_j \le T/2\} \ge 1/2$. Outside its improved region, that game's per-episode gap is a fixed positive multiple of ϵ. It therefore incurs $\Omega(T\epsilon)$ regret with probability at least $1/2$. For $r = 0$, a two-alternative construction with uncertain leader transitions yields $\Omega(\sqrt{T})$ regret with probability at least $1/4$. All alternatives at a given scale have common feature maps and satisfy the same fixed budgets. The class contains the features for every scale, so this construction has the minimax quantifiers specified in Section 3.5. The following subsections supply the constructions and both probability arguments.

The upper and lower arguments describe the same exploration problem from two directions. At resolution ϵ, up to order ϵ^{-r} different regions can require exploration, while distinguishing a local return difference of size ϵ through random transitions costs order ϵ^{-2} informative visits. The critical scale therefore satisfies

$$T \asymp \epsilon^{-r}\epsilon^{-2}, \qquad \frac{\mathfrak{R}^*_{\mathrm{SO}}(T)}{T} = \widetilde{\Theta}\Big(T^{-1/(r+2)}\Big).$$

The denominator $r + 2$ adds the geometric exponent r to the statistical precision exponent 2. At the resulting resolution, accumulating an order-ϵ loss over T episodes gives $T\epsilon = T^{1-1/(r+2)}$, which explains the numerator $r+1$. This balance yields the continuum-bandit form of the exponent (Bubeck et al., 2011); the governing dimension in (16) comes from the Stackelberg response structure.

C.2 A COMMON FIXED-BUDGET IMPLEMENTATION

Write $R = R_U > 1$ and choose constants

$$\max\{R^{-1}, 2^{-1/2}\} < c_0 < 1, \qquad 0 < c_y \le \tfrac{1}{2}\sqrt{1 - c_0^2}, \qquad 0 < \rho \le \min\{1, \tfrac{1}{2}(R - c_0^{-1})\}. \tag{32}$$

Let $b : \Delta(\mathcal{Z}) \to [0, a_0]$ have a fixed Lipschitz constant L_b and vanish at the joint distribution $\mathcal{P}^\circ$ induced by the uniform follower response associated with a fixed population μ°. Choose a fixed $c_b > 0$ with $c_b a_0 \le c_y$ and $c_b L_b \le L_\phi$. Use two leader states y_0, y_1 and define the first two leader feature coordinates by

$$\phi(y_0, a, \mathcal{P}) = (c_0, c_b b(\mathcal{P})), \qquad \phi(y_1, a, \mathcal{P}) = (c_0, c_y). \tag{33}$$

Extra leader states have feature $(c_0, 0)$, and extra coordinates are zero. At step one let ω_1^* have vector masses $(1/(2c_0), 1/4)$ and $(1/(2c_0), -1/4)$ at (y_1, μ°) and (y_0, μ°), respectively. At every later step it has the single vector mass $(1/c_0, 0)$ at (y_0, μ°). The only nonzero leader reward coefficient is $\xi_2^* = (0, \rho)$. Set follower parameters at steps $h \ge 2$ to zero and start from (y_0, μ°).

These features have norm at most one and satisfy the required population Lipschitz bound. At step one, the probability of y_1 is $1/2 + \phi_2/4 \in [1/2, 3/4]$ for every joint distribution in $\Delta(\mathcal{Z})$, and the complementary probability reaches y_0. Later transitions are probability measures as well, and all population transitions reset to μ°, yielding the required causal factorization. Rewards take values in $[0, 1]$ and have coefficient norm at most one. For arbitrary $|v_0|, |v_1| \le 1$, the norm of the integrated step-one transition coefficient is

$$\left\| \left(\frac{v_1 + v_0}{2c_0}, \frac{v_1 - v_0}{4} \right) \right\|_2 \le \max\{c_0^{-1}, 1/2\} = c_0^{-1}.$$

The inequality follows by maximizing the convex norm over the vertices of the square. Later transition coefficients obey the same bound. Every Borel continuation in $[0, H]$ therefore has a backup coefficient of norm at most

$$\rho + H/c_0 \le RH,$$

which verifies Assumption 2 on the full domain. At step two the response is uniform and $b(\mathcal{P}^\circ) = 0$, so the only realized reward is $\rho c_y Y$, where Y indicates the first-step transition to y_1. For every full policy with initial commitment p, its value is consequently

$$J_\theta(p) = \frac{\rho c_y}{2} + c_{\mathrm{sig}} b(\mathcal{P}_\theta(z_1, p)), \qquad c_{\mathrm{sig}} = \frac{\rho c_y c_b}{4} > 0. \tag{34}$$

Later commitment choices do not change this value.

C.3 POSITIVE EFFECTIVE DIMENSION

Take $D = r \geq 1$, select D distinct nonbaseline follower state–action slots, and index D leader actions by $1, \ldots, D$, with a further action indexed by 0. For slot $i = (x_i, b_i)$ set its follower feature to $\kappa e_i \mathbf{1}\{a = i\}$, and set all baseline and unused features to zero. Choose fixed $\tau, \kappa > 0$ such that $2\tau\sqrt{D} < B_\Theta$, $\kappa \leq 1$, and $2\kappa\tau \leq 1$. The hidden first-step follower parameter belongs to $[\tau, 2\tau]^D \cup \{0\}$, padded with zeros. All true follower rewards are then in $[0, 1]$. Use the uniform population μ°.

Every relevant entry of a joint distribution induced by the constructed follower response is at least $1/(XB\exp(2\eta\kappa\tau))$. Let λ be half this positive constant and define on the entire ambient simplex

$$\ell_i(\mathcal{P}) = \mathrm{clip}_{[0,2\tau]}\left(\frac{\log\max\{\mathcal{P}(x_i, b_i), \lambda\} - \log\max\{\mathcal{P}(x_i, b_0), \lambda\}}{\eta\kappa}\right),$$

where b_0 is the baseline action in row x_i. This function is globally Lipschitz with a constant independent of the localization scale, and equals $\theta_i p_i$ on the joint distributions induced by the follower responses in this family. Fix $0 < c_* < \tau/(4D)$ and $0 < \epsilon_0 < c_*/4$. For $0 < \epsilon \leq \epsilon_0$, use $b = b_\epsilon$ from (30) in (33). The entire family has one common Lipschitz constant, and $b_\epsilon(\mathcal{P}^\circ) = 0$.

The optimal coordinates $p_i^* = c_*/\theta_i$ lie in a fixed D-dimensional box, and their sum is less than $1/4$. The map from the parameter cube to this box is bi-Lipschitz with fixed constants. Choose a packing at scale ϵ with $K_\epsilon \geq c_D \epsilon^{-D}$ centers so that their positive-bump regions U_j are pairwise disjoint. Separation in the first D commitment coordinates makes the regions disjoint in the full simplex, including any extra leader actions. Their centers achieve bump height ϵ; outside U_j the bump in alternative j is zero. The null parameter $\theta = 0$ has zero bump everywhere.

For an arbitrary SO learner, let $\mathbb{Q}_j$ denote the complete transcript law, including its chosen policies and internal randomization, under alternative j. At a common history the learner uses the same action-selection kernel in every testing model. Its own action is sampled from its announced commitment and contributes no conditional model divergence. The only model-dependent observation is the first-step Bernoulli transition, whose mean differs from the null by at most $C\epsilon\mathbf{1}\{p \in U_j\}$ and stays bounded away from zero and one. The population is fixed and the later observed reward is a common function of that transition. The chain rule therefore gives

$$\mathrm{KL}(\mathbb{Q}_0 \,\|\, \mathbb{Q}_j) \leq C\epsilon^2 \mathbb{E}_0 N_j, \qquad N_j = \sum_{t=1}^{T} \mathbf{1}\{p_1^t \in U_j\}.$$

Pinsker's inequality for the bounded statistic N_j/T, followed by Cauchy–Schwarz and $\sum_j N_j \leq T$, yields

$$\frac{1}{K_\epsilon}\sum_{j=1}^{K_\epsilon} \frac{\mathbb{E}_j N_j}{T} \leq \frac{1}{K_\epsilon} + C\epsilon\sqrt{\frac{T}{K_\epsilon}}. \tag{35}$$

Choose $\epsilon = c_{\mathrm{loc}} T^{-1/(D+2)}$ with a sufficiently small fixed $c_{\mathrm{loc}} > 0$ so that $\epsilon \leq \epsilon_0$ and the right-hand side of (35) is at most $1/4$ for every $T \geq 1$. There is an alternative j with $\mathbb{E}_j N_j \leq T/4$. Markov's inequality implies $\Pr_j\{N_j \leq T/2\} \geq 1/2$. Equation (34) gives, on every transcript,

$$\mathrm{Reg}_j(T) \geq c_{\mathrm{sig}}\epsilon(T - N_j).$$

Hence $\Pr_j\{\mathrm{Reg}_j(T) \geq c_{\mathrm{sig}}\epsilon T/2\} \geq 1/2$, proving (14) for $r \geq 1$ with the stated uniform probability $1/4$. The feature maps at this scale are common to the null and all alternatives. Allowing every $0 < \epsilon \leq \epsilon_0$ gives one fixed model class with the same numerical bounds and the quantifiers over feature maps specified in Section 3.5.

C.4 THE ZERO-DIMENSIONAL ENDPOINT

When $r = 0$, set every follower feature to zero and use two nonempty leader-action groups indexed by $I(a) \in \{0, 1\}$. Choose the constants in (32) and fix $0 < c_a < c_y$. Set

$$\phi(y_0, a, \mathcal{P}) = (c_0, c_a I(a)), \qquad \phi(y_1, a, \mathcal{P}) = (c_0, c_y).$$

The first-step transition has vector masses $(1/(2c_0), v)$ and $(1/(2c_0), -v)$ at (y_1, μ°) and (y_0, μ°), with hidden $v = \pm\Delta$ and $\Delta \leq 1/4$. Use only the step-two reward coefficient $(0, \rho)$ and the common

later transitions above. The norm and closure verification is unchanged, since the square-vertex transition bound is now $\max\{c_0^{-1}, 2\Delta\} = c_0^{-1}$.

For an initial commitment placing mass $p(I = 1)$ on group one, the value with optimal continuation is

$$\frac{\rho(c_y + c_a)}{2} + Lvp(I = 1), \qquad L = \rho c_a(c_y - c_a) > 0.$$

A suboptimal later decision only increases regret. Writing $F = \sum_{t=1}^{T} p_1^t(I = 1)$, the two models therefore satisfy on every transcript

$$\mathrm{Reg}_+(T) \geq L\Delta(T - F), \qquad \mathrm{Reg}_-(T) \geq L\Delta F.$$

The complete transcript KL divergence between them is at most $CT\Delta^2$. Choose $\Delta = c/\sqrt{T}$ with $c > 0$ sufficiently small so that their total-variation distance, in the convention $\sup_E |\mathbb{Q}_+(E) - \mathbb{Q}_-(E)|$, is at most $1/2$. For $E = \{F \geq T/2\}$, it follows that

$$\mathbb{Q}_-(E) + \mathbb{Q}_+(E^{\mathrm{c}}) \geq 1 - \mathrm{TV}(\mathbb{Q}_-, \mathbb{Q}_+) \geq \frac{1}{2}.$$

In at least one model, regret is at least $L\Delta T/2$ with probability at least $1/4$. This proves (14) also for $r = 0$. Multiplying the threshold in that equation by its probability lower bound yields (15), completing Theorem 3.

D PROBABILITY TRANSFER FOR THE MO LEARNER

D.1 THE DIAGNOSTIC LEARNER AND ITS REGRET GUARANTEE

The retained labels satisfy the main-text constraints in (18) and the same fixed-class excess-loss test. We detail the fiber geometry and executed diagnostic schedule before proving the probability transfer.

The remaining response uncertainty. Fix a step h and an augmented state $z = (s, \mu)$, and write

$$\alpha_z^\theta(p)_b = \sum_x \mu(x)\nu^{p,\theta}(b \mid x; s).$$

The parameters that agree with the true action marginal at every commitment form the fiber

$$F_z = \{\theta \in \Theta_{\mathrm{ad}} : \alpha_z^\theta(p) = \alpha_z^{\theta_h^*}(p) \text{ for every } p \in \Delta(\mathcal{A})\}. \tag{36}$$

Once a finite diagnostic grid is completed at (h, z), the learner can impose these constraints through the observed action marginals, as described below. The geometry of F_z then gives two ways to control the remaining exploration complexity.

The fibers in (36) are defined by finitely many analytic softmax equations and admit the uniform covering bound

$$N(\epsilon, F_z, \|\cdot\|_2) \le C\epsilon^{-\dim F_z}, \qquad 0 < \epsilon \le 1, \tag{37}$$

where N denotes the minimum number of radius-ϵ Euclidean balls needed to cover the fiber. Here $\dim F_z$ is its subanalytic dimension. The constant is uniform over admissible population weights and follower features, including fibers with singularities. The bound follows by applying uniform parameterization of bounded subanalytic families (Cluckers et al., 2020) to the finite action-marginal equations; Lemma 12 gives the uniform construction.

If $\dim F_z \le d_F - 1$, (37) reduces the parameter-side exploration dimension. If the fiber has full dimension, it contains an open subset on which each coordinate of the action marginal is constant as a function of θ. Analytic continuation then makes them constant for all θ, and their value at $\theta = 0$ is uniform. In this case the joint span of the bilinear logit map in (θ, p) lies in a blind linear space of dimension at most $C_{\mathrm{lin}}(X, B)$. The response-preserving optimization in Lemma 1 therefore admits a greedy commitment supported on at most $1 + C_{\mathrm{lin}}(X, B)$ actions. Minimum-support selection makes this reduced commitment dimension available to the learner as well, as proved in Lemmas 15 and 16.

These two possibilities control the same residual sequence. Together with the original SO bounds, they yield

$$r_{\mathrm{MO}} = \min\{d_F, A - 1, X(B - 1), \max\{C_{\mathrm{lin}}(X, B), d_F - 1\}\}. \tag{38}$$

The inner maximum accounts for the two possible fiber geometries, and the outer minimum retains the original SO complexity bounds. In (17), $C_{\mathrm{lin}}(X, B) < d_F$ implies $r_{\mathrm{MO}} = d_F - 1$.

Obtaining the constraints from executed feedback. There is a fixed finite grid $\mathcal{G}$, determined by the model dimensions, in the relative interior of $\Delta(\mathcal{A})$ on which equality of two action-marginal maps implies equality at every commitment. It has $K = (2XB^{2X-1})^{A-1}$ points; Lemma 9 gives the grid and proves this finite zero test. Thus, completing the grid at (h, z) identifies the fiber in (36) using the action marginals observed there.

The learner executes these commitments when it visits a step–state pair that has not completed the grid. Its episode-start policy specifies the next grid point at every unfinished pair and the first grid point at a newly encountered pair. At completed pairs it uses the optimistic greedy commitment, with minimum-support tie breaking. We call these greedy decisions made after grid completion non-diagnostic decisions. The probe schedule and fitted tuple are fixed from completed episodes, so this policy obeys the original commitment protocol. There are at most $Z = \max\{1, Sd_L\}$ reachable augmented states per step under the model assumptions. Consequently at most HZK diagnostic decisions are executed, as formalized in Lemmas 10 and 11. In the trajectory-level Bellman-error bound, each diagnostic decision contributes at most H, so their total contribution is at most H^2ZK, independently of T. The same martingale concentration argument used for SO then converts this bound into a high-probability regret guarantee.

The fitting targets continue to use the greedy candidate value $V_{Q,h}$ from (6) during diagnostic episodes. In the telescoping argument for regret, a non-diagnostic decision is charged through its Bellman residual, while a diagnostic decision is charged at most H in unnormalized return. This accounts for the exploration cost using the observations actually generated by the learner.

Regret guarantee. Using the same confidence radius $\beta_T(\delta)$, the learner consequently satisfies, with probability at least $1-\delta$,

$$\mathrm{Reg}_M(T) \le H^2ZK + CT^{\frac{r_{\mathrm{MO}}+1}{r_{\mathrm{MO}}+2}}\left[\log\frac{e(T+1)}{\delta}\right]^q. \tag{39}$$

Here C and q depend only on the fixed dimensions and bounds. Lemma 17 proves the residual-dimension bound; the next subsection supplies the probability transfer while charging all executed diagnostic decisions. Taking $\delta=(T+1)^{-2}$ gives an expected regret bound with the same temporal order for the minimax statements. The global confidence event still comes from the fixed classes used for SO; fiber covers sharpen the deterministic counting of compatible residuals. Combining (39) with $r_{\mathrm{MO}} = d_F - 1$ proves the improved upper bound in Theorem 5 under its stated capacity condition.

D.2 Trajectory-to-value probability transfer

The MO learner uses the same global confidence event G from Lemma 8 and the same radius $\beta_T(\delta)$. For every episode $t \ge 2$, use the filtration $\mathcal{F}_{t-1}$ from Appendix B and let Q^t be the tuple selected before the episode. When the retained set is empty, use the zero candidate tuple to define the analysis variables and execute the algorithm's fallback policy. Let $D_{t,h}$ indicate that the executed decision at step h is a scheduled diagnostic probe. Define the normalized greedy-decision gap at that step by

$$\Gamma_{t,h}(z) = V_{Q^t,h}(z) - \sum_a \pi_h^t(a \mid z) Q_h^t(z, \pi_h^t(z), a).$$

By Lemma 1, $V_{Q^t,h}$ is the maximum over the full simplex. Consequently $0 \le \Gamma_{t,h} \le 1$, and the gap is zero at every non-diagnostic decision. The Bellman telescoping argument and optimism therefore give, on the confidence event,

$$\frac{J_M^* - J_M(\pi^t)}{H} \le \mathbb{E}\left[\sum_h \{g_{t,h}(\chi_h^t) + \Gamma_{t,h}(z_h^t)\} \,\middle|\, \mathcal{F}_{t-1}\right].$$

At a diagnostic decision the signed summand has the useful identity

$$\mathbb{E}[g_{t,h}(\chi_h^t) + \Gamma_{t,h}(z_h^t) \mid z_h^t, \mathcal{F}_{t-1}] = V_{Q^t,h}(z_h^t) - \sum_a \pi_h^t(a \mid z_h^t)(\mathcal{B}_h V_{Q^t,h+1})(z_h^t, p_h^t, a) \le 1.$$

Thus the preceding conditional regret is at most the conditional expectation of

$$Y_t = \sum_{h=1}^H \big[(1-D_{t,h})|g_{t,h}(\chi_h^t)| + D_{t,h}\big], \qquad 0 \le Y_t \le H.$$

This identity charges one unit per diagnostic decision without requiring its Bellman residual to be small.

The number of executed probes is at most HZK on every trajectory. On the non-diagnostic decisions, the compatible residual dimension proved in Lemma 17 has exponent r_{MO}. The squared past-error bound still uses all observations, so it remains valid on the subsequence of non-diagnostic decisions. Applying the counting argument in (28) to this subsequence and summing over steps yields, on the confidence event,

$$\sum_{t=2}^T Y_t \le HZK + CH\left[T\epsilon + e_{\mathrm{MO}}(\epsilon) + \sqrt{\beta_T(\delta) T e_{\mathrm{MO}}(\epsilon)}\right], \qquad e_{\mathrm{MO}}(\epsilon) \le C\epsilon^{-r_{\mathrm{MO}}}[\log(C/\epsilon)]^q.$$

Under the original trajectory law, $\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] - Y_t$, for $t = 2, \dots, T$, is a martingale-difference sequence bounded in absolute value by H. The concentration bound in (29) therefore holds for its sum on an event G'_{MO} of probability at least $1-\delta/2$. On $G \cap G'_{\mathrm{MO}}$, including the first-episode cost H gives the following bound with probability at least $1-\delta$:

$$\mathrm{Reg}_M(T) \le H^2ZK + CH^2\left[1 + T\epsilon + e_{\mathrm{MO}}(\epsilon) + \sqrt{\beta_T(\delta) T e_{\mathrm{MO}}(\epsilon)} + \sqrt{T\log(2/\delta)}\right].$$

Taking $\epsilon = T^{-1/(r_{\mathrm{MO}}+2)}$ proves (39). Taking $\delta = (T+1)^{-2}$ and using $\mathrm{Reg}_M(T) \le HT$ on failure supplies the expected upper bound used in Theorem 5 and Corollary 7.

E Exact Marginal Diagnostics and Uniform Fiber Complexity

We first show that finitely many executed probes determine the action-marginal map at each encountered augmented state. We then bound the covering number of the compatible parameter fiber and use it to count sequential residual witnesses. These deterministic arguments combine with the fixed-class confidence event in Appendix B and the probability transfer in Appendix D.

E.1 Finite diagnostics under episode-start commitment

Fix a step and an augmented state $z = (s, \mu)$, and let $q = A - 1$. For $\theta \in \Theta_{\mathrm{ad}}$, write

$$\ell^{\theta}_{x,b}(p) = \eta \left\langle \sum_{a \in \mathcal{A}} p(a)\varphi(s, a, x, b), \theta \right\rangle, \qquad Z^{\theta}_{x}(p) = \sum_{c \in \mathcal{B}} e^{\ell^{\theta}_{x,c}(p)}.$$

Thus

$$\alpha^{\theta}_{z}(p)_b = \sum_{x \in \mathcal{X}} \mu(x) \frac{e^{\ell^{\theta}_{x,b}(p)}}{Z^{\theta}_{x}(p)}.$$

Lemma 9 (Finite zero test for action-marginal maps)**.** *Let*

$$M_{\mathrm{exp}} = 2XB^{2X-1}.$$

If $q = 0$, let $\mathcal{G}$ contain the unique commitment. If $q \geq 1$, choose the same M_{exp} distinct, equally spaced values in $[1/(4A), 1/(2A)]$ for each of the first q commitment coordinates, and let the last coordinate be the remaining probability mass. The resulting product set $\mathcal{G}$ lies in the relative interior of $\Delta(\mathcal{A})$ and contains

$$K = M_{\mathrm{exp}}^{A-1}$$

points. For any $\theta, \theta' \in \Theta_{\mathrm{ad}}$, it satisfies

$$\alpha^{\theta}_{z}(p) = \alpha^{\theta'}_{z}(p) \quad \textit{for every } p \in \mathcal{G} \quad \Longleftrightarrow \quad \alpha^{\theta}_{z}(p) = \alpha^{\theta'}_{z}(p) \quad \textit{for every } p \in \Delta(\mathcal{A}). \tag{40}$$

Proof of Lemma 9. For $q = 0$, the grid is the full commitment simplex and $K = 1$, so the equivalence holds directly. Suppose henceforth that $q \geq 1$. Write $p_A = 1 - \sum_{j=1}^{q} p_j$. Every logit $\ell^{\theta}_{x,b}(p)$ is then an affine function of $(p_1, \ldots, p_q)$. For a fixed follower action b, multiply the difference $\alpha^{\theta}_{z}(p)_b - \alpha^{\theta'}_{z}(p)_b$ by the strictly positive factor

$$D(p) = \prod_{x \in \mathcal{X}} Z^{\theta}_{x}(p) Z^{\theta'}_{x}(p).$$

For the summand indexed by x in $\alpha^{\theta}_{z}(p)_b$, cancellation of $Z^{\theta}_{x}(p)$ leaves the numerator exponential and $2X - 1$ row partition functions. Expanding those partition functions produces at most B^{2X-1} exponential terms. The same count holds for each of the X summands from θ'. Consequently, after collecting repeated exponents and deleting zero coefficients, $D(p)(\alpha^{\theta}_{z}(p)_b - \alpha^{\theta'}_{z}(p)_b)$ is a real exponential polynomial with at most $M_{\mathrm{exp}} = 2XB^{2X-1}$ distinct terms.

We use the following elementary zero bound. If $f(t) = \sum_{j=1}^{m} c_j e^{\lambda_j t}$ has distinct real exponents and is not identically zero, then it has at most $m - 1$ distinct real zeros. Indeed, after ordering the exponents and multiplying by $e^{-\lambda_1 t}$, differentiation leaves at most $m - 1$ exponential terms. Induction on m together with Rolle's theorem proves the claim; if the derivative vanishes identically, the normalized function is a nonzero constant. The same statement applies after repeated exponents of an exponential polynomial have been combined.

Now suppose that the action-marginal maps agree on $\mathcal{G}$. Fix all grid coordinates except p_1. The preceding univariate bound and the M_{exp} distinct values of p_1 imply that the cleared difference vanishes for every real value of p_1 while the remaining coordinates stay on their grid values. Fix such an arbitrary p_1 and repeat the argument for p_2. Iterating over all q coordinates shows that the cleared difference vanishes on all of $\mathbb{R}^q$. Since $D(p) > 0$, the difference between the corresponding action-marginal coordinates also vanishes wherever p is a commitment. Doing this for every $b \in \mathcal{B}$

proves the forward implication in (40). The reverse implication is immediate. Finally, every grid point lies in the relative interior of the simplex because

$$\sum_{j=1}^{A-1} p_j \le \frac{A-1}{2A} < 1,$$

while every displayed coordinate is strictly positive. □

The grid therefore identifies the action-marginal map and its compatible parameter fiber. Parameters in the same fiber may induce different joint state–action distributions, which is why the subsequent analysis studies the complexity of that fiber.

Lemma 10 (Finite reachable augmented states)**.** *At every step, the set of augmented states reachable under a fixed game in Definition 1 has cardinality at most*

$$Z = \max\{1, Sd_L\}.$$

Proof of Lemma 10. Fix a transition step h and push each of the d_L coordinate signed measures of ω_h^* forward under the projection $(s', \mu') \mapsto \mu'$. For every current input $(s, a, \mathcal{P})$, the population marginal of the augmented transition kernel belongs to the linear span of these d_L fixed signed measures by Definition 1. By the causal factorization (1), that same marginal is the Dirac measure

$$\delta_{\Phi_h(s,a,\mu,\nu)}.$$

Singletons in the population simplex are Borel sets. Any finite collection of distinct Dirac measures is therefore linearly independent: evaluating a linear relation on the singleton containing one of their distinct atoms forces the corresponding coefficient to vanish. Hence at most d_L distinct population distributions can occur at the next step. Combining them with the S possible leader states gives at most Sd_L augmented states at every noninitial step. The initial augmented state is deterministic, which yields the stated bound $Z = \max\{1, Sd_L\}$ uniformly over steps. □

Lemma 11 (Episode-frozen diagnostic schedule)**.** *The grid in Lemma 9 can be acquired using actual MO feedback while preserving the episode-start commitment protocol. Along every trajectory, the total number of executed diagnostic decisions is at most HZK.*

Proof of Lemma 11. Before every episode $t \ge 2$, use only completed episodes to record, for every previously encountered pair (h, z), the number $n_{h,z}^{t-1} \in \{0, \dots, K\}$ of grid probes already executed there. When the retained set is nonempty, the full policy announced for episode t is defined as follows. At a stored unfinished pair it selects grid point $n_{h,z}^{t-1} + 1$. At a previously unseen pair it selects the first grid point. At a completed pair it uses the optimistic greedy commitment with the minimum-support tie breaking specified in the main text. If the retained set is empty, the learner uses the initialization fallback inherited from Algorithm 1; on the simultaneous confidence event this case never occurs because the true tuple is retained. The fitted tuple, the counters, and the entire contingent rule used in the episode are fixed before the episode starts.

For each step h, only finitely many singleton augmented states have been stored before episode t. The diagnostic assignments on those singletons, the default first probe off the stored set, and the assumed Borel greedy selection together define a Borel map of the augmented state. Thus the rule belongs to the policy class in Section 3.3. A fixed step is visited only once in one episode, so an (h, z) counter never needs to be updated and reused within the same episode. The newly observed action marginal is stored only after its executed commitment and can therefore enter the compatibility constraints only in later episodes.

By Lemma 10, there are at most HZ reachable step–state pairs. Each pair receives at most the K probes in $\mathcal{G}$. This gives the pathwise bound HZK. The count charges each probe when it is executed, so it also covers pairs whose grids are only partially completed by episode T. The continuation fitted on diagnostic episodes remains the greedy candidate continuation $V_{Q,h}$ from (6); the resulting diagnostic cost is therefore exactly the one handled in Appendix D. □

E.2 Uniform covering of compatible parameter fibers

Once all probes at z have been executed, Lemma 9 turns the observed grid equalities into the full compatibility condition in (36). The next result justifies the uniform covering-number bound (37), including singular fibers and population vectors on the boundary of the simplex.

Lemma 12 (Uniform proper covers of compatible fibers)**.** *There is a constant $C_{\mathrm{fib}} < \infty$, depending only on the fixed dimensions and numerical budgets defining $\mathfrak{C}$, such that every fiber F_z in* (36) *satisfies*

$$N(\epsilon, F_z, \|\cdot\|_2) \le C_{\mathrm{fib}} \epsilon^{-\dim F_z}, \qquad 0 < \epsilon \le 1. \tag{41}$$

The covering centers may be chosen in F_z itself. The constant is uniform over the follower features, the population weights, and the true follower parameter allowed by the model class.

Proof of Lemma 12. By Lemma 9, membership in F_z is equivalent to the finitely many equations

$$\alpha_z^{\theta}(g)_b = \alpha_z^{\theta_h^*}(g)_b, \qquad g \in \mathcal{G}, \quad b \in \mathcal{B},$$

together with the constraints defining Θ_{ad}. Treat the finitely many entries of the follower feature map, the population vector μ, and the reference parameter θ_h^* as parameters of this family. Their ranges satisfy the specified feature and parameter bounds. For $\theta \in \Theta$, all softmax logit arguments lie in a fixed bounded interval depending only on ηB_Θ. After clearing the positive row denominators, every grid equality is therefore an equation formed from field operations and exponential functions restricted to a fixed compact interval. After an affine rescaling, these restricted exponentials are restricted analytic functions. The Euclidean norm constraint, the simplex constraints, and the finitely many follower-reward inequalities defining Θ_{ad} are semialgebraic. Hence the collection of fibers is one bounded family definable in the restricted analytic structure $\mathbb{R}_{\mathrm{an}}$. The defining equations involve only the follower features, population weights, and follower parameters.

If $d_F = 0$, the nonempty fiber is a singleton, so the claimed bound holds directly. Assume henceforth that $d_F \ge 1$ and choose a fixed affine map S that sends the bounded ambient parameter box containing Θ into a compact subbox of $(0,1)^{d_F}$. Write $\widetilde{F}_z = \mathsf{S}(F_z)$. The dimension of a fiber in a definable family is a definable integer-valued function of the external parameters, so the parameter space can be partitioned into finitely many definable pieces on which the fiber dimension is a fixed value j. On such a piece, Cluckers et al. (2020, Theorem 2.1.3), with $m = j$, $n = d_F$, and differentiability order $r = 1$, supplies a family-uniform finite number of analytic maps

$$\psi_i : (0,1)^j \longrightarrow \widetilde{F}_z$$

whose ranges cover $\widetilde{F}_z$ and whose C^1 norms are at most one. The ranges cover the entire rescaled fiber, including its singular points. For $j = 0$, the domain $(0,1)^0$ is a singleton and the same theorem gives a uniform bound on the fiber cardinality.

The maps $\mathsf{S}^{-1} \circ \psi_i$ take values in F_z and have one common Euclidean Lipschitz constant $L < \infty$ depending only on the fixed budgets and dimensions. For $j \ge 1$, set

$$N_\epsilon = \max\left\{1, \left\lceil \frac{2L\sqrt{j}}{\epsilon} \right\rceil \right\}$$

and use the product grid with one-dimensional coordinates $(r - 1/2)/N_\epsilon$, $r = 1, \dots, N_\epsilon$. Every point of $(0,1)^j$ lies within Euclidean distance $\sqrt{j}/(2N_\epsilon)$ of a grid point, so the grid mapped by $\mathsf{S}^{-1} \circ \psi_i$ is an ϵ-net for the range of this map. Every center is an image point and therefore belongs to F_z, while the number of grid points is at most $C_j \epsilon^{-j}$. Multiplying by the uniform number of charts and taking the maximum over the finitely many dimensions $j = 0, \dots, d_F$ proves (41) with a constant uniform over the entire admissible family. The bound applies uniformly to all dimension strata, including boundary and singular fibers. □

E.3 From fiber covering numbers to sequential residual complexity

We now use the cover in Lemma 12 to bound the sequential dimension defined in Appendix B.4. Within each covering cell, centering the nonlinear feature reduces residual prediction to a linear problem with controlled approximation error.

Lemma 13 (Compatible residual counting from a fiber cover)**.** *Fix a diagnosed augmented state z at a step h and suppose $\dim F_z \le k$. Consider the non-diagnostic residual queries at this same (h, z) for which the candidate labels satisfy the exact MO compatibility constraints. Their sequential dimension at resolution ϵ is at most*

$$C\epsilon^{-k}\big[\log(C/\epsilon)\big]^q, \qquad 0 < \epsilon \le 1, \tag{42}$$

for constants depending only on the fixed dimensions and bounds specified in Section 3.5. The same conclusion remains valid after restricting to any further subsequence of non-diagnostic queries.

Proof of Lemma 13. At an augmented state with a completed diagnostic grid, every retained candidate parameter agrees with the true action marginal at all grid points. By Lemma 9, its parameter therefore belongs to F_z. The enlarged residual in (21) can be written as a linear prediction in the $2d_L$-dimensional feature

$$x_\theta(z, p, a) = \big(\phi(s, a, \mathcal{P}_\theta(z, p)), \phi(s, a, \mathcal{P}_{\theta_h^*}(z, p))\big)$$

with coefficient $(w, -v)$ of norm at most $R_b = \sqrt{2}W$. The softmax map is uniformly Lipschitz in θ on the bounded parameter set. Combining this with Assumption 1 gives a finite constant L_θ such that replacing θ by $\bar\theta$ changes every residual prediction by at most $L_\theta\|\theta - \bar\theta\|_2$. Set $\bar L_\theta = \max\{1, L_\theta\}$.

Fix a common independence scale $\epsilon' \ge \epsilon$, set $m = 2d_L$, and put

$$L_\epsilon = \lceil C_0 m \log(C_0/\epsilon)\rceil, \qquad \delta_\epsilon = \frac{\epsilon}{16\bar L_\theta\sqrt{L_\epsilon}}.$$

By Lemma 12, F_z has a proper δ_ϵ-cover of size at most

$$C\delta_\epsilon^{-k} \le C\epsilon^{-k}L_\epsilon^{k/2}.$$

Assign each witness parameter to one of these centers and suppose that one center $\bar\theta$ receives $L_\epsilon + 1$ witnesses from an ϵ'-independent sequence. Keep the first $L_\epsilon + 1$ such witnesses in their original order and write their coefficients as $\beta_i = (w_i, -v_i)$, so $\|\beta_i\|_2 \le R_b$. For a query χ, let

$$\bar g_i(\chi) = \langle x_{\bar\theta}(\chi), \beta_i\rangle.$$

The approximation error at every query is at most $\epsilon/(16\sqrt{L_\epsilon})$. For every retained position i, the original independence inequalities and the triangle inequality therefore give

$$\left(\sum_{j<i}\bar g_i(\chi_j)^2\right)^{1/2} \le \frac{17}{16}\epsilon', \qquad |\bar g_i(\chi_i)| > \frac{15}{16}\epsilon'.$$

Indeed, there are at most L_ϵ preceding approximation errors, whose Euclidean norm is at most $\epsilon/16 \le \epsilon'/16$, and the current approximation error is at most $\epsilon'/16$.

Let $x_i = x_{\bar\theta}(\chi_i)$ and set

$$V_0 = \lambda I, \qquad V_i = V_{i-1} + x_i x_i^\top, \qquad \lambda = \frac{(17/16)^2(\epsilon')^2}{R_b^2}.$$

The preceding-prediction bound and $\|\beta_i\|_2 \le R_b$ imply

$$\|\beta_i\|_{V_{i-1}}^2 = \lambda\|\beta_i\|_2^2 + \sum_{j<i}\bar g_i(\chi_j)^2 \le 2\left(\frac{17}{16}\right)^2(\epsilon')^2.$$

Cauchy–Schwarz in the V_{i-1} norm and the lower bound on $|\bar g_i(\chi_i)|$ yield

$$x_i^\top V_{i-1}^{-1}x_i \ge c_* := \frac{(15/16)^2}{2(17/16)^2} > 0.$$

The matrix determinant lemma gives one factor $1 + x_i^\top V_{i-1}^{-1}x_i$ per retained query, whereas $\|x_i\|_2 \le \sqrt{2}$ and the arithmetic–geometric mean inequality give

$$M\log(1 + c_*) \le \log\frac{\det V_M}{\det(\lambda I)} \le m\log\left(1 + \frac{CM}{m(\epsilon')^2}\right) \le m\log\left(1 + \frac{CM}{m\epsilon^2}\right).$$

To make the last comparison explicit, set $u = \log(C_0/\epsilon)$, so $u \geq \log C_0$ and $L_\epsilon + 1 \leq 2C_0 mu$ after enlarging C_0 once. The right-hand side above is then at most $m(3u + C_1)$ for a fixed C_1, whereas the left-hand side is at least $C_0 mu \log(1 + c_*)$. Choosing the fixed constant C_0 so that $C_0 \log(1 + c_*) > 3 + C_1/\log C_0$ rules out $M = L_\epsilon + 1$ simultaneously for all $0 < \epsilon \leq 1$. Thus each cover center receives at most L_ϵ independent witnesses, and multiplying by the number of centers proves (42). The argument applies to every subsequence of compatible queries: its witnesses satisfy the same independence inequalities, and discarding earlier queries can only decrease their squared-prediction sums. □

F Blind Response Geometry and the MO Upper Bound

We now prove the linear-capacity facts used when a diagnosed fiber has full dimension and then combine the two fiber branches into the effective MO complexity in (38).

F.1 Balanced partitions and the geometric capacity

Fix $1 \leq m \leq X$ occupied follower states with positive weights $\mu_1, \ldots, \mu_m$ summing to one, and let $\mathcal{I}_m = [B]^m$. For $i = (i_1, \ldots, i_m) \in \mathcal{I}_m$ and a gauged logit table u with $u_{xB} = 0$, define

$$L_i(u) = \sum_{x=1}^{m} u_{x,i_x}.$$

A partition Π of $\mathcal{I}_m$ is called balanced for μ if

$$B \sum_{x=1}^{m} \mu_x \#\{i \in G : i_x = b\} = |G| \qquad \text{for every } G \in \Pi \text{ and } b \in [B]. \tag{43}$$

Associated with such a partition is the linear space

$$V_\Pi = \{u : u_{xB} = 0, \quad L_i(u) = L_j(u) \text{ whenever } i, j \in G \in \Pi\}. \tag{44}$$

Lemma 14 (Balanced-partition characterization of linear blindness)**.** *The geometric capacity $C_{\mathrm{lin}}(X, B)$ in Definition 2 equals the maximum of $\dim V_\Pi$ over all $m \leq X$, all positive population weights μ, and all partitions Π satisfying* (43)*. Moreover,*

$$C_{\mathrm{lin}}(X, B) \leq X - 1. \tag{45}$$

Proof of Lemma 14. Fix m and positive weights μ. For an action b, the uniform-aggregate identity from Definition 2 is

$$B \sum_{x=1}^{m} \mu_x \frac{e^{u_{xb}}}{\sum_c e^{u_{xc}}} = 1.$$

Multiplying by the positive denominator $\prod_x \sum_c e^{u_{xc}}$ and expanding gives

$$\sum_{i \in \mathcal{I}_m} \left(B \sum_{x=1}^{m} \mu_x \mathbf{1}\{i_x = b\} - 1 \right) e^{L_i(u)} = 0. \tag{46}$$

If $u \in V_\Pi$, all exponentials indexed by one block G have the same exponent. The sum of their coefficients in (46) is

$$B \sum_x \mu_x \#\{i \in G : i_x = b\} - |G|,$$

which vanishes by (43). Thus every balanced space V_Π is blind in the sense of Definition 2.

Conversely, let V be any blind linear space for the fixed positive weights and partition $\mathcal{I}_m$ by equality of the linear functionals $L_i|_V$. If $V = \{0\}$, this produces one block and (43) follows directly by counting, so suppose $\dim V > 0$. Choose $v \in V$ outside the finite union of hyperplanes on which two distinct functionals $L_i|_V$ and $L_j|_V$ take the same value. Along the line $u = tv$, equation (46) becomes a univariate exponential polynomial whose distinct exponents are indexed by the blocks of the partition. Linear independence of distinct real exponentials forces the summed coefficient of

every block to vanish. This is exactly (43). By construction, equality of $L_i|_V$ within a block also gives $V \subseteq V_\Pi$. Taking dimensions and then maximizing proves the equivalence with Definition 2.

It remains to prove (45). Let V be a blind space of dimension r on m positive-mass rows, and choose a linear isomorphism $t \mapsto u(t)$ from $\mathbb{R}^r$ onto V. Write

$$u_{xb}(t) = f_{xb}^\top t, \qquad t \in \mathbb{R}^r, \qquad f_{xB} = 0.$$

Let $q_{xb}(t) = \mathrm{softmax}_b(u_{x\cdot}(t))$. At $t = 0$, every row response is uniform. Set

$$\bar{f}_x = \frac{1}{B}\sum_{c=1}^{B} f_{xc}, \qquad v_{xb} = f_{xb} - \bar{f}_x.$$

Differentiating the constant aggregate $\sum_x \mu_x q_{xb}(t) = 1/B$ once and twice at zero gives, for every b,

$$\sum_{x=1}^{m} \frac{\mu_x}{B} v_{xb} = 0 \tag{47}$$

and

$$\sum_{x=1}^{m} \frac{\mu_x}{B} v_{xb} v_{xb}^\top = \sum_{x=1}^{m} \frac{\mu_x}{B}\, \mathrm{Cov}_{c\sim\mathrm{Unif}([B])}(f_{xc}). \tag{48}$$

The matrix on the right is positive definite on $\mathbb{R}^r$. Indeed, if y lies in its kernel, positivity of the weights implies that $y^\top f_{xb}$ is constant in b for every occupied row x. Since $f_{xB} = 0$, all these constants are zero, so $u(y) = 0$. Injectivity of the parametrization forces $y = 0$. Hence the matrix on the left of (48) also has rank r.

That matrix is generated by the m vectors $\sqrt{\mu_x/B}\, v_{xb}$. Equation (47) is a nontrivial linear relation among these vectors with strictly positive coefficients $\sqrt{\mu_x/B}$. These vectors therefore span a space of dimension at most $m-1$. Thus $r \le m-1 \le X-1$, proving (45). The positive-weight convention in Definition 2 applies this argument to the occupied rows of any population. □

F.2 A COMMON BLIND SPACE FOR ALL BILINEAR COEFFICIENT COLUMNS

The full-dimensional fiber branch requires a space that contains the response variation of every commitment. The next lemma identifies such a common blind space for the complete bilinear logit map.

Lemma 15 (Joint coefficient-column containment)**.** *Fix an augmented state $z = (s, \mu)$ and remove the zero-mass follower rows. Suppose that*

$$\alpha_z^\theta(p)_b = \frac{1}{B} \qquad \textit{for every } \theta \in \mathbb{R}^{d_F}, \quad p \in \Delta(\mathcal{A}), \quad b \in [B]. \tag{49}$$

Gauge the row logits by follower action B and write

$$u_{xb}(\theta, p) = \sum_{a\in\mathcal{A}} \sum_{\ell=1}^{d_F} p(a)\theta_\ell L_{xb,a\ell}, \qquad u_{xB} = 0. \tag{50}$$

Then all coefficient tables $L^{a,\ell} = (L_{xb,a\ell})_{x,b}$ lie in one balanced space V_Π. Consequently their joint span has dimension at most $C_{\mathrm{lin}}(X, B)$. For every fixed θ, the restriction of the linear map $p \mapsto u(\theta, p)$ to the simplex tangent space $\{v \in \mathbb{R}^A : \mathbf{1}^\top v = 0\}$ has rank at most $\min\{A-1, C_{\mathrm{lin}}(X, B)\}$.

Proof of Lemma 15. Use the denominator-cleared identity (46). For a multi-index $i \in \mathcal{I}_m$, its exponent under (50) can be written

$$L_i(u(\theta, p)) = \theta^\top r_i(p),$$

where $r_i(p) \in \mathbb{R}^{d_F}$ is a vector-valued affine function on the simplex. For every pair i, i', the coincidence set $\{p : r_i(p) = r_{i'}(p)\}$ is either the whole affine hull of the simplex or a proper affine subset. There are only finitely many pairs. Choose a commitment p° in the relative interior of the simplex outside the union of all proper coincidence sets. When $A = 1$, the affine hull is a singleton

and the proper coincidence sets are empty, so the unique commitment has this property. At p°, two exponent vectors $r_i(p^\circ)$ and $r_{i'}(p^\circ)$ are equal if and only if the corresponding affine functions agree identically on the simplex.

By (49), the cleared identity holds for every θ at p°. Group the multi-indices by equality of $r_i(p^\circ)$. Distinct multivariate exponentials $e^{\theta^\top r}$ are linearly independent: choose a direction in θ-space avoiding the finitely many hyperplanes on which two distinct exponent vectors have the same projection, and reduce to the univariate exponential independence used in Lemma 14. Hence the summed coefficient of every group in (46) vanishes. The resulting partition Π is balanced.

If i, i' belong to one block, the generic choice of p° implies $r_i(p) = r_{i'}(p)$ for every commitment p. Evaluate this affine identity at the pure commitment $p = e_a$. Since it holds for every θ, equality of the coefficient of each θ_ℓ gives

$$\sum_x L_{x,i_x,a\ell} = \sum_x L_{x,i'_x,a\ell} \qquad \text{for every } a, \ell.$$

This is exactly the condition $L^{a,\ell} \in V_\Pi$ from (44). Thus every bilinear coefficient column lies in the same balanced space, whose dimension is at most $C_{\mathrm{lin}}(X, B)$ by Lemma 14. For fixed θ, every action column of the map in p is a linear combination of the tables $L^{a,\ell}$ and hence also lies in V_Π. Restricting its domain to the $(A-1)$-dimensional simplex tangent space gives the final rank bound. □

Lemma 16 (Small-support greedy commitments on full-dimensional fibers)**.** *If a diagnosed fiber F_z has dimension d_F, then the action marginal at z is uniform for every $\theta \in \mathbb{R}^{d_F}$ and every commitment p. For every candidate parameter θ, the greedy optimization at z therefore has a maximizer supported on at most*

$$1 + \min\{A - 1, C_{\mathrm{lin}}(X, B)\}$$

leader actions. The minimum-support tie breaking used by the learner selects a maximizer obeying this support bound without knowing in advance that the fiber is full dimensional.

Proof of Lemma 16. A definable subset of $\mathbb{R}^{d_F}$ with dimension d_F contains a nonempty open cell. Thus F_z contains a nonempty open set U. Fix any commitment p and follower action b. The function

$$\theta \longmapsto \alpha_z^\theta(p)_b - \alpha_z^{\theta_h^*}(p)_b$$

is real analytic on all of $\mathbb{R}^{d_F}$ and vanishes on U by the definition of the fiber. The real-analytic identity theorem makes it identically zero. Since p and b were arbitrary, the action marginal is independent of θ for every commitment. Evaluating this analytic identity at $\theta = 0$ gives the uniform action marginal $1/B$.

Apply Lemma 15 to the occupied rows. For a fixed candidate θ, the logit-contrast map has rank at most $\min\{A - 1, C_{\mathrm{lin}}(X, B)\}$ on the simplex tangent space. Together with the normalization $\mathbf{1}^\top p = 1$, the response-preserving equalities therefore have rank at most one plus this quantity. Maximizing the leader objective over the response-preserving polytope from the proof of Lemma 1 has an extreme-point optimizer with at most this many positive coordinates. Since $C_{\mathrm{lin}}(X, B) \le X(B-1)$ by Definition 2, this support size is at most $k + 1$, so the optimizer belongs to the original search set $\mathcal{P}_k$. The learner chooses, among all greedy maximizers in $\mathcal{P}_k$, one with minimum support, so its selected non-diagnostic commitment cannot have larger support. □

F.3 Residual complexity and the MO upper bound

Lemma 17 (MO residual dimension)**.** *For the non-diagnostic decisions of the MO learner, the sequential residual dimension satisfies*

$$e_{\mathrm{MO}}(\epsilon) \le C\epsilon^{-r_{\mathrm{MO}}} \left[\log(C/\epsilon)\right]^q, \qquad 0 < \epsilon \le 1, \tag{51}$$

where

$$r_{\mathrm{MO}} = \min\{d_F, A - 1, X(B-1), \max\{C_{\mathrm{lin}}(X, B), d_F - 1\}\}.$$

Proof of Lemma 17. Partition the non-diagnostic queries by step and realized augmented state. By Lemma 11, a non-diagnostic decision is made at a pair only after its diagnostic grid has been completed. There are only finitely many such pairs by Lemma 10.

If $\dim F_z \le d_F - 1$, apply Lemma 13 with $k = d_F - 1$. This bounds the number of compatible witnesses at that augmented state by

$$C\epsilon^{-(d_F-1)}[\log(C/\epsilon)]^q.$$

If instead $\dim F_z = d_F$, Lemma 16 places every non-diagnostic minimum-support greedy commitment in the union of simplex faces of dimension at most $\min\{A-1, C_{\mathrm{lin}}(X,B)\}$. For each fixed (h,z,a) and each such face, the enlarged residual class is uniformly Lipschitz in the commitment with a finite common constant L_p. Set $\bar{L}_p = \max\{1, L_p\}$ and cover the face by $O(\epsilon^{-\min\{A-1, C_{\mathrm{lin}}(X,B)\}})$ cells of ℓ_1 diameter at most $\epsilon/(2\bar{L}_p)$. If five queries in one cell belonged to an independent sequence at a common scale $\epsilon' \ge \epsilon$, consider the witness g for the fifth query. For each of the four earlier queries in the same cell, commitment Lipschitzness gives

$$|g(\chi_i)| \ge |g(\chi_5)| - L_p\|p_i - p_5\|_1 > \epsilon' - \epsilon/2 \ge \epsilon'/2.$$

The four preceding squared predictions would therefore sum to more than $(\epsilon')^2$, contradicting the independence condition. Thus each cell contributes at most four witnesses. The finite number of augmented states, actions, and faces changes only the fixed constant. The two fiber types together are therefore bounded by

$$C\epsilon^{-\max\{C_{\mathrm{lin}}(X,B), d_F-1\}}[\log(C/\epsilon)]^q,$$

where the separate cap $A-1$ may be retained outside this expression.

The original SO parameter, commitment, and response-preserving support arguments remain valid for the same learner because the MO compatibility equations only remove candidate labels and diagnostic points are omitted from the non-diagnostic sequence. They give the additional exponents d_F, $A-1$, and $X(B-1)$. The preceding bounds concern the deterministic sequential dimension of the residual class. For the actual fitted residuals selected by the learner, the global event in (24) bounds the squared error against all previous observations at the same step, including diagnostic observations. Deleting diagnostic indices and then restricting the remaining indices to any one of the finitely many groups indexed by augmented state, realized leader action, and support face used above can only decrease that squared sum. Thus (28) applies on each non-diagnostic subsequence with the same bound $C\beta_T(\delta)$ on accumulated squared Bellman errors over all past observations. Summing the resulting bounds over the fixed number of analysis classes changes only a constant depending on the fixed dimensions and bounds. Taking the strongest of the four exponent bounds proves (51) with the stated r_{MO}. □

Proof of the MO upper bound in (39). Use the confidence event from Lemma 8 with the unchanged radius $\beta_T(\delta)$. The exact marginal equations only intersect the retained candidate set with additional deterministic compatibility constraints, so the true tuple remains retained and the squared historical residual bound remains the one proved for the fixed global classes. On non-diagnostic decisions, Lemma 17 supplies the complexity bound required in Appendix D. On diagnostic decisions, Lemma 11 gives the pathwise count HZK, and the fitted continuation remains the greedy candidate continuation as required by that appendix. Therefore Appendix D gives, on an event of probability at least $1-\delta$,

$$\mathrm{Reg}_M(T) \le H^2ZK + CH^2\left[1 + T\epsilon + e_{\mathrm{MO}}(\epsilon) + \sqrt{\beta_T(\delta)Te_{\mathrm{MO}}(\epsilon)} + \sqrt{T\log(2/\delta)}\right].$$

Taking $\epsilon = T^{-1/(r_{\mathrm{MO}}+2)}$ and using (51) absorbs the fixed powers of H and all logarithmic terms into constants C, q depending only on the fixed dimensions and bounds specified in Section 3.5. This is exactly (39). Finally, choosing $\delta = (T+1)^{-2}$ and using $0 \le \mathrm{Reg}_M(T) \le HT$ on the failure event gives the expected upper bound of the same temporal order used in Theorem 5 and Corollary 7. □

G MO Lower Bounds, the Capacity Criterion, and Separation

This section completes the information-theoretic side of the MO results. The construction combines the fixed-budget leader construction from Appendix C with a linear blind space whose action marginal is identical throughout the testing family. We verify the resulting information bound for the complete MO transcript and then derive the capacity criterion and the matching separation.

G.1 Why the action marginal preserves the testing difficulty

Why cancellation can preserve the lower bound. For any $1 \le D \le \min\{d_F, A-1, C_{\mathrm{lin}}(X,B)\}$, choose D independent logit directions in a blind linear space. The lower-bound construction of Section 5.1 can encode its hidden coordinates in these directions while keeping every executed action marginal uniform. Row-wise shifts of follower rewards preserve the logit contrasts and ensure nonnegative follower rewards. The extra MO observation is then identical throughout the testing family, and the complete-transcript information bound remains (31). Lemma 18 gives the full construction and its constant-probability guarantee, which implies the lower bound $\Omega(T^{(D+1)/(D+2)})$. When $C_{\mathrm{lin}}(X,B) \ge d_F$, taking $D = d_F$ proves the remaining direction of Theorem 5. The criterion thus reflects either a reduction in compatible parameter variation or a residual family of strategically relevant changes that cancel in the aggregate. Appendix G.3 combines these two arguments into the proof of Theorem 5.

G.2 Blind lower-bound families

Lemma 18 (MO lower bound inside a blind linear family)**.** *Assume $S, H, d_L \ge 2$ and let*

$$1 \le D \le \min\{d_F, A-1, C_{\mathrm{lin}}(X,B)\}.$$

There is a constant $c > 0$, depending only on the fixed dimensions and bounds specified in Section 3.5, such that for every MO algorithm A and every $T \ge 1$, some $(\sigma, M) \in \mathfrak{C}$ satisfies

$$\Pr_M^{\mathsf{A}(\sigma)}\left\{\mathrm{Reg}_M(T) \ge cT^{\frac{D+1}{D+2}}\right\} \ge \frac{1}{4}. \tag{52}$$

Consequently,

$$\inf_{\mathsf{A}\in\mathcal{A}_{\mathrm{MO}}} \sup_{(\sigma,M)\in\mathfrak{C}} \mathbb{E}_M^{\mathsf{A}(\sigma)}[\mathrm{Reg}_M(T)] \ge \frac{c}{4} T^{\frac{D+1}{D+2}}. \tag{53}$$

At each localization scale, every testing model uses the same feature maps, and the executed action marginal is the uniform distribution on $\mathcal{B}$ for every possible commitment.

Proof of Lemma 18. We construct admissible blind responses, decode their hidden coordinates into a leader feature, and construct a packing of the resulting commitment-value functions before bounding the information in the complete transcript.

Admissible blind responses. By the definition of $C_{\mathrm{lin}}(X,B)$ and Lemma 14, choose $m \le X$, positive weights $\mu_1, \dots, \mu_m$, and D linearly independent gauged tables

$$w_i \in \mathbb{R}^{m\times B}, \qquad w_i(x,B) = 0, \qquad i \in [D],$$

whose span V is blind for the population $\mu^\circ = (\mu_1, \dots, \mu_m, 0, \dots, 0)$. Thus

$$\sum_{x=1}^{m} \mu_x \,\mathrm{softmax}_b(u_{x\cdot}) = \frac{1}{B} \qquad \text{for every } u \in V \text{ and } b \in [B]. \tag{54}$$

Choose distinct leader actions $a_1, \dots, a_D$ and a further baseline action a_0, which is possible because $D \le A-1$. For each basis table define the row-common shift

$$m_i(x) = -\min_{b\in\mathcal{B}} w_i(x,b) \ge 0.$$

At the initial leader state y_0, define follower-feature coordinate i by

$$\varphi_i(y_0, a, x, b) = \kappa \mathbf{1}\{a = a_i\}\big(w_i(x,b) + m_i(x)\big), \qquad i \in [D], \quad x \in [m]. \tag{55}$$

Set all follower features on unoccupied follower states to zero. Set the remaining follower-feature coordinates, other leader actions, and follower features at other leader states to zero as well. Because only one of the coordinates in (55) can be nonzero at a fixed leader action, a fixed sufficiently small $\kappa > 0$ makes the follower feature have Euclidean norm at most one. Choose a fixed $\tau > 0$ such that $2\tau\sqrt{D} < B_\Theta$ and, after decreasing κ if necessary,

$$0 \le 2\tau\kappa\big(w_i(x,b) + m_i(x)\big) \le 1 \qquad \text{for every } i, x, b.$$

Hence every parameter in $[\tau, 2\tau]^D$, padded by zeros, belongs to Θ and generates follower rewards in $[0, 1]$. The null parameter 0 is legal as well. The row shifts use the same follower-feature coordinates and therefore consume no additional reward coordinate.

For a commitment p, the logit contrast in occupied row x between follower action b and the baseline action B is

$$\eta \langle \varphi^p(y_0, x, b) - \varphi^p(y_0, x, B), \theta \rangle = \eta\kappa \sum_{i=1}^{D} \theta_i p(a_i) w_i(x, b). \tag{56}$$

The row-common shifts cancel in this expression. The gauged logit table therefore lies in V for every p and every testing parameter. Equation (54) gives

$$\alpha^{\theta}_{(y_0,\mu^\circ)}(p)_b = \frac{1}{B} \qquad \text{for every } p, \theta, b. \tag{57}$$

At all other leader states the follower features are zero, and, as in the leader construction in Appendix C, set the true follower parameters at steps $h \geq 2$ to zero. Hence the action marginal is uniform at every step, including any later visit to y_0. Thus the additional MO observation is exactly model-independent throughout the full episode.

A globally Lipschitz decoder. We recover the hidden local coordinates from a joint state–action distribution to define the leader feature. Let W be the matrix whose ith column is the vectorization of $w_i(x, b)$ over occupied rows and nonbaseline follower actions. The columns are independent, so choose a fixed left inverse L with $LW = I_D$. All follower rewards in the testing family lie in $[0, 1]$ and all occupied population weights are fixed and positive. Consequently there is a constant $c_{\mathrm{cell}} > 0$, depending only on the fixed dimensions and bounds, such that each joint distribution induced by the constructed follower responses satisfies, on every occupied row,

$$\mathcal{P}(x, b) \geq c_{\mathrm{cell}}.$$

For example, one may take $c_{\mathrm{cell}} = \mu_{\min}/(Be^{\eta})$, where $\mu_{\min} = \min_{x \leq m} \mu_x$. Put $\lambda = c_{\mathrm{cell}}/2$ and, for an arbitrary joint distribution $\mathcal{P} \in \Delta(\mathcal{Z})$, define

$$\widetilde{r}_{xb}(\mathcal{P}) = \frac{1}{\eta\kappa} \left[\log \max\{\mathcal{P}(x, b), \lambda\} - \log \max\{\mathcal{P}(x, B), \lambda\}\right]$$

for occupied x and $b < B$. Define

$$\ell_i(\mathcal{P}) = \mathrm{clip}_{[0,2\tau]}\big[(L\widetilde{r}(\mathcal{P}))_i\big], \qquad i \in [D]. \tag{58}$$

The map in (58) is globally Lipschitz on the ambient simplex because $v \mapsto \log \max\{v, \lambda\}$ is $1/\lambda$-Lipschitz and coordinate clipping is nonexpansive. On every joint distribution induced by the constructed follower responses, no clipping occurs in the cell log-ratios, and (56) yields

$$\ell_i\big(\mathcal{P}_\theta((y_0, \mu^\circ), p)\big) = \theta_i p(a_i).$$

Let $L_\ell < \infty$ be a fixed Lipschitz constant for this decoder on the ambient joint simplex.

Shared feature maps and packing. Fix $0 < c_* < \tau/(4D)$ and then fix $0 < \epsilon_0 < c_*/4$. When invoking the fixed-budget leader construction below, take its range constant to be $a_0 = \epsilon_0$. For $0 < \epsilon \leq \epsilon_0$, define on the full joint-distribution simplex

$$b_\epsilon(\mathcal{P}) = \left[\epsilon - \max_{1 \leq i \leq D} |\ell_i(\mathcal{P}) - c_*|\right]_+. \tag{59}$$

Its Lipschitz constant is at most L_ℓ, uniformly in ϵ. At the joint distribution induced by the null parameter, all decoder coordinates are zero, so $b_\epsilon = 0$. For a testing parameter $\theta \in [\tau, 2\tau]^D$, the maximizing coordinates satisfy

$$p(a_i) = p_i^*(\theta) := \frac{c_*}{\theta_i}, \qquad i \in [D],$$

and their total mass is less than $1/4$. The map $\theta \mapsto (p_1^*(\theta), \ldots, p_D^*(\theta))$ is bi-Lipschitz from the fixed parameter cube onto the fixed coordinate box $[c_*/(2\tau), c_*/\tau]^D$, and every point of this box has coordinate sum less than $1/4$. Choose a maximal subset of this image box whose centers are separated

by $4\epsilon/\tau$ in ℓ_∞ distance, and take the corresponding testing parameters in $[\tau, 2\tau]^D$. Maximality makes the ℓ_∞ balls of radius $4\epsilon/\tau$ around the centers cover the box, so a volume comparison gives at least

$$K_\epsilon \geq c_D \epsilon^{-D}$$

centers for all sufficiently small ϵ. If the positive-bump regions for two distinct testing parameters intersected, then at an intersecting commitment p the inequalities $|\theta_i p(a_i) - c_*| < \epsilon$ and $|\theta'_i p(a_i) - c_*| < \epsilon$ would imply

$$|p_i^*(\theta) - p_i^*(\theta')| < 2\epsilon/\tau \qquad \text{for every } i,$$

contradicting the chosen separation. Thus the positive-bump regions $U_1, \ldots, U_{K_\epsilon}$ are pairwise disjoint in the full commitment simplex. At the center of U_j the bump has height ϵ, while it vanishes outside U_j.

Insert (59) into the fixed-budget leader construction (32)–(33) from Appendix C. The function b_ϵ obeys $0 \leq b_\epsilon \leq \epsilon \leq a_0$, vanishes at the joint distribution induced by the null parameter, and has a Lipschitz constant uniform in ϵ. These are exactly the properties of b used by the verification of the leader construction. With L_ℓ and a_0 now fixed, choose the multiplier $c_b > 0$ so that $c_b a_0 \leq c_y$ and $c_b L_\ell \leq L_\phi$, as required by that appendix. Its verification then gives the leader-feature norm, population Lipschitzness, valid probability normalization, the factorization (1), the reward bounds, and Bellman closure for the fixed $R_U > 1$. The follower feature (55) and the leader feature using b_ϵ are shared by the null and every alternative at that scale. All constants are independent of T and ϵ, and allowing every $0 < \epsilon \leq \epsilon_0$ keeps the construction inside one fixed-budget minimax class in the sense of Section 3.5.

Complete-transcript information and regret. We finally verify the information calculation for the full MO transcript. Let $\mathbb{Q}_0$ be the complete observable transcript law under the null model and $\mathbb{Q}_j$ the law under alternative j, including the learner's internal randomization, committed policies, observed state marginal, observed action marginal, leader action, leader reward, and next augmented state. Conditional on any common observed history, the algorithm uses the same randomized policy-selection kernel in all testing models, and the realized leader action is sampled from that announced commitment under the same kernel. The state marginal is the fixed μ° and, by (57), the action marginal is the same uniform vector in all models for every possible commitment. In this leader construction, the only model-dependent conditional observation is the first-step Bernoulli leader-state transition, whose mean differs from the null by at most $C\epsilon \mathbf{1}\{p \in U_j\}$ and remains in a fixed compact subinterval of $(0, 1)$. The later observed reward is a common deterministic function of this transition, and all later observations have the same conditional laws. The chain rule for relative entropy therefore gives exactly the same complete-transcript estimate as in Appendix C:

$$\mathrm{KL}(\mathbb{Q}_0 \,\|\, \mathbb{Q}_j) \leq C\epsilon^2 \mathbb{E}_0 N_j, \qquad N_j = \sum_{t=1}^{T} \mathbf{1}\{p_1^t \in U_j\}.$$

The observed action marginal contributes zero conditional divergence at every step.

Pinsker's inequality applied to the bounded statistic N_j/T, followed by Cauchy–Schwarz and the disjointness relation $\sum_j N_j \leq T$, gives the same average-visits inequality as (35):

$$\frac{1}{K_\epsilon} \sum_{j=1}^{K_\epsilon} \frac{\mathbb{E}_j N_j}{T} \leq \frac{1}{K_\epsilon} + C\epsilon \sqrt{\frac{T}{K_\epsilon}}.$$

Choose $\epsilon = c_{\mathrm{loc}} T^{-1/(D+2)}$ with a fixed sufficiently small $c_{\mathrm{loc}} > 0$ so that, for every $T \geq 1$, the packing estimate applies, $\epsilon \leq \epsilon_0$, and the right-hand side is at most $1/4$. Some alternative then satisfies $\mathbb{E}_j N_j \leq T/4$, and Markov's inequality gives

$$\Pr_j\{N_j \leq T/2\} \geq \frac{1}{2}.$$

The value identity (34) implies pathwise on that alternative that

$$\mathrm{Reg}_j(T) \geq c_{\mathrm{sig}} \epsilon (T - N_j).$$

Consequently the regret is at least a fixed multiple of $T^{(D+1)/(D+2)}$ with probability at least $1/2$, which implies (52) with the stated uniform constant $1/4$. Multiplying the threshold in (52) by its probability lower bound proves (53). □

G.3 THE CAPACITY CRITERION AND THE DIVISIBLE-ACTION SEPARATION

Proof of Theorem 5. Under (17), the SO effective dimension in (10) is $r = d_F$. Suppose first that $C_{\mathrm{lin}}(X, B) \geq d_F$. The MO learner may always ignore the additional action marginal, so the SO upper bound gives

$$\mathfrak{R}^*_{\mathrm{MO}}(T) \leq \widetilde{O}\left(T^{\frac{d_F+1}{d_F+2}}\right).$$

Conversely, $D = d_F$ is admissible in Lemma 18 because (17) also gives $d_F \leq A - 1$. Equation (53) therefore gives the matching lower bound, proving the first conclusion.

Suppose next that $C_{\mathrm{lin}}(X, B) < d_F$. Since both quantities are integers,

$$\max\{C_{\mathrm{lin}}(X, B), d_F - 1\} = d_F - 1.$$

The other three entries in (38) are at least d_F under (17), so $r_{\mathrm{MO}} = d_F - 1$. The high-probability bound (39), followed by the expectation conversion in Appendix D, yields

$$\mathfrak{R}^*_{\mathrm{MO}}(T) \leq \widetilde{O}\left(T^{\frac{d_F}{d_F+1}}\right).$$

Because

$$\frac{d_F}{d_F+1} < \frac{d_F+1}{d_F+2},$$

this is strictly smaller, up to logarithmic factors, than the SO minimax order from Corollary 4. Hence SO and MO have the same minimax temporal order exactly when $C_{\mathrm{lin}}(X, B) \geq d_F$ in the stated parameter regime. □

Size of the exponent improvement. Under (17), when $C_{\mathrm{lin}}(X, B) < d_F$, the comparison of the MO upper bound with the SO minimax rate yields a quantitative gain. The resulting exponent improves by at least

$$\frac{d_F+1}{d_F+2} - \frac{d_F}{d_F+1} = \frac{1}{(d_F+1)(d_F+2)}.$$

Realizing the divisible-action capacity. The capacity admits a simple exact value for a family with arbitrarily many follower actions. In general, $C_{\mathrm{lin}}(X, B) \leq X - 1$. This bound follows by differentiating the constant-aggregate identity: the first derivative gives a positive weighted dependence among row response directions, while the second derivative gives a positive-definite covariance on their effective span. A Gram matrix of X vectors with such a dependence has rank at most $X - 1$. When X divides B, cyclically balanced action groups attain this bound.

Proof of Proposition 6. Write $B = nX$ and partition the actions into X groups of n actions. Give the X follower states equal population mass. Assign to each state a cyclic permutation of an arbitrary X-vector of group logits, repeating each group logit within its group. If the softmax probabilities of the group logits are $q_1, \ldots, q_X$, an individual action has conditional probability q_j/n in the corresponding group. Averaging all cyclic permutations gives $1/(nX) = 1/B$ for every action. After removing row-wise additive constants, this family has dimension $X - 1$, which attains the general upper bound in Lemma 14. □

Proof of Corollary 7. Here $d_F = X$, $A - 1 \geq X$, and $X(B - 1) \geq X$, so (10) gives the SO dimension $r = X$. The SO minimax characterization Corollary 4 therefore yields

$$\mathfrak{R}^*_{\mathrm{SO}}(T) = \widetilde{\Theta}\left(T^{\frac{X+1}{X+2}}\right).$$

By Proposition 6, $C_{\mathrm{lin}}(X, B) = X - 1$. Substitution into (38) gives $r_{\mathrm{MO}} = X - 1$, so (39) and Appendix D imply

$$\mathfrak{R}^*_{\mathrm{MO}}(T) \leq \widetilde{O}\left(T^{\frac{X}{X+1}}\right).$$

For the reverse inequality, take $D = X - 1$ in Lemma 18. This choice is admissible because $X - 1 \leq d_F$, $X - 1 \leq A - 1$, and $X - 1 = C_{\mathrm{lin}}(X, B)$. Equation (53) gives

$$\mathfrak{R}^*_{\mathrm{MO}}(T) \geq \Omega\left(T^{\frac{X}{X+1}}\right).$$

Combining the two bounds proves (19) for every multiple $B = nX$. □

H EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

This appendix presents the numerical experiments, followed by the environment, implementation, and evaluation details. The construction separates the information supplied by action marginals from the leader's underlying decision problem, so the SO/MO comparison isolates the effect of that information.

H.1 OVERVIEW AND MAIN RESULTS

We study two mechanisms underlying the theory: the effect of a larger effective dimension under SO, and the benefit of action marginals that distinguish candidate responses. The experiments compare learning trajectories in controlled linear Stackelberg mean-field games, with the same leader rewards and transitions across the two marginal-information constructions.

Experimental setup. We use horizon length $H = 3$, state and action cardinalities $S = A = X = B = 4$, leader feature dimension $d_L = 5$, and follower feature dimension $d_F \in \{2, 3\}$. The SO effective dimension is therefore $r = d_F$. The population state distribution is uniform, and the leader starts in a state where the hidden follower response controls the probabilities of reaching subsequent states with different continuation values. Two constructions vary the information carried by action marginals: cyclically permuted responses across follower states produce an *uninformative* marginal, whereas aligned responses produce an *informative* marginal. For any fixed commitment and parameter, both constructions induce the same leader rewards and transitions. We implement candidate-response Bellman optimism using a finite, adaptively refined set of candidate parameters; MO additionally enforces consistency with the observed action marginals. All population observations are exact, and each policy is fixed before its episode begins.

Each setting is evaluated for $T = 10,000$ episodes on ten new seeds after fixing the numerical confidence coefficient at 0.002 in a separate calibration experiment. We pair the games and trajectory random numbers across observation regimes and report cumulative pseudo-regret as defined in (3), including all initialization and diagnostic episodes. The optimal commitment value and each executed policy value are evaluated by exact dynamic programming over the continuous commitment benchmark. Appendices H.2 to H.4 give the environment, implementation, and evaluation details.

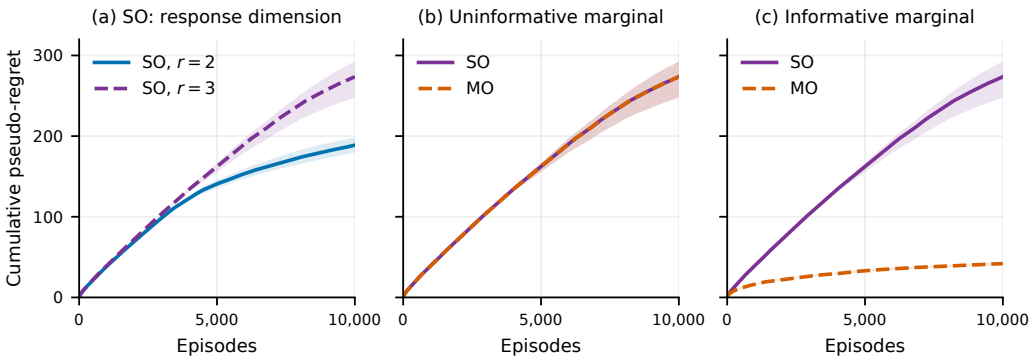

Figure 1: Learning from population marginals over ten confirmation seeds. Curves show mean cumulative pseudo-regret, with pointwise 95% bootstrap intervals. (a) SO with effective dimensions $r = 2$ and $r = 3$. (b,c) SO and MO with uninformative and informative action marginals, respectively, at SO effective dimension $r = 3$. The paired SO/MO trajectories coincide in (b); informative marginals substantially reduce regret in (c).

Effective dimension under SO. Figure 1(a) shows a clear dimension effect: the mean cumulative regret increases from 188.77 at $r = 2$ to 273.57 at $r = 3$, a 44.92% increase. Both learners improve their commitments over time, with mean per-episode regret decreasing from 0.03865 to 0.00721 for $r = 2$ and from 0.03929 to 0.01552 for $r = 3$ when comparing the first and last 1,000 episodes. The slower improvement at $r = 3$ illustrates the greater exploration burden associated with a larger effective dimension, the mechanism quantified by Corollary 4.

When the extra marginal helps. In the uninformative construction, every executed action marginal is uniform, and the paired SO and MO trajectories coincide for all ten seeds in Figure 1(b). The informative construction supplies a different outcome: a full-support commitment identifies the follower parameter from the exact action marginal, after which MO continues learning the leader's unknown reward and transition coefficients. Its mean cumulative regret is 41.97, an 84.66% reduction relative to SO, with a paired bootstrap 95% interval of $[83.08\%, 85.83\%]$. Every confirmation seed benefits from the additional information. The exact regret decomposition in Appendix H.5 attributes 98.84% of the mean improvement to better commitments at the initial state, connecting the gain directly to learning the hidden response. Together, these comparisons show how cancellation can remove the information in action marginals and how informative marginals can accelerate learning, complementing the geometric analysis in Section 6.

The experiments exhibit greater learning difficulty with a larger effective dimension and show how informative action marginals improve commitments in otherwise matched games.

H.2 Two games with matched leader rewards and transitions

For this construction, index the leader and follower states and actions by $\{0, 1, 2, 3\}$, and write the follower parameter coordinates as $\theta_0, \ldots, \theta_{d_F-1}$. Every episode starts at leader state 0, and $\mu_h(x) = 1/4$ at every step. Set $\eta = 5$ and use a common follower parameter across steps, drawn coordinatewise from $\mathrm{Unif}[0.25, 0.55]$. To pair dimensions, we draw three coordinates for each seed and use the first two for the $d_F = 2$ game. All these parameters lie in the nonnegative unit ball. The follower features reveal that this is the part of the unit ball satisfying the follower-reward constraints.

Follower responses and their marginals. For a commitment p, let $\ell_j(p) = \theta_j p(j)$ for $0 \le j < d_F$ and $\ell_j(p) = 0$ for $d_F \le j \le 3$. Define the base response

$$\rho_j^\theta(p) = \frac{\exp(\eta \ell_j(p))}{\sum_{k=0}^{3} \exp(\eta \ell_k(p))}.$$

In the informative construction, every follower state uses $\nu^{p,\theta}(b \mid x; s) = \rho_b^\theta(p)$. In the uninformative construction, it uses $\nu^{p,\theta}(b \mid x; s) = \rho^\theta_{(b+x) \bmod 4}(p)$. Both are quantal responses with unit-norm features: writing $j = b$ in the informative case and $j = (b + x) \bmod 4$ in the uninformative case, set $\varphi(s, a, x, b) = e_j$ if $j < d_F$ and $a = j$, and set it to zero otherwise. Here e_j is the corresponding coordinate vector in $\mathbb{R}^{d_F}$. Thus the follower rewards lie in $[0, 1]$.

Uniform population weights imply

$$\alpha^\theta_{\text{uninformative}}(p) = (1/4, 1/4, 1/4, 1/4), \qquad \alpha^\theta_{\text{informative}}(p) = \rho^\theta(p).$$

The first identity holds at every commitment and parameter because summing the four cyclic permutations sums every response coordinate once. For the second construction, any full-support commitment gives the exact inverse

$$\theta_j = \frac{\log\big(\alpha(j)/\alpha(3)\big)}{\eta p(j)}, \qquad 0 \le j < d_F. \tag{60}$$

Both constructions have the same conditional response at follower state $x = 0$. We use this common row to define the leader features, ensuring matched leader rewards and transitions across constructions.

Leader features and transitions. Let $c = 0.075$, $w = 0.035$, and $\varepsilon_0 = 10^{-3}$. For a joint distribution $\mathcal{P}$, define the response score

$$q_j(\mathcal{P}) = \frac{1}{\eta} \log \frac{\max\{\mathcal{P}(0, j), \varepsilon_0\}}{\max\{\mathcal{P}(0, 3), \varepsilon_0\}}, \qquad \psi(\mathcal{P}) = \left[1 - \frac{\max_{0 \le j < d_F} |q_j(\mathcal{P}) - c|}{w}\right]_+,$$

where $[v]_+ = \max\{v, 0\}$. For every nonnegative unit-ball parameter and every commitment, $\ell_j(p) \ge 0$ and $\sum_{j=0}^{3} \ell_j(p) \le 1$. Convexity of the exponential sum on this simplex gives $\sum_j e^{5\ell_j(p)} \le e^5 + 3$, and hence every entry in row $x = 0$ of the induced joint distribution satisfies

$$\mathcal{P}(0, j) \ge \frac{1}{4(e^5 + 3)} > 10^{-3} = \varepsilon_0.$$

Thus the floor is inactive in either construction and $q_j(\mu \otimes \nu^{p,\theta}) = \theta_j p(j)$ throughout this response family. The score therefore measures how closely a commitment places the unknown response at the target c in every relevant direction.

At the initial leader state, use

$$\phi(0, a, \mathcal{P}) = (0.7,\, 0.3\psi(\mathcal{P}),\, 0,\, 0,\, 0).$$

For $s \in \{1, 2, 3\}$, set $(g_1, g_2, g_3) = (-1, 0.2, 1)$, $\zeta_a = -1 + 2a/3$, and

$$\phi(s, a, \mathcal{P}) = (0.7,\, 0,\, 0.2g_s,\, 0.2\zeta_a,\, 0.2g_s\zeta_a).$$

The unknown leader reward coefficient at step $h \in \{1, 2, 3\}$ is

$$\xi_h^* = \big(5/7,\, 0,\, 0.30 + 0.015(h-1),\, 0.18,\, 0.15\big).$$

Writing $\phi_1, \ldots, \phi_5$ for the five leader-feature coordinates, define

$$\Delta(s, a, \mathcal{P}) = 0.5\phi_2(s, a, \mathcal{P}) + 0.15\phi_3(s, a, \mathcal{P}) + 0.08\phi_4(s, a, \mathcal{P}) + 0.06\phi_5(s, a, \mathcal{P}).$$

The next leader state has probabilities

$$\mathcal{T}_{L,h}(\cdot \mid s, a, \mathcal{P}) = \big(0,\, 1/3 - \Delta(s, a, \mathcal{P}),\, 1/3,\, 1/3 + \Delta(s, a, \mathcal{P})\big),$$

and the next population distribution is uniform. The constant terms are linear in $\phi_1 = 0.7$, so this is a linear augmented transition kernel. The transition never returns to leader state 0. Consequently, the response affects the initial commitment decision, while subsequent steps require ordinary reward and transition learning.

These features have Euclidean norm below one, the reward coefficients have norm below one, and rewards lie in $[0, 1]$. For any joint distribution, $|\Delta| \le 0.15 < 1/3$, so the transition probabilities are valid on the full feature domain. The clipped log-ratio map is $1/(\eta\varepsilon_0)$-Lipschitz from the ℓ_1 norm on joint distributions to the ℓ_∞ norm on response coordinates, so the population Lipschitz constant of the leader feature is at most

$$\frac{0.3}{w\eta\varepsilon_0} = \frac{12000}{7} < 2000.$$

For Bellman closure, take any Borel continuation v with values in $[0, H]$ and write $v_s = v(s, \mu)$ at the uniform population. The coefficient vector representing $\mathbb{P}_h v$ is

$$a(v) = \big(\tfrac{10}{21}(v_1 + v_2 + v_3),\, 0.5(v_3 - v_1),\, 0.15(v_3 - v_1),\\ 0.08(v_3 - v_1),\, 0.06(v_3 - v_1)\big).$$

It satisfies $\|a(v)\|_2 \le H\sqrt{(10/7)^2 + 0.2825} < 2H$. Consequently $\|\xi_h^* + a(v)\|_2 \le 1 + 2H \le 3H < 4H$, verifying the stated budget $R_U = 4$ on the full input domain. All five leader-feature coordinates are active. The leader rewards and transitions coincide across the informative and uninformative constructions because their row at $x = 0$ coincides and all other leader features are independent of the response.

H.3 Numerical implementation of Bellman optimism

The implementation follows the candidate-response evaluation principle of Section 4.1, using a finite set of candidate parameters, ball-constrained ridge fits, and a finite commitment search. These choices specify the numerical optimization used for the experiments. Both SO and MO use the same fitting, planning, and refinement settings; their candidate-response constraints reflect their respective observations.

Candidate parameters and profiled fits. The initial candidate set contains 256 scrambled Sobol points mapped into the nonnegative unit ball, together with zero and the d_F coordinate vectors. Returns are normalized by H. At a planning update after n completed episodes, let $f_{i,h}^\theta = \phi(s_h^i, a_h^i, \mu_h^i \otimes \nu^{p_h^i,\theta})$ and $y_{i,h} = u_h^i/H + V_{h+1}(z_{h+1}^i)$, where V_{h+1} is the current optimistic continuation. For each candidate, compute

$$\widehat{w}_h^\theta \in \arg\min_{\|v\|_2 \le R_U} \left\{ \sum_{i=1}^n \big(\langle f_{i,h}^\theta, v\rangle - y_{i,h}\big)^2 + \lambda\|v\|_2^2 \right\}, \qquad \lambda = 10^{-3}.$$

Let $L_h(\theta)$ denote the minimum objective and $G_h^\theta = \lambda I + \sum_{i=1}^n f_{i,h}^\theta (f_{i,h}^\theta)^\top$. For the candidate set size K used at that step, the numerical confidence threshold is

$$\beta_{n,K} = 0.002 \log \frac{2KH(n+1)^2}{0.05} + \lambda R_U^2.$$

The remaining confidence budget for a candidate is $b_h(\theta) = \beta_{n,K} - [L_h(\theta) - \min_\vartheta L_h(\vartheta)]$. SO retains candidates with $b_h(\theta) \geq 0$. For each retained candidate θ and query (z, p, a), write $f = \phi(s, a, \mathcal{P}_\theta(z, p))$ with $z = (s, \mu)$. The planner evaluates the confidence score

$$\left[\langle f, \widehat{w}_h^\theta \rangle + \sqrt{b_h(\theta)\, f^\top (G_h^\theta)^{-1} f} \right]_{[0,(H-h+1)/H]},$$

where $[v]_{[a,b]} = \min\{b, \max\{a, v\}\}$ denotes clipping. The planner compares these scores over retained candidate parameters and candidate commitments. At leader state 0, it selects the largest unclipped score among commitments whose clipped scores are within 10^{-12} of the maximum. At states $1, 2, 3$, it selects the first maximizing pure action in the fixed action order. The coefficient fits are step-specific, and a shared candidate set represents the common follower parameter of these games.

Commitments, updates, and refinement. The commitment set contains the uniform commitment, all four pure actions, a simplex grid of denominator 12 on actions $0, \ldots, d_F$, and feasible score-maximizing commitments associated with the candidate parameters. For a positive candidate θ, the latter assign $p(j) = c/\theta_j$ for $j < d_F$ and place the remaining mass on action d_F, whenever these probabilities are feasible. For $d_F = 2$, unused coordinates of the simplex grid are zero-padded to the four-action simplex. At noninitial states, the features are independent of the follower response and the action objective is linear, so the planner selects a pure action.

The first episode uses the uniform commitment at every state and step, followed by four episodes that cycle through pure actions. Planning starts after these five episodes and is then updated after every ten completed episodes. Unless the action marginal has identified the parameter, the learner proposes up to two additional candidate parameters every 250 episodes using sequential least squares programming (SLSQP), with at most 30 iterations per start and tolerance 10^{-8}. The proposal objective profiles the ridge loss at the initial step over unconstrained coefficients; candidate evaluation uses the ball-constrained fit above. Their sufficient statistics are reconstructed from the complete observed history, and their retained losses are evaluated by the same norm-constrained ridge fit. Within each backward planning pass, subsequent steps are processed before parameter refinement at the initial step. The full policy remains fixed during each episode.

Using action marginals. For informative MO, the first uniform commitment yields (60); the inferred parameter is added to the candidate set, and the action-marginal compatibility constraints retain this candidate to numerical tolerance. Parameter refinement is then unnecessary. For uninformative MO, the cancellation identity implied by the feature construction leaves the candidate set unchanged. The auxiliary minimum $\min_\vartheta L_h(\vartheta)$ always uses the entire numerical candidate set before the action-marginal compatibility constraints are imposed. Every observed action marginal is checked against the corresponding response prediction with tolerance 10^{-10}; matching an inferred parameter to an existing candidate parameter uses tolerance 10^{-8}. The learner receives the features and its permitted feedback, while the true coefficients are confined to the simulator and offline evaluator.

H.4 Continuous benchmark and statistical evaluation

For every executed policy, the evaluator computes its expected return by backward dynamic programming using the true transition probabilities and rewards. Write $W_h^*(s)$ for the optimal unnormalized continuation value at the uniform population in this experiment. At states $s \in \{1, 2, 3\}$, the step-3 action objective is

$$0.5 + 0.066 g_s + (0.036 + 0.030 g_s)\zeta_a.$$

The coefficient of ζ_a is at least $0.006 > 0$, so $a = 3$ is optimal and $W_3^*(s) = 0.536 + 0.096 g_s$. Substituting these values into the step-2 transition gives the action objective

$$1.0424 + 0.06876 g_s + (0.039072 + 0.032304 g_s)\zeta_a.$$

Quantity	Setting
Horizon length and cardinalities	$H = 3$, $S = A = X = B = 4$
Feature dimensions	$d_L = 5$, $d_F \in \{2, 3\}$
Model parameters and bounds	$\eta = 5$, $B_\Theta = 1$, $R_U = 4$, $L_\phi = 2000$
Episodes per run	$10,000$
Ridge and confidence coefficient	10^{-3} and 0.002, respectively
Initial Sobol candidates	256, plus zero and coordinate vectors
Commitment grid denominator	12
Planning schedule	After 5 completed episodes, then every 10 episodes
Refinement schedule	Every 250 episodes, at most 2 starts
Calibration seeds	$0, 1, 2$
Confirmation seeds	$10, \ldots, 19$
Environment and Sobol master seeds	314159 and 271828, respectively
Main-figure uncertainty	Pointwise 95% percentile bootstrap; 10,000 resamples

Table 3: Frozen experimental settings used for all confirmation runs.

Response construction	Regime and SO effective dimension	Mean regret	Seed SD
Uninformative	SO, $r = 2$	188.77	15.92
Uninformative	SO, $r = 3$	273.57	39.27
Uninformative	MO, $r = 3$	273.57	39.27
Informative	SO, $r = 3$	273.57	39.27
Informative	MO, $r = 3$	41.97	3.33

Table 4: Cumulative pseudo-regret at $T = 10,000$ over all ten confirmation seeds. The standard deviations describe seed-to-seed variation.

Its coefficient of ζ_a is at least $0.006768 > 0$, so again $a = 3$ is optimal and $W_2^*(s) = 1.081472 + 0.101064 g_s$. Thus an initial commitment with response score ψ followed by optimal continuation has value

$$\begin{aligned} J^*_{\mathrm{root}}(\psi) &= \frac{1}{2} + \frac{1}{3}\sum_{s=1}^{3} W_2^*(s) + 0.15\psi\big(W_2^*(3) - W_2^*(1)\big) \\ &= 1.5882096 + 0.0303192\psi. \end{aligned}$$

For the sampled parameters, the commitment $p(j) = c/\theta_j$ for $j < d_F$, with the remaining mass on action d_F, attains $\psi = 1$ because $\sum_{j<d_F} c/\theta_j \le 0.9$. Since $0 \le \psi \le 1$, the exact continuous-commitment optimum is 1.6185288, independent of the learner's numerical search grid. The slope $0.15\big(W_2^*(3) - W_2^*(1)\big) = 0.0303192$ also supplies the constant C in the regret decomposition below. We record the difference between this optimum and the exact value of each executed policy, and sum these differences to obtain (3).

The numerical confidence coefficient was selected using SO calibration runs with seeds $0, 1, 2$ and coefficients 0.02, 0.002, and 0.0002. The value 0.002 improved learning while retaining the true Bellman coefficients in the checked calibration trajectories, and was fixed before evaluating the ten confirmation seeds. Each confirmation seed defines a follower parameter and independent trajectory random streams; the same streams are paired across SO/MO and across the two response constructions. Thus the reported variability includes both instance variation and trajectory randomness. The shared Sobol sampling seed is held fixed.

The five settings are uninformative SO at $d_F = 2$, uninformative SO and MO at $d_F = 3$, and informative SO and MO at $d_F = 3$. All 50 runs complete $10,000$ episodes and all are included in the reported results. For Figure 1, we bootstrap the ten seeds with 10,000 resamples and report pointwise percentile intervals for the mean cumulative regret. Relative improvements use a paired bootstrap of the ratio of group means, preserving seed correspondence in each resample. The bootstrap random seed is 20260925. Table 4 reports the terminal means and sample standard deviations.

H.5 Learning the hidden response

Figure 2 tracks the mean per-episode regret and the response score of the executed initial commitments in consecutive 500-episode blocks. The score increases as SO learns a useful commitment, with slower progress in three dimensions. Informative MO reaches an average score of 0.9930 over the last 1,000 episodes, compared with 0.5328 for three-dimensional SO and 0.8004 for two-dimensional SO.

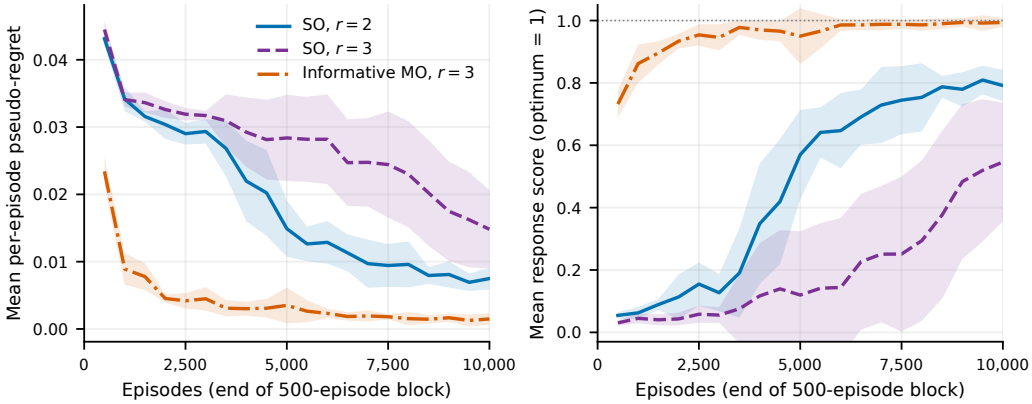

Figure 2: Learning progress over consecutive 500-episode blocks. Left: mean per-episode pseudo-regret. Right: mean response score ψ, whose maximum is one. Lines average all ten confirmation seeds, and shading shows one sample standard deviation across seeds.

The environment also permits an exact attribution of regret. Let ψ_t be the response score of the initial commitment in episode t, and let D_t be its expected continuation loss relative to optimal decisions at subsequent steps, under that commitment's next-state distribution. The initial reward is constant, and the optimal continuation value depends affinely on ψ_t. Consequently,

$$\mathrm{Reg}_M(T) = C\sum_{t=1}^{T}(1-\psi_t) + \sum_{t=1}^{T} D_t, \qquad C = 0.0303192, \quad D_t \geq 0.$$

The constant C is the difference between the optimal-continuation returns when the response score at the initial step is one and zero. The two terms have mean values 243.19 and 30.38 for three-dimensional SO, compared with 14.28 and 27.70 for informative MO. Thus 98.84% of the mean SO/MO improvement comes from the initial commitment, while the subsequent learning costs remain similar. This decomposition links the benefit of the extra marginal to its information about the hidden follower response.